Étude de la variabilité inter-individuelle du transcriptome soumis à un stimulus

Nicolas Derian

THÈSE DE DOCTORAT DE

L'UNIVERSITÉ PIERRE ET MARIE CURIE

Spécialité

Biologie des Systèmes

École doctorale Complexité du Vivant

Présentée par

Nicolas DERIAN

Pour obtenir le grade de

DOCTEUR de L'UNIVERSITÉ PIERRE ET MARIE CURIE

Sujet de thèse :

Étude de la variabilité inter-individuelle du transcriptome soumis à un stimulus.

Soutenue le 21 septembre 2016, devant un jury composé de :

M. Philippe SEKSIK	Président du jury
M. Adrien SIX	Directeur de thèse
M. Éric VICAUT	Directeur de thèse
M. Vassili SOUMELIS	Rapporteur
M. Pierre BOUDINOT	Rapporteur
M. Thomas DERRIEN	Examinateur

ABRÉVIATIONS ..... **VI**

INTRODUCTION **1**

CONTEXTE GÉNÉRAL **1**

LA VARIABILITÉ EN BIOLOGIE **3**

ORIGINES ET RÔLE **3**

Origines *4*

Un moyen de créer de la diversité *10*

Le système immunitaire *12*

VARIABILITÉ INTER-INDIVIDUELLE **19**

De l'évolution à la médecine personnalisée *19*

Variabilité inter-individuelle et réponse *25*

APPROCHES MÉTHODOLOGIQUES **34**

LE TRANSCRIPTOME **34**

ORIGINES DES DONNÉES **40**

LPS *41*

Tolérance fœto-maternelle *44*

Lymphocytes T régulateurs et IL-2 ..... *47*

MÉTHODES D'ANALYSE CLASSIQUES **51**

Les données *51*

Analyse des données *56*

LES INDICES DE DIVERSITÉ **68**

Mesures d'entropie et vraie diversité *69*

Application au transcriptome : un cas bibliographique *72*

Spécificité et coefficient de variation *75*

TRAITEMENT DES DONNÉES **83**

DIVERSITÉ **88**

RÉSULTATS **88**

DISCUSSION **100**

SPÉCIALISATION **106**

RÉSULTATS **106**

DISCUSSION **113**

AUTRES APPLICATIONS **118**

DONNÉES APPARIÉES **118**

CLASSER LES TRANSCRITS **121**

DISCUSSION GÉNÉRALE **128**

BIBLIOGRAPHIE **134**
ANNEXE #1 ..... **144**
ANNEXE #2 ..... **162**
ANNEXE #3 ..... **176**

FIGURE 1. VARIABILITÉ TECHNIQUE : PRÉCISION ET EXACTITUDE. 5
FIGURE 2. VARIABILITÉ TECHNIQUE POUR UN JEU DE PUCE À ADN. 6
FIGURE 3. LES CELLULES DU SYSTÈME IMMUNITAIRE 14
FIGURE 4. LES RÉCEPTEURS DES LYMPHOCYTES T. 16
FIGURE 5. DIVERSITÉ ET PLASTICITÉ CELLULAIRE. 18
FIGURE 6. MODES D'INTERACTION DE DEUX STIMULI. 19
FIGURE 7. VARIABILITÉ INTRA-INDIVIDUELLE CHEZ *SCELOPORUS OCCIDENTALIS*. 21
FIGURE 8. CLAIRANCE DU LOPINAVIR CHEZ L'ENFANT. 23
FIGURE 9. SPECTRATYPES DE CDR3. 24
FIGURE 10. RELATION ENTRE FLUCTUATION ET RÉPONSE PAR K. KANEKO. 26
FIGURE 11. RELATION FLUCTUATION D'EXPRESSION ET RÉPONSE CHEZ LA LEVURE 27
FIGURE 12. ANALYSE EN COMPOSANTES PRINCIPALES : EXPÉRIENCE 1. 28
FIGURE 13. CASCADE DE RÉACTIONS INDUITES PAR LE LPS. 30
FIGURE 14. ANALYSE EN COMPOSANTES PRINCIPALES : EXPÉRIENCE 2. 31
FIGURE 15. ANALYSE EN COMPOSANTES PRINCIPALES : EXPÉRIENCE 3. 32
FIGURE 16. LA TRANSCRIPTION. 36
FIGURE 17. LES PUCES À ADN. 38
FIGURE 18. IMPACT DE LA PROFONDEUR DE SÉQUENÇAGE ET DE LA TECHNOLOGIE EN RNASEQ. 40
FIGURE 19. MODULES DE GEVADSS. 43
FIGURE 20. ENRICHISSEMENT DE SIGNATURES DANS L'ENVIRONNEMENT FŒTAL. 45
FIGURE 21. RELATION ENTRE ENVIRONNEMENTS FŒTAL ET TUMORAL. 46
FIGURE 22. RÔLES DE L'IL-2 SUR LES LYMPHOCYTES. 50
FIGURE 23. IMPACT DE LA CORRECTION DES BIAIS TECHNIQUES SUR UN JEU DE TRANSCRIPTOME. 53
FIGURE 24. DISTRIBUTION DES DONNÉES AVANT ET APRÈS NORMALISATION. 54
FIGURE 25. EFFET DU BOOTSTRAP SUR LA CLASSIFICATION DES VECTEURS. 67
FIGURE 26. DIVERSITÉ ET SPÉCIFICITÉ. 74
FIGURE 27. DIVERSITÉ ET SPÉCIALISATION DES ORGANES HUMAINS AU NIVEAU TRANSCRIPTOMIQUE 74
FIGURE 28. DISTRIBUTIONS DES DONNÉES SIMULÉES PAR *QUANTROSIM* 76
FIGURE 29. RELATION ENTRE SPÉCIFICITÉ ET COEFFICIENT DE VARIATION #1. 77
FIGURE 30. RELATION ENTRE SPÉCIFICITÉ ET COEFFICIENT DE VARIATION #2. 78
FIGURE 31. RELATION ENTRE SPÉCIFICITÉ ET COEFFICIENT DE VARIATION #3. 79
FIGURE 32. RELATION ENTRE SPÉCIFICITÉ ET COEFFICIENT DE VARIATION #4. 80
FIGURE 33. RELATION ENTRE SPÉCIFICITÉ ET COEFFICIENT DE VARIATION #5. 81
FIGURE 34. PVCA SUR DONNÉES LPS AVANT ET APRÈS NORMALISATION. 86
FIGURE 35. PVCA SUR DONNÉES TOLÉRANCE FŒTO-MATERNELLE AVANT ET APRÈS NORMALISATION. 87
FIGURE 36. DIVERSITÉ AU SEIN DU JEU DE DONNÉES LPS SELON L'INDICE DE SHANNON. 90

FIGURE 37. DIVERSITÉ AU SEIN DU JEU DE DONNÉES LPS SELON L'INDICE DE SIMPSON. 91
FIGURE 38. DIVERSITÉ AU SEIN DU JEU DE DONNÉES TOLÉRANCE FŒTO-MATERNELLE SELON L'INDICE DE SHANNON. 92
FIGURE 39. DIVERSITÉ AU SEIN DU JEU DE DONNÉES TOLÉRANCE FŒTO-MATERNELLE SELON L'INDICE DE SIMPSON. .. 93
FIGURE 40. DIVERSITÉ AU SEIN DU JEU DE DONNÉES TREG SELON L'INDICE DE SHANNON. 94
FIGURE 41. DIVERSITÉ AU SEIN DU JEU DE DONNÉES TREG SELON L'INDICE DE SIMPSON. 95
FIGURE 42. SIMILARITÉ DANS LES GROUPES EXPÉRIMENTAUX DU JEU DE DONNÉES LPS. 97
FIGURE 43. SIMILARITÉ DANS LES GROUPES EXPÉRIMENTAUX DU JEU DE DONNÉES TOLÉRANCE FŒTO-MATERNELLE. 98
FIGURE 44. SIMILARITÉ DANS LES GROUPES EXPÉRIMENTAUX DU JEU DE DONNÉES TREG. 99
FIGURE 45. SPÉCIALISATION AU SEIN DU JEU DE DONNÉES LPS. 107
FIGURE 46. SPÉCIALISATION AU SEIN DU JEU DE DONNÉES TOLÉRANCE FŒTO-MATERNELLE. 109
FIGURE 47. SPÉCIALISATION AU SEIN DU JEU DE DONNÉES TREG. 110
FIGURE 48. SIGNATURES SPÉCIFIQUES DANS LE JEU DE DONNÉES TREG : PATIENTS 1, 6 ET 7. 112
FIGURE 49. SIGNATURES SPÉCIFIQUES DANS LE JEU DE DONNÉES TREG : PATIENTS 4,7 ET 9 112
FIGURE 50. DIVERSITÉ DES LOG-RATIOS DU JEU DE DONNÉES TREG. 118
FIGURE 51. DIVERSITÉ DES LOG-RATIO SELON DIFFÉRENTES VALEURS DE Q. 119
FIGURE 52. SPÉCIALISATION DES VARIATIONS DE GÈNES AU SEIN DU JEU DE DONNÉES TREG. 120
FIGURE 53. ANALYSE DE LA DISTRIBUTION DES TAILLES DE GÈNES. 124
FIGURE 54. LONGUEUR DES GÈNES : DIVERSITÉ SELON L'INDICE DE SHANNON. 125
FIGURE 55. LONGUEUR DES GÈNES : SIMILARITÉ. 126
FIGURE 56. LONGUEUR DES GÈNES : SPÉCIALISATION. 127

TABLE 1 : SYSTÈME IL-2/IL-2R DANS LES LTREG ET LES LT EFFECTEURS (CHENG *ET AL*, 2011) 48
TABLE 2. LES INDICES DE DIVERSITÉ (SOURCE : JOST *ET AL*). 71
TABLE 3. ENRICHISSEMENT FONCTIONNEL DES GÈNES LES PLUS ABONDANTS 104
TABLE 4. VALEURS DE SPÉCIALISATION DES GROUPES EXPÉRIMENTAUX. 114

ABRÉVIATIONS

ACI : Analyse en composantes indépendantes

ACP : Analyse en composantes principales

ADN : Acide désoxyribonucléique

ARN : Acide ribonucléique

BCR : Récepteur des lymphocyte B

CD4 : Cluster de différenciation 4

CD8 : Cluster de différenciation 8

CD25 : Cluster de différenciation 25

CDR3 : Complementary-determining region 3

DC : Cellule dendritique

ES : Score d'enrichissement

FOXP3 : Forkhead box P3

FDR : False discovery rate (taux de faux positifs)

GSEA : Gene Set Enrichment Analysis

IL-2 : Interleukine-2

IL-4 : Interleukine-4

IL-7 : Interleukine-7

IL-12 : Interleukine-12

IL-17 : Interleukine-17

LB : Lymphocyte B

LPS : Lipopolysaccharide

LT : Lymphocyte T

LTCD4 : Lymphocyte T exprimant le CD4

LTCD8 : Lymphocyte T exprimant le CD8

LTreg : Lymphocyte T régulateur

NES : Score d'enrichissement normalisé

NFkB : Nuclear factor-kappa B

PBMC : Cellules mononucléaires du sang périphérique

PBS : Tampon phosphate salin

PCR : Réaction de polymérisation en chaîne

PDE3B : Phosphodiestérase 3B

PVCA : Analyse en composantes de variation principales

RIN : RNA integrity number

SIDA : Syndrôme d'immuno-déficience acquise

STAT5 : Signal transductor and activator of transcription 5

TCR : Récepteur des lymphocytes T

INTRODUCTION

CONTEXTE GÉNÉRAL

Le laboratoire Immunologie-Immunopathologie-Immunothérapie (i3, www.i3-immuno.fr) est un laboratoire d'immunologie, dirigé par le Pr David Klatzmann, dont les compétences de recherche concernent la vaccinologie, les processus de tolérance immunitaire et l'étude des pathologies auto-immunes et auto-inflammatoires. Nous sommes par conséquent un laboratoire mêlant recherche fondamentale en biologie, recherche clinique et compétences en analyse de données.

Depuis quelques années, i3 s'attache à analyser les pathologies auto-immunes et auto-inflammatoires au travers de nombreux essais cliniques visant soit à comprendre ces pathologies d'un point de vue moléculaire soit de les traiter via un protocole mis en place par i3 et consistant à l'injection à faibles doses de la molécule d'interleukine-2 (IL-2).

- Le projet Transimmunom regroupe un consortium de laboratoire au sein d'un Laboratoire d'Excellence (LabEx) avec pour objectif la caractérisation moléculaire de 15 pathologies auto-immunes et/ou auto inflammatoires. Historiquement, le projet est né de l'expertise du laboratoire dans l'analyse des processus inflammatoires, auto-immuns et de la description de ces maladies comme appartenant à un continuum de pathologies partageant des composantes auto-immunes et auto-inflammatoires de manière plus ou moins importantes (McGonagle and McDermott, 2006). La caractérisation moléculaire se fait par l'analyse systématique du génome par la cartographie des polymorphismes nucléotidiques – du transcriptome, en utilisant les dernières innovations en matière de séquençage haut-débit – du répertoire des LT, là aussi grâce au séquençage haut-débit – du cytome, l'étude des caractéristiques cellulaires, par l'utilisation de treize panels de marqueurs de cellules, fournissant des informations sur plus de cinquante populations cellulaires – du microbiome en utilisant les techniques de méta-génomique – et enfin toutes les données clinico-biologiques liées aux patients.
- Le laboratoire conduit six essais cliniques portant sur un total de quinze pathologies associées au système immunitaire. Les patients inclus sont tous traités avec de l'IL-2 à

faible dose dont l'efficacité fut rapportée en 2011 par notre laboratoire dans un essai clinique portant sur des patients atteints de vascularites induites par le virus de l'hépatite C (Saadoun et al., 2011). Conjointement, une équipe américaine montrait les effets bénéfiques de cette molécule dans des cas de patient atteints de la maladie dite *graft-versus-host* (GVH, (Koreth et al., 2011).

C'est notamment dans ces essais que l'intérêt du laboratoire pour la variabilité inter-individuelle a émergé. Au cours d'un essai clinique portant sur patients atteints de diabète de type-1 et traités à l'IL-2 faible dose (Rosenzwajg et al., 2015), nous avons constaté que les augmentations des lymphocyte T régulateurs (LTreg) étaient très différentes d'un patient à l'autre permettant de séparées les patients considérés comme fortement répondeurs et les autres non répondeurs. Il est alors fortement envisagé que les études à venir, effectuées sur des cohortes plus importantes, fournissent le même type de résultat. L'importance de ces variations dans le développement de molécules médicamenteuses est encore de nos jour un challenge (Thakkar et al., 2016; Verbeeck et al., 2016). Elle participe au mouvement de la dernière décennie qui tend vers la médecine personnalisée (Ginsburg and Willard, 2009). Chaque individu est unique et répond de manière unique à un traitement donné. Trouver au sein d'une cohorte de patients les sous-groupes fonctionnellement similaires apparaît aujourd'hui fondamentale, d'autant que nous avons en notre possession un ensemble d'instruments et de méthodes qui permettent de traquer ces variations, de la génomique à la phénomique.

Le projet Transimmunom ouvre par ailleurs un champ de questions concernant la variabilité :

- Qu'est qu'un continuum en biologie ?
- Les pathologies du continuum sont-elles des variations plus ou moins importantes d'une forme de pathologie plus générale ? La classification de ces maladies serait alors fortement modifiée car les distinguer les unes des autres ne serait peut-être plus dans l'intérêt de la recherche.
- Si classifications il y a, les variations au sein d'une pathologie sont-elles de mêmes natures que celles d'une autre pathologie ?

Cette thèse s'inscrit dans cette démarche, développée au laboratoire, de compréhension d'un système biologique et de l'impact d'un changement d'état sur celui-ci. Pour cela je me suis intéressé au transcriptome et nous verrons qu'à travers des jeux de données simulés et issus de la littérature, nous évaluons des stratégies d'analyse de cette variabilité inter-individuelle.

LA VARIABILITÉ EN BIOLOGIE

On peut définir la variabilité comme la dispersion (énergétique, spatiale, structurelle…) d'objets les uns par rapport aux autres. Cette dispersion peut se représenter comme une évaluation des distances qui séparent ces objets deux à deux. Cette dispersion est généralement mise en évidence par l'étude de la variation d'une mesure qui caractérise l'objet d'intérêt (intensité énergétique, position dans l'espace, taille…). Plus cette dispersion est grande, plus on considérera que les objets sont différents eux aussi.

Cette dispersion prend tout son sens en biologie lorsqu'on compare des objets structurellement ou fonctionnellement proches. En comparant deux êtres humains de même sexe, nous constatons qu'il existe d'innombrables différences, la taille, le poids, la pilosité et la pigmentation de la peau, la forme des oreilles, la couleur des yeux, etc. sans que cela ne change quoi que ce soit à l'organisme dans son ensemble. Quels que soient sa taille ou son poids, un individu est, et restera, un être humain avec l'ensemble des attributs inhérents à son espèce. On peut donc voir la variabilité en biologie comme la variation de paramètres biologiques autour de valeurs moyennes de l'organisme étudié, sans s'éloigner de manière extrême de ces valeurs moyennes. L'organisme n'est évidemment pas invariant dans le temps, les paramètres qui le définissent bougent au cours de l'évolution.

La variabilité est mesurable à toutes les échelles de l'organisme, du génome à l'individu et, sous l'apparente stabilité des micro- et macrostructures de l'organisme, se cache l'inexorable jeu du hasard. Nous pouvons considérer que les structures sont stables quel que soit le niveau biologique que l'on observe ; rien ne ressemble plus à un brin d'acide désoxyribonucléique (ADN) qu'un autre brin d'ADN, un foie qu'un autre foie, une main qu'une autre main alors même qu'à y regarder de plus près, il n'existe pas deux mains identiques. La fonction rend la forme générale de l'objet constante (téléonomie, (Monod, 1971; Pittendrigh, 1958) ; le hasard en modifie sensiblement les contours. Ainsi, nous nous retrouvons avec une série de mains, identiques dans les grandes lignes, mêmes nombres d'os, de doigts, mêmes insertions des tendons, et si la taille de la main, par exemple, peut grandement différer d'un individu à l'autre, la fonction reste la même.

ORIGINES ET RÔLE

ORIGINES

Dans le domaine des sciences biologiques, la variabilité a deux grandes origines : l'une biologique, l'autre technique. Une mesure biologique peut être modélisée comme une combinaison linéaire d'informations telles que :

Équation 1:

$$x = \alpha X + \beta Y$$

Où X et Y sont respectivement le signal d'origine biologique et le signal d'origine technique, α et β étant les coefficients représentant la part de ces deux types d'information dans la mesure x. Bien évidemment, cela n'est pas aussi simple. En effet, X et Y peuvent être eux-mêmes décomposés en un sous-ensemble de signaux provenant de tel ou tel système biologique ou de telle ou telle étape du protocole de mesure de la donnée x. Ainsi, X peut contenir l'information sur la relation d'un gène avec une voie de signalisation ou son interaction avec un autre gène. C'est par ailleurs l'idée sous-jacente de certains outils d'analyse du signal dont nous verrons, dans cette thèse, deux exemples avec l'analyse en composantes indépendantes (ACI, (Chiappetta et al., 2004) et l'analyse en composantes de variance principales (PVCA, (Boedigheimer et al., 2008). La science développe en permanence de nouvelles technologies qui améliorent considérablement la qualité de la mesure. Couplées aux stratégies d'analyse qui s'appuient sur les méthodes de nettoyage des données, il est raisonnable de penser que le bruit technique est maintenant bien contenu pour ces technologies (voir Figure 1 pour une explication des deux grandes catégories de variabilité technique).

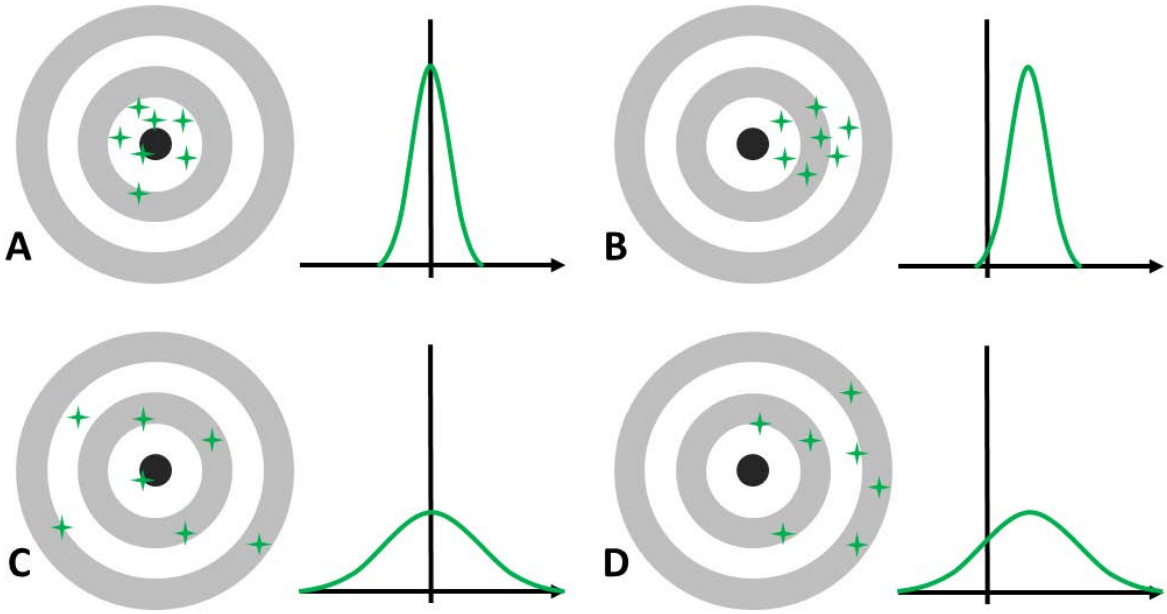


Figure 1. Variabilité technique : précision et exactitude.
Comportement de différentes séries de mesures. Le centre de la cible représente la valeur attendue pour la mesure. À droite des cibles, les courbes de densité de probabilités schématiques associées aux mesures. A : Série de mesures telles que souhaitées, précises et exactes. B : mesures inexactes mais précises : les points de mesure sont proches les uns des autres mais il y a un biais systématique qui empêche la mesure de s'approcher du centre de la cible. C : les mesures sont exactes car concentrées autour du centre de la cible mais imprécises car la dispersion des points est importante, ce qui se traduit par la courbe de distribution bien centrée mais avec une forte variance. D : Série de mesures ni précises, ni exactes. D'une manière générale, le manque d'exactitude est un biais corrigeable par la calibration des appareils, la précision quant à elle peut-être grandement améliorée mais jamais complètement éliminée.

PVCA permet de mesurer la part de variance expliquée par chaque facteur pour un échantillon donné. Nous n'enregistrons que trop rarement les informations liées à ces différents facteurs, et quand cela est le cas, la communauté ne partage ces informations que dans de trop rares occasions. Pour s'en convaincre, il suffit de regarder le niveau d'information fourni pour un jeu de données de transcriptome sur un des sites de dépôt de données publiques, comme *Gene Expression Omnibus* (http://www.ncbi.nlm.nih.gov/geo), dépôt rendu obligatoire par les journaux scientifiques pour permettre le contrôle et la pérennité des données. De mon expérience il ressort qu'un jeu de données sur ce type de site contient, bien souvent, seulement trois niveaux d'information par étude. Vous saurez quel est l'objet biologique analysé (Organe, cellules triées…). On y trouve obligatoirement la plate-forme technologique qui a permis de produire les données. Enfin on y trouve un résumé

du protocole technique consistant généralement à l'énoncé du kit utilisé pour chaque étape. Mais qui peut raisonnablement penser que chaque étude se passe sans modifications, même mineures, du protocole fournit par le fabriquant du kit ? De la même manière, il est rare qu'une seule et même personne soit toujours responsable de la même étape du protocole ou encore que le kit servant à l'étape soit lui aussi toujours le même. Or c'est bien dans ces détails que le diable de la variabilité technique se cache.

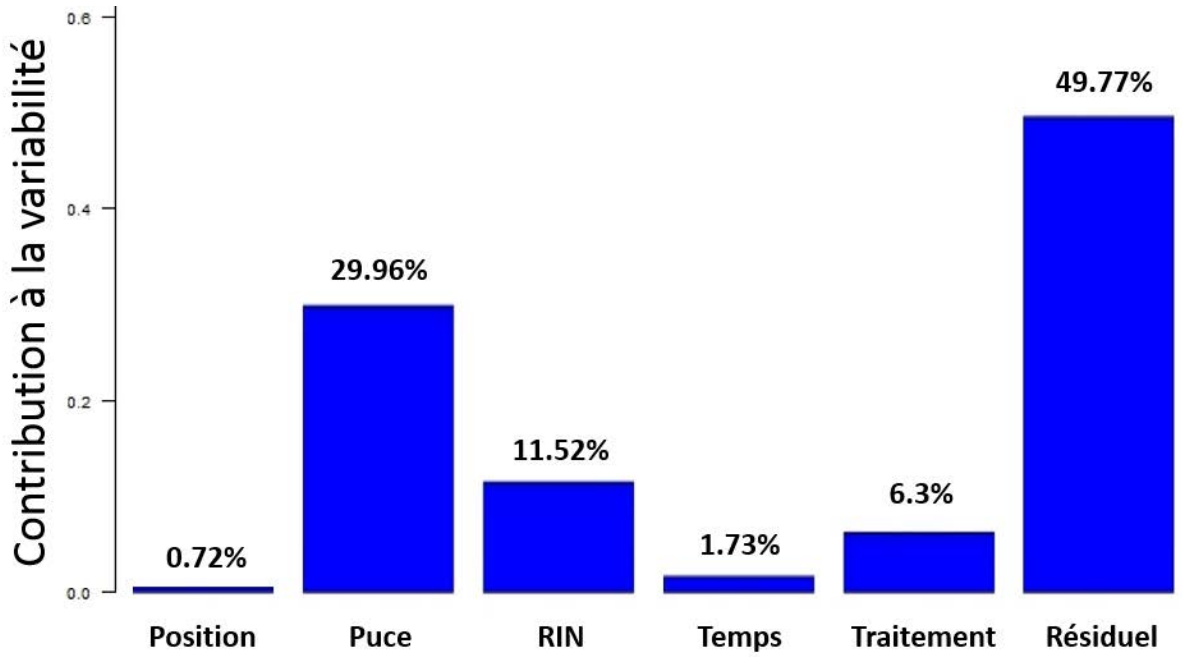


Figure 2. Variabilité technique pour un jeu de puce à ADN.
Contribution à la variabilité de différents facteurs dans un jeu de données de transcriptome produit au laboratoire. Expérience : Souris ayant subi une vaccination ou une injection de solution saline. Étude transcriptomique des cellules mononucléaires du sang périphérique (PBMC) 6 heures, 24 heures et 48 heures après l'injection, sur puces Illumina WG-6 v1. Le facteur « Puce » correspond au numéro de la lame contenant six « Positions », donc six échantillons. Le « RIN » correspond à une mesure de la qualité de l'ARN déposé. Le « Temps » correspond à la cinétique et le « Traitement » à l'injection du vaccin ou de la solution saline. Alors que l'on s'attend à avoir une forte contribution de ces deux derniers facteurs, ils s'avèrent très faibles en comparaison des autres facteurs et notamment du numéro de la lame. Le « Résiduel » est la partie de la variabilité qui n'est pas expliquée par les facteurs étudiés. Résultats obtenus au laboratoire dans le cadre du projet CompuVac.

La Figure 2 nous permet d'apprécier quelle est la part de la variabilité pouvant être expliquée par la variabilité technique dans un jeu de données de transcriptome. Des souris reçoivent une injection d'un vecteur viral, exprimant un antigène à sa surface (le gp33-41), ou une solution saline.

L'activité des gènes des cellules polynucléaires du sang est suivie au cours d'une cinétique (6 heures, 24 heures et 48 heures après injection). L'algorithme PVCA (Boedigheimer et al., 2008) indique la part de différents facteurs techniques et biologiques expliquant la variabilité du jeu de données : effrayant. Alors que l'on s'attend raisonnablement à ce que les facteurs biologiques expliquent une large part de cette variabilité, il est flagrant, dans cet exemple, que les lames et la qualité des ARN perturbent grandement le signal, ce qui n'est pas un phénomène isolé sur ce type de lame produites par Illumina (Schurmann et al., 2012). Comment considérer que les analyses de la variabilité biologique peuvent être concluantes sur de telles données ? Nous verrons plus loin que si nous ne pouvons pas toujours contrôler les informations sur les données, nous pouvons choisir avec soin ces données pour limiter l'impact de ce manque sur la qualité de l'analyse.

On l'a dit précédemment, la variabilité est un phénomène général qui s'inscrit dans le processus de développement des systèmes biologiques. Elle est par conséquent observable à toutes les échelles du vivant, du génome à l'individu. Je vais maintenant donner quelques exemples de cette variabilité aux différentes échelles pour montrer son caractère fondamental.

L'acide désoxyribonucléique (ADN), support de l'information génétique, est considéré comme une molécule particulièrement stable. Néanmoins, un certain nombre de modifications sont à mettre sur le compte de cette molécule pour expliquer l'origine de la variabilité inter-organisme et intra-organisme. La première tient à notre mode reproduction où le hasard des rencontres se mêle au hasard du brasage génétique (Masel, 2011). Ainsi, la production des gamètes par méiose est soumise à trois événements impliquant le hasard. La méiose est composée de deux étapes (méiose I et méiose II) incluant quatre phases (prophase, métaphase, anaphase et télophase). Durant la prophase I, les paires de chromosomes homologues s'apparient en plusieurs points (chiasmas) permettant l'échange ou le transfert de tout ou partie d'un chromosome vers un autre. On parle alors de brassage intra-chromosomique. Les étapes d'anaphase I et II sont sujettes au brassage inter-chromosomique. Durant ces deux étapes le matériel génétique se sépare de manière aléatoire de chaque côté de la cellule. S'ensuit la production d'une grande variété de gamètes pour la reproduction. Si l'individu ressemble à ces géniteurs, il n'en reste pas moins que la composition de son génome est donc tout à fait unique. Nous sommes ainsi nous-mêmes des facteurs de variabilité, générant des dérives génétiques au sein de la population humaine. Nous créons à chaque génération une nouvelle diversité.

À cela s'ajoute l'apparition de mutations plus ou moins influentes sur le phénotype de l'individu. Les mutations *de novo* expliquent une grande part de la variabilité génétique entre individus et peuvent impliquer des changements dans l'expression des gènes ou la fonctionnalité des protéines (Fahrner et al., 2016; Lin et al., 2016). Les modifications structurelles comme les délétions, insertions et translocations de morceaux de chromosomes jouent eux aussi dans la perpétuation d'une variabilité biologique au sein d'une population (Bochtler et al., 2015; Keim et al., 2009; Lannoy and Hermans, 2016). La découverte plus récente de l'épigénome a montré qu'il existait encore d'autres niveaux de régulation de la diversité génétique (Liu et al., 2015; Patel and Wang, 2013). Un point intéressant concernant l'épigénome est son caractère transitoire, s'il semble se transmettre de génération en génération, il se modifie avec le temps. L'épigénome modifie ainsi durablement mais pas définitivement l'expression des gènes et il suffit de quelques générations chez la souris pour voir une modification de l'épigénome (Manikkam et al., 2014). L'épigénome est un des facteurs de régulation de la transcription des gènes (Piunti and Shilatifard, 2016), c'est-à-dire du passage de l'information portée par le gène vers une structure mobile, l'acide ribonucléique messager (ARN messager). La stabilité apparente des chromosomes, et de l'ADN plus généralement, laissait penser que l'expression des gènes était un phénomène tout à fait déterministe mais l'étude des quantités d'ARN messagers ont conduit à l'idée que l'expression des gènes est en réalité un phénomène probabiliste. Les études initiatrices de cette théorie remontent à (Novick and Weiner, 1957). Il faudra attendre la fin des années 90 pour voir apparaître des études s'intéressant exclusivement aux variations de l'expression des gènes dans les populations cellulaires. Ces premières études se concentraient essentiellement sur les systèmes bactériens, plus faciles à manipuler mais surtout plus simples dans la structure de leur génome que les systèmes eucaryotes. En 2002, M. B. Elowitz (Elowitz et al., 2002) propose l'expérience suivante : l'étude de l'expression de deux gènes, codant pour des molécules fluorescentes différentes, insérés dans le génome de bactérie la *Escherichia coli* sous la dépendance de deux promoteurs identiques. Les hypothèses alors établies sont que si l'expression des gènes est parfaitement déterministe, les deux fluorochromes seront alors exprimés de la même manière et à une intensité déterminée dans toutes les bactéries de la culture expérimentale. M. B. Elowitz postule que si l'expression de ces gènes est soumise à du bruit extrinsèque, c'est-à-dire influencée par des événements extérieurs, l'intensité de l'expression des deux gènes évoluera avec le temps par exemple, mais gardera une expression d'intensité identique entre les deux gènes à un instant t. Leur expression sera alors parfaitement

corrélée. Dans un système probabiliste l'expression des deux gènes sujets à du bruit extrinsèque sera également soumis à du bruit intrinsèque, c'est-à-dire propre à chaque cellule et de nature aléatoire, induisant une expression non corrélée des deux gènes, et ce quel que soit le point d'observation. L'observation effectuée lors de l'étude est conforme à cette dernière hypothèse et Elowitz montre que l'expression des deux gènes est à la fois sujette aux bruits intrinsèque et extrinsèque. L'expression des gènes dans les systèmes procaryotes est alors définie comme étant probabiliste. Notons qu'à lui seul, le bruit extrinsèque peut engendrer une variabilité inter-individuelle pour peu que les échantillons soient soumis à un environnement sensiblement différent. Mais le bruit intrinsèque vient définitivement sceller l'apparition d'une variabilité inter-individuelle, les expressions de gènes étant soumises aux lois du hasard pour chaque cellule. Néanmoins, ce qui est valable chez les procaryotes ne l'est pas forcément chez les cellules eucaryotes.

Les premières études de l'aspect probabiliste de l'expression de gènes sur les levures (Blake et al., 2003; Raser and O'Shea, 2004) font état de la faiblesse des modèles procaryotes pour expliquer les phénomènes intervenant chez les eucaryotes. En revanche, elles ont mis en évidence l'adéquation de modèles d'expression par « *burst* », que l'on traduira ici par impulsion, où le gène est en constante transition entre des états d'activité d'expression et de non activité, un phénomène déjà observé chez les bactéries (Golding et al., 2005) mais existant avec une toute autre intensité chez les eucaryotes. On retrouve ce phénomène chez les eucaryotes supérieurs comme chez *Dictyostelium discoideum* (Chubb et al., 2006) où les impulsions sont moins violentes mais plus longues que dans les résultats décrits pour les bactéries. De manière intéressante, on constate aussi que ces variations impliquent une corrélation d'expression de gènes géographiquement proches sur le chromosome, attestant de l'hypothèse d'un contrôle de l'accès des gènes par la capacité de la chromatine à se contracter et se décontracter, via des facteurs spécifiques, permettant ou non l'accès des régions promotrices de la transcription aux complexes ribosomiques. Les périodes d'inactivité entre deux impulsions ont été associées à des phénomènes de modification de l'accès de la chromatine locale (Harper et al., 2011). Un autre phénomène permet d'expliquer l'impulsion, la transcription ne s'effectue pas de manière homogène mais par l'intermédiaire de régions spécifiques où se concentrent les éléments nécessaires à la transcription et où sont recrutés les gènes actifs. La compétition entre ces différentes régions pourrait alors expliquer en partie le bruit intrinsèque observé.

En 2002, E. M. Ozbudak démontre que, chez *Bacillus subtilis*, la variabilité de l'expression d'un gène est liée aux taux de transcription et de traduction (Ozbudak et al., 2002). En résumé, quand le taux de transcription est fort, le niveau de protéine est stable et peu variable. Mais quand le taux de transcription est faible et que le taux de traduction est fort, la variabilité dans la production de protéine est très forte. D'une manière générale, la variabilité de l'expression protéique est directement liée à certain nombre de caractéristiques du transcriptome comme l'écrit Warren (Warren et al., 2006) « *Messenger transcripts generally turn over much faster than the proteins they encode, which implies that protein expression may be buffered against stochastic fluctuations at the mRNA level* ». Les protéines dont la durée de vie est courte vont avoir tendance à suivre très fortement les niveaux d'expression des ARN messagers qui les codent, contrairement à celles qui ont une durée de vie plus longue qui verront leur accumulation rattraper le niveau d'ARN messagers.

L'analyse de la transmission de la variabilité d'expression des gènes le long d'une cascade transcriptionnelle montre la transmission de cette variabilité au sein des divers éléments qui composent la cascade. Les effets observés vont alors dans le sens d'une accentuation de la variabilité des éléments les plus bas dans la cascade en ajoutant du bruit au bruit qui leur est propre. Mais les modèles mathématiques prédisent aussi la possibilité que cette transmission du bruit vienne perturber cette variabilité propre au gène en ayant pour conséquence une diminution de celle-ci (Paulsson and Ehrenberg, 2000).

UN MOYEN DE CRÉER DE LA DIVERSITÉ

Imaginons un système biologique parfaitement immuable de génération en génération. Que se passe-t-il quand des changements environnementaux importants s'exercent sur le système ? La marge de manœuvre du système est contrainte par la structure intrinsèque du système. Cette rigidité ne pourrait en aucun cas faire face à des perturbations extérieures telles que nous pouvons les imaginer dans la nature (prédations, infections…). L'hypothèse est alors que la diversité permet de proposer un ensemble de réponses pour réagir à des phénomènes extérieurs.

Cela implique que le système, dont nous apprécions l'absence de diversité, évolue dans un système plus vaste qui lui fluctue. Imaginons maintenant que notre système immuable prospère dans un système tout aussi immuable que lui, alors il n'y aurait pas nécessité pour lui d'augmenter le champ

des possibles. Il faut alors considérer la variabilité comme une conséquence de la pression évolutive naturelle, une propriété qui aurait émergé pour devenir un outil de plus pour l'évolution.

Au-delà des aspects purement évolutifs considérés à l'échelle d'une population ou d'un règne entier, le hasard joue un rôle fondamental dans la génération de la diversité cellulaire au sein même de l'organisme. Un exemple flagrant est celui des cellules souches embryonnaires. La revue de deux chercheurs de l'université de Strasbourg, publiée en 2014, explique que ces cellules pluripotentes présentent une expression génique particulièrement diverse au sein d'une population cellulaire pourtant semblable phénotypiquement et évoluant dans un environnement stable (Torres-Padilla and Chambers, 2014). Les cellules souches embryonnaires sont des cellules dites pluripotentes, c'est-à-dire qu'elles peuvent être à l'origine de plusieurs types cellulaires en fonction des conditions dans lesquelles elles évoluent. Ce statut de cellules pluripotentes est contrôlé biologiquement par des régulations de l'expression de certains gènes qui participent à la stabilité du phénotype pluripotent. L'exemple du facteur de transcription NANOG, ayant un rôle connu dans la détermination du statut pluripotent des cellules, est particulièrement révélateur de l'existence d'une variabilité intrinsèque. Une première série d'expériences montrent que ce type de facteur de transcription a des niveaux d'expression très différents d'une cellule à une autre. Plus intéressant encore, ces niveaux ne sont pas stables dans la descendance. Une cellule n'exprimant pas NANOG, repiquée seule, donnera des cellules filles dont certaines exprimeront NANOG. L'atténuation expérimentale de l'hétérogénéité d'expression de NANOG, par l'inactivation du gène *nanog* originel et l'expression constitutive d'un gène *nanog* transgénique, montre une diminution de l'hétérogénéité globale des cellules pluripotentes. Cela montre bien le rôle important du niveau d'expression de ce facteur de transcription dans l'établissement et le maintien de phénotypes pluripotents particuliers. Plus encore, les travaux de Wu et Tzanakakis (Wu and Tzanakakis, 2013) ont proposé qu'une population de cellules souches embryonnaires conservait un équilibre entre des cellules exprimant NANOG d'un seul allèle, des cellules l'exprimant à partir des deux allèles et des cellules n'exprimant pas NANOG, faisant du contrôle de l'expression allélique de *nanog*, un acteur central de l'hétérogénéité de la population cellulaire. L'augmentation de l'expression de NANOG est par ailleurs corrélée avec une plus grande fréquence des événements d'impulsion et une plus grande probabilité d'expression via les deux allèles en même temps (Miyanari and Torres-Padilla, 2012). Enfin, sachant que NANOG contrôle sa propre expression par rétrocontrôle négatif, son expression est supportée par une série de phénomènes

aléatoires (le changement d'allèle et la fréquence des impulsions) pour maintenir une hétérogénéité globale dans la population de cellules souches embryonnaires. On se posera alors la question de l'intérêt d'une telle hétérogénéité et les auteurs nous orientent vers ce que je décrivais déjà plus haut en discutant des différentes tailles de mains : « *Heterogeneity of gene expression might have a functional role in cell fate decisions. This notion was put forward for in vivo embryonic development some years ago, and stipulated that stochastic changes in gene expression, including that of nanog, might allow a window of opportunity to direct lineage allocation* » (Torres-Padilla and Chambers, 2014). En sus de cette explication, cette hétérogénéité individuelle dans la population cellulaire engendre, presque de manière contre-intuitive, une plus grande stabilité du phénotype de la population globale et améliore les possibilités de réponse à des éléments extérieurs (Paszek et al., 2010).

Nous retrouvons cette diversité dans un modèle d'infection par le virus de l'hépatite C. Le travail que j'ai effectué il y a 10 ans avec R. Sobesky a montré qu'un patient atteint d'une infection par le virus de l'hépatite C voit la souche virale se différencier en colonies disséminées dans les nodules cirrhotiques du foie (Sobesky et al., 2007). S'y développent alors des quasi-espèces présentant des différences au niveau des protéines F et Core du virus. Il s'avère que la diversité et la complexité des variations des quasi-espèces est plus forte dans les nodules cancéreux que les nodules non cancéreux. Les quasi-espèces évoluent donc différemment selon les nodules où elles se situent et engendrent des variations qui favorise ou non l'apparition de cellules tumorales.

LE SYSTÈME IMMUNITAIRE

Le système immunitaire assure la surveillance et la garantie de l'intégrité de l'organisme qui exploite la variabilité pour accroître son efficacité. Dans le système immunitaire, la variabilité est présente à tous les niveaux d'échelle biologique. On peut, par exemple constater, la formidable diversité de cellules qui composent cet « organe de défense » (voir Encadré « Les cellules du système immunitaire » et Figure 3). Une diversité qui continue de dévoiler son ampleur avec les progrès technologiques, jusqu'à donner l'impression que finalement, il existerait presque autant de catégories de cellules immunitaire qu'il existe de cellules immunitaires.

Les cellules du système immunitaire. On distingue deux grands lignages cellulaires, les cellules de la lignée lymphoïde et les cellules de la lignée myéloïde, toutes deux issues d'une même origine cellulaire, les cellules souches hématopoïétiques (Figure 3).

La lignée lymphoïde produit plusieurs types cellulaires dont les lymphocytes T (LT), lymphocytes B (LB) et les cellules *Natural Killer* (NK)**.** Si leur production est issue d'un même précurseur présent dans la moelle osseuse, le lieu de leur maturation diffère (moelle osseuse pour les LB et les NK, le thymus pour les LT). Les cellules LT se divisent elles-mêmes en deux grandes classes en fonction de l'expression de marqueurs spécifiques : les LTCD4 et les LTCD8. Ces différentes populations possèdent des fonctions particulières :

- Production d'anticorps pour les LB
- Cytotoxique pour les LTCD8 et cellules NK, même si leur mécanisme est différent
- Stimulateur de la réponse humorale par sécrétion de cytokines pour les LTCD4

La lignée myéloïde produit un grand nombre de cellules différentes dont les cellules présentatrices d'antigène que sont les monocytes qui se différencient en macrophages lors de leur infiltration dans les tissus. Monocytes et macrophages ont la faculté de phagocyter et lyser des pathogènes ou cellules infectées. Les monocytes peuvent aussi se différentier en cellules dendritiques qui sont des activateurs de la réponse adaptative par la présentation des antigènes. Les granulocytes neutrophiles sont en général les premiers arrivés sur le site de l'infection. Ils possèdent eux aussi la capacité de phagocytose mais libèrent également des médiateurs qui participent à l'initiation de la résolution de l'inflammation locale (Levy et al., 2001; Sugimoto et al., 2016). Les granulocytes basophiles ont un rôle dans l'établissement de la réponse inflammatoire. Notamment, ils orientent la différenciation des LTCD4 en cellules dites Th2, qui sécrète alors des cytokines particulières (Otsuka et al., 2016). Les granulocytes éosinophiles ont un rôle pro-inflammatoire par la sécrétion de cytokines et détruisent les parasites sans phagocytose (Stone et al., 2010). La définition des cellules immunitaires ne saurait s'arrêter aux classes précédemment décrites tant elles sont riches d'une diversité caractérisée par l'expression de clusters de différentiation (CD) différents à leur surface. Sur la base de l'expression différentielle de ces CDs, il est possible de caractériser de nouvelles sous-populations (Passlick et al., 1989).

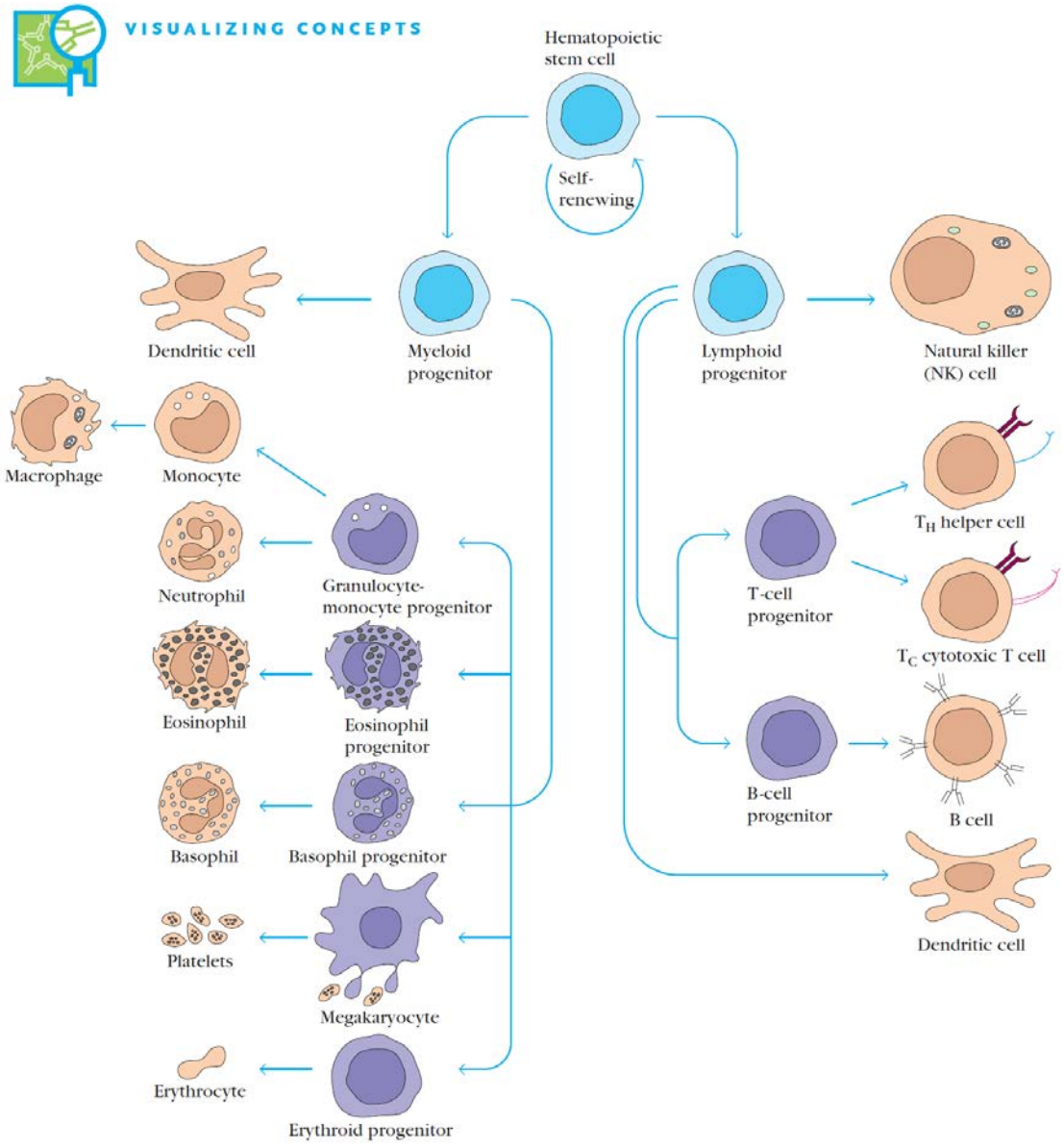


Figure 3. Les cellules du système immunitaire

Description schématique des grandes populations cellulaires du système immunitaire. Les deux lignées cellulaires (myéloïde et lymphoïde) dérivent de cellules souches hématopoïétiques. Chaque lignée se compose de différents types cellulaires ayant des caractéristiques morphologiques et fonctionnelles qui leur sont propres (source : Kindt, et al. Kuby immunology 6th edition, 2006. New York, W.H. Freeman).

Même sans considérer cet extrême, la variabilité au sein d'une population cellulaire que l'on décrit sur la base de ses caractéristiques structurales comme homogène, est très importante. On pourra se reporter aux écrits de (Mishalian et al., 2016; Sallusto, 2016) pour appréhender l'hétérogénéité des LTCD4, d'une part, et des cellules neutrophiles d'autre part.

Le système immunitaire est donc un système cellulaire complexe, mais aussi dynamique. Il est constitué de différents organes dont on distingue deux classes, ceux dits primaires (thymus et moelle osseuse) et ceux dits secondaires (rate, ganglions lymphatiques et les formations lymphoïdes associées aux muqueuses). Les organes lymphoïdes primaires sont les sites de développement et de maturation des cellules immunitaires. Les cellules immunocompétentes, ainsi produites, migrent ensuite vers les organes secondaires où elles participent à l'initiation de la réponse adaptative. On notera que les organes dits secondaires ne sont pas indispensables à la survie de l'organisme ; il arrive encore assez régulièrement de se faire retirer les végétations ou les amygdales en cas d'inflammations chroniques, ou de perdre la rate lors d'un choc violent à l'abdomen. Il se compose aussi d'une vaste collection de cellules, comme nous l'avons mentionné précédemment, ayant des phénotypes et des fonctions particulières. C'est donc un système multi-échelle (organe, cellules mais aussi molécules solubles) interagissant avec l'extérieur et l'intérieur de l'organisme pour maintenir son intégrité.

Afin de garantir cette fonction de protection, le système immunitaire possède un certain nombre de caractéristiques qui lui confère des avantages indéniables pour l'accomplissement de sa tâche. Il est tout d'abord dynamique, les cellules immunitaires, une fois produites, migrent vers les organes lymphoïdes secondaires en utilisant le réseau sanguin. Les lymphocytes entrent alors au contact des antigènes drainés par les réseaux lymphatique et sanguin, ou présentés par des cellules présentatrices d'antigène. Ces interactions, dites spécifiques d'antigène dans le cas des lymphocytes, car elle nécessite la présentation d'un antigène, induisent l'activation et la différentiation des cellules lymphocytaires et leur migration vers les sites effecteurs. Les réseaux fluidiques de l'organisme sont mis à contribution pour accélérer le transfert des cellules d'un site à un autre. Par la suite, les cellules immunitaires vont pouvoir s'infiltrer dans les tissus pour répondre au plus près à l'agression. Les organes lymphoïdes secondaires concentrent des quantités importantes de cellules qui sont mobilisées rapidement après l'agression par le jeu de migrations cellulaires depuis le site de l'agression vers les organes lymphoïdes, puis des organes lymphoïdes vers le site de l'agression. Cette dynamique est une composante de la résilience du système immunitaire (voir les définitions dans les encadrés « Résilience »). Une perturbation locale peut engendrer des répercutions à différentes échelles de l'organisme (organes, cellules et molécules solubles), donnant lieu à l'activation et la migration de populations cellulaires vers le site de la

perturbation. Ces caractéristiques ne pourraient pas exister sans un transfert d'information efficace, impliquant un système hautement dynamique.

Le système immunitaire est un système hautement adaptatif puis qu'il doit répondre de manière efficace à une infinité de stimuli. Dans ces conditions, il est probable que les acteurs de la réponse immunitaire soient eux-mêmes très variables pour répondre à ces défis. C'est ce qu'un article met en évidence en 2014. Ferraro *et al* a montré que l'expression des gènes liés au phénotype de lymphocytes T régulateurs était en moyenne plus variable que celle des gènes non liés à ce phénotype (Ferraro et al., 2014). Comme pour l'exemple des cellules souches cité plus haut, les gènes impliqués, qui participent au maintien du phénotype voient leur expression varier de manière importante. Cette variabilité d'expression des gènes induit une population de LTreg pouvant répondre à un stimulus de manière optimale.

Les lymphocytes T et B possèdent de surcroît un élément régit par des phénomènes aléatoires : le récepteur à l'antigène (respectivement TCR et BCR). Le TCR assure la reconnaissance spécifique de l'antigène. Chaque lymphocyte, qu'il soit B ou T, n'exprime qu'un seul type de récepteur en plusieurs exemplaires. Le récepteur est un hétérodimère constitué de deux chaînes (Figure 4).

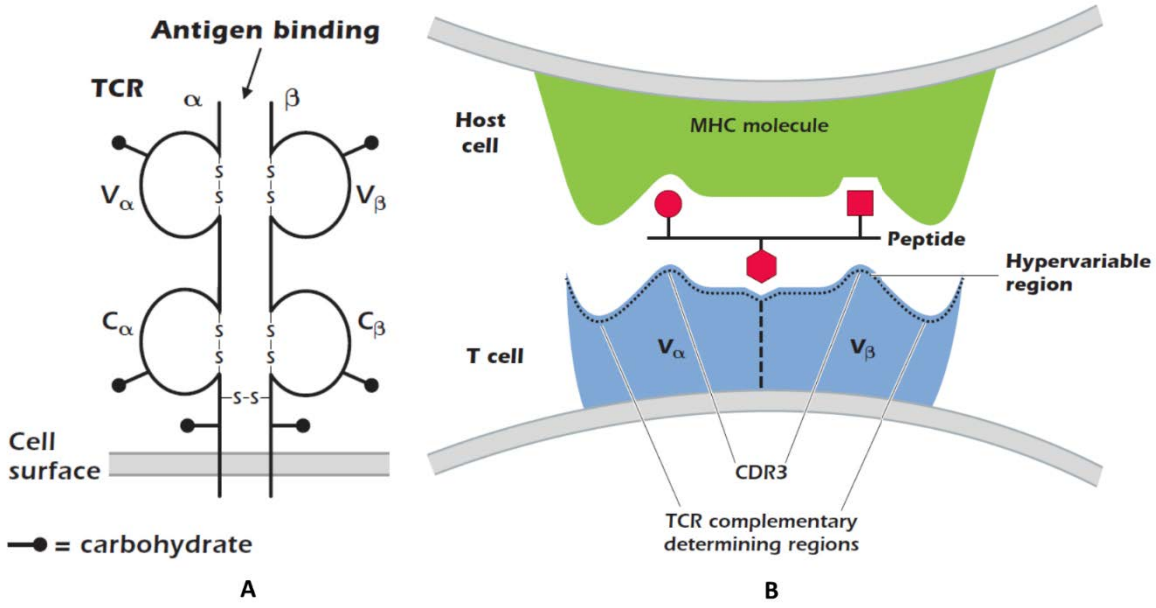


Figure 4. Les récepteurs des lymphocytes T.

A : Les TCR dits αβ sont constitués des chaînes α et β. Chacune des chaînes est elle-même constituée d'une région variable V, d'une région constante C, d'une région transmembranaire et d'une région intra-cytoplasmique. B :

Interaction entre le TCR et une molécule du complexe majeur d'histocompatibilité (MHC) présentant un antigène. Source : Coico and Sunshine, Immunology short course, 7th edition, 2015, Wiley Blackwell.

Chacune de ces chaînes possède une région dite variable et une autre dite constante et est codée par un gène unique. Ce gène est le fruit de réarrangements somatiques de plusieurs segments de gènes classés en famille V (variable), D (diversité), J (jonction) et C (constante). Au cours de l'ontogénie des lymphocytes, les segments de gènes vont se réarranger aléatoirement :

- Chaînes α et γ : recombinaison de deux gènes des familles V et J suivi de l'ajout d'un gène C.
- Chaînes β et δ : recombinaison de deux gènes des familles D et J, puis d'un gène V et enfin l'ajout du gène C.

Le produit de ces réarrangements géniques conduit notamment à la création d'une région appelée CDR3 (*complementarity-determining region 3*) qui contribue particulièrment à la reconnaissance de l'antigène (Figure 4). Le fait que les réarrangements sont aléatoires génère un répertoire de TCR différents (parmi 10^{18} TCR possibles), déterminant ainsi la capacité de l'organisme à reconnaître des antigènes. De la diversité naît l'efficacité d'une réponse en adéquation avec un stimulus, et donc l'adaptation au milieu et par conséquence une augmentation des chances de survie.

Résilience. La résilience est la capacité d'un système à retourner à un état d'équilibre après une perturbation. Ce statut n'est pas nécessairement identique à celui qui définissait le système avant la perturbation. Prenons l'exemple d'un écosystème, comme un étang, soudainement soumis à une pression extérieure comme un agent polluant. L'écosystème se voit alors fortement impacté par le polluant, modifiant profondément les relations entre les différentes espèces présentes par la disparition d'une espèce importante de la chaîne alimentaire par exemple. Si, avec le temps, le polluant disparaît, l'écosystème peut éventuellement se reconstruire autour des espèces qui ont survécu. Il redevient un écosystème fonctionnel sans être nécessairement identique au précédent. La résilience va donc caractériser la capacité d'un système à absorber le choc d'une perturbation.

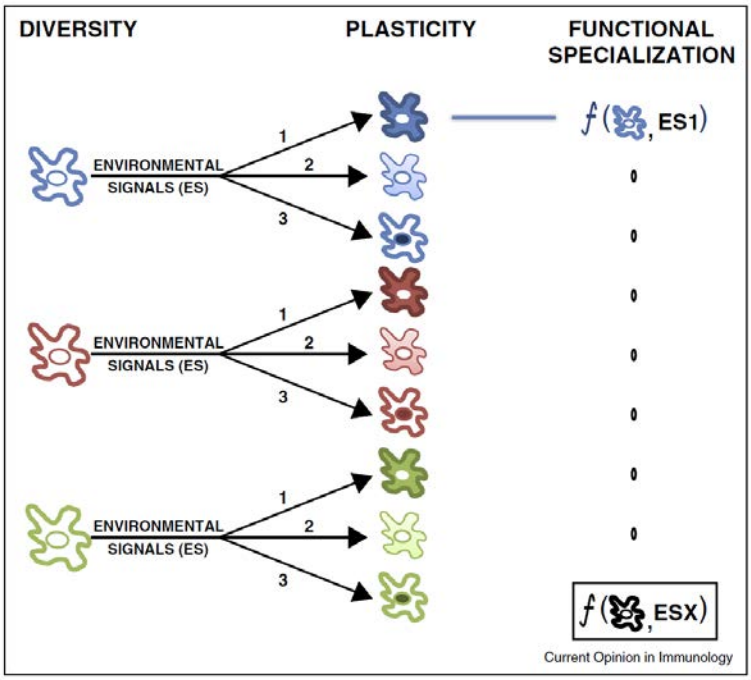


Figure 5. Diversité et plasticité cellulaire.

La diversité des cellules est amplifiée par la plasticité de ces dernières. Divers sous-ensembles (rouge, bleu et vert) d'une même population cellulaire répondent de manières particulières à des stimuli (1, 2 et 3). Le sous-ensemble cellulaire et la nature du stimulus conditionnent la spécialisation fonctionnelle. Source : (Soumelis et al., 2015).

Ces notions sont rappelées dans la publication de V. Soumelis *et al.* (Soumelis et al., 2015) où il étudie la diversité des cellules dendritiques (DC) et leur plasticité, c'est-à-dire leur capacité à changer de statut en fonction de la nature du stimulus. Combinées, la diversité et la plasticité, permettent d'engendrer un panel très important de réponse possible pour une population cellulaire données (Figure 5).

Une autre publication proposée par les mêmes auteurs (Cappuccio et al., 2015) référence par ailleurs les différentes possibilités de réponses d'un système à des stimuli. Dans leur modèle, les auteurs déterminent quatre-vingt-deux profils d'interaction entre deux stimuli. Ces profils sont résumés sous la forme de dix profils généraux, cinq dits positifs et cinq dits négatifs (Figure 6).

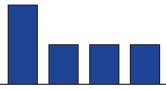
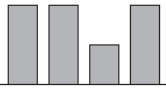
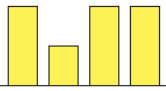
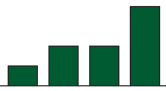
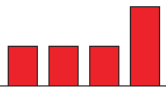
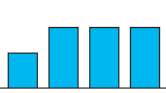
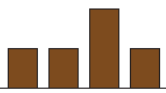
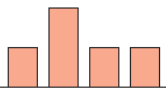
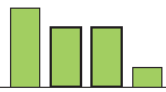
Interaction type	Interaction mode	Mathematical definitions	No.	Example profiles
Positive : $\Delta e_{X+Y} > \Delta e_X + \Delta e_Y$	Low stabilization	$e_\phi > \max(e_X, e_Y, e_{X+Y})$	13	
	X restores Y	$e_X \geq e_{X+Y}$ and $e_X > e_Y$ and $e_X \geq e_\phi$	6	
	Y restores X	$e_Y \geq e_{X+Y}$ and $e_Y > e_X$ and $e_Y \geq e_\phi$	6	
	Positive synergy	$e_{X+Y} \geq \max(e_\phi, e_X, e_Y)$ and $\Delta e_X > 0$ OR $\Delta e_Y > 0$	7	
	Emergent pos. synergy	$\text{sign}(\Delta e_{X+Y}) = +1$ and $\text{sign}(0) = +1$ and $\Delta e_X \leq 0$ and $\Delta e_Y \leq 0$	9	
Negative : $\Delta e_{X+Y} < \Delta e_X + \Delta e_Y$	High stabilization	$e_\phi < \min(e_X, e_Y, e_{X+Y})$	13	
	X inhibits Y	$e_x \leq e_{X+Y}$ and $e_X < e_Y$ and $e_X \leq e_\phi$	6	
	Y inhibits X	$e_Y \leq e_{X+Y}$ and $e_Y < e_X$ and $e_Y \leq e_\phi$	6	
	Negative synergy	$e_{X+Y} < \min(e_\phi, e_X, e_Y)$ and $\Delta e_x < 0$ OR $\Delta e_Y < 0$	7	
	Emergent neg. synergy	$\text{sign}(\Delta e_{X+Y}) = -1$ and $\text{sign}(0) = -1$ and $\Delta e_X \geq 0$ and $\Delta e_Y \geq 0$	9	
				Ø X Y X+Y

Figure 6. Modes d'interaction de deux stimuli.
Classification de 82 profils d'interaction en 10 modes d'interaction. Chaque mode comprend un ensemble de profils d'interaction qui répondent à des une interprétation biologique et mathématique particulière. Le nombre de profils inclus dans chaque mode est noté en colonne No. Source : (Cappuccio et al., 2015)

VARIABILITÉ INTER-INDIVIDUELLE

DE L'ÉVOLUTION À LA MÉDECINE PERSONNALISÉE

L'idée de variation dans le monde du vivant n'est pas nouvelle et des générations de biologistes se sont succédé pour analyser et comprendre ces variations. Elles furent longtemps étudiées pour comprendre les différences entre les espèces mais aussi entre individus d'une même espèce

géographiquement éloignés. Finalement, la variabilité inter-individuelle d'un groupe homogène d'individus ne fut étudiée que plus tardivement et il faudra attendre 1987 pour A.F. Bennet sonne l'alarme du peu d'intérêt porté par les sciences biologiques pour cette notion (Bennet, 1987).

Il publie, dans le livre *New direction in ecological physiology*, un article interpellant les chercheurs sur le manque de considération de la variabilité dans les analyses biologiques : *Interindividual variability : an underutilized resource*. Il y décrit comment la variabilité inter-individuelle prend racine dans la variabilité intra-individuelle. Si pour un paramètre donné, un individu varie d'un jour à l'autre, ces différences vont engendrer une variabilité au sein d'un groupe d'individus, car ceux-ci ne seront pas nécessairement en synchronisation parfaite. Par ailleurs, l'amplitude de la variabilité intra-individuelle ne sera pas la même pour chaque individu. L'exemple pris est une expérience consistant à étudier les capacités de course d'une espèce de lézard (*Sceloporus occidentalis*). Quinze lézards effectuent une course, cinq jours consécutifs, le résultat de leur classement est présenté dans la Figure 7.

Ceux-ci mettent en évidence une variabilité intra-individuelle d'une ampleur propre à chaque lézard. Globalement, l'ordre des individus est stable pour les cas extrêmes, les premiers et les derniers, moins pour les individus dans le milieu du classement. Deux idées émergent de ces résultats. Il existe des différences importantes entre les individus qui expliquent que certains soient plus rapides que d'autres. Ces différences sont stables : un individu qui court vite, court toujours vite ; un individu lent court toujours lentement. La seconde idée est que la mesure d'un paramètre ne représente qu'un ensemble de données aléatoires, dans un espace contraint, impossible à répéter du fait de la variabilité intra-individuelle. Néanmoins, Bennet insiste sur le fait que ce dernier point ne doit pas faire oublier que l'ordre des classements reste extrêmement stable à travers les jours (p-value < 0.001 selon le test de coefficient de concordance de Kendall). La variabilité intra-individuelle induit donc des variations à l'échelle de l'organisme qui expliquent l'importance de la variabilité inter-individuelle à l'échelle du groupe d'individus.

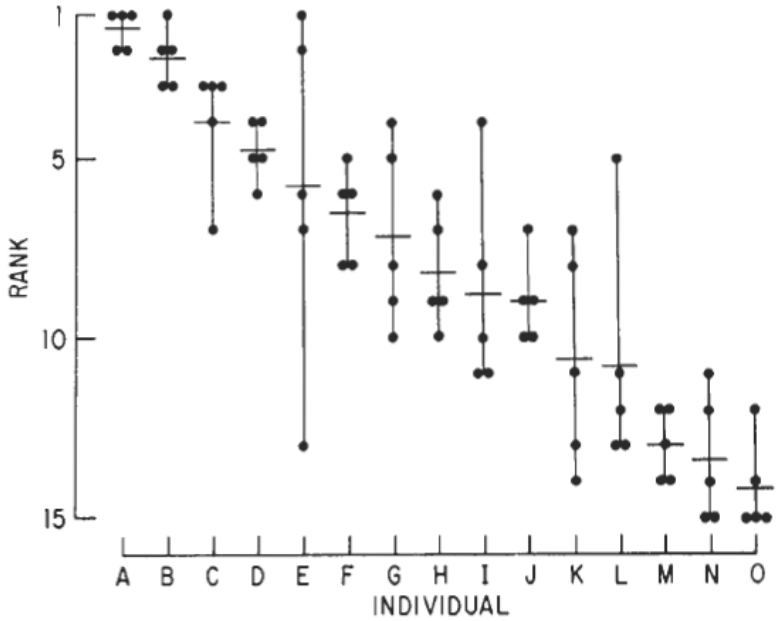


Figure 7. Variabilité intra-individuelle chez *Sceloporus occidentalis*.
Classement des performances de quinze lézards adultes de la famille *Sceloporus occidentalis* mesurées sur cinq jours consécutifs. Le rang 1 est l'animal le plus rapide, le rang 15 indique l'animal le plus lent. Les points indiquent les performances pour chaque jour ; les barres horizontales le rang moyen de chaque individu et les barres verticales l'écart entre les extremums. Les différents classements sont robustes ($p<0.001$ selon le test de coefficient de concordance de Kendall). Données de Bennet, 1980. Source : (Bennet, 1987).

L'un des arguments majeurs visant à minimiser la variabilité inter-individuelle dans les expériences consiste à dire que les points extrêmes ne sont pas représentatifs du groupe, ou qu'ils sont anormaux. C'est un point intéressant que critique Bennet en mettant en évidence la contradiction des chercheurs qui considéreraient que seuls les points autour de la moyenne sont vrais et de confiance. Ils devraient alors pondérer les mesures par leur distance à la moyenne et ne pourraient donc pas utiliser les tests classiques. Un cas extrême n'a en réalité d'extrême que le nom car il fait partie d'un continuum de possibilités pour l'échantillonnage observé. Reprenons un exemple basé sur la mesure de la vitesse et imaginons une expérience prenant pour cadre l'épreuve phare de l'athlétisme, le « cent mètres ». L'expérience consisterait à prendre de manière aléatoire dans la population cinquante individus et de mesurer leur performance sur cette distance. Il pourrait alors arriver que le choix inclue un coureur professionnel de « cent mètres ». La mesure de sa performance serait alors bien au-delà des mesures effectuées pour le reste de la population, son résultat serait-il pour autant une aberration ou une erreur ? Non. Ses capacités physiques lui autorisent à surpasser la majorité des individus mais elles n'ont rien d'aberrantes. Exclure ce

coureur, c'est enlever de l'information sur les capacités humaines en matière de « cent mètres » et donc fausser la mesure. Il est fort probable qu'avec un échantillonnage suffisamment grand, cette mesure extrême trouverait sa place dans un continuum de valeurs et ne semblerait alors plus aussi exceptionnelle.

Le deuxième argument avancé est que la variabilité inter-individuelle serait due en grande partie à la technologie qui induit des erreurs techniques qui viennent influencer la mesure biologique. La variabilité technique est un paramètre à prendre en compte. Son influence est limitée par les évolutions techniques et méthodologiques. Cependant, garantir un signal spécifique supérieur à un bruit de fond pour n'importe quelle mesure, reste un challenge.

Encore aujourd'hui, la majorité des articles de recherche ne raisonne que par la significativité de la moyenne, c'est ce que Bennet appelle « *the golden mean tyranny* ». Alors même que nous sommes tous conscients de l'existence de la variabilité inter-individuelle, nous avons tendance à la négliger ou à la considérer comme nuisible à la compréhension du système alors qu'elle est une caractéristique de ce dernier.

L'intérêt pour la variabilité inter-individuelle est très prononcé en pharmacologie. Et c'est notamment un article de pharmacologie qui va remettre l'article de Bennet au goût du jour. T. D. Williams le cite dès la première ligne pour souligner le faible engouement des spécialistes de son domaine de prédilection, le système endocrinien, pour la prise en compte de la variabilité inter-individuelle (Williams, 2008). Les doses sont prescrites par rapport à un patient moyen, l'écart-type n'étant bien souvent pas commenté. Or, comme l'écrivait déjà le médecin Sir William Osler en 1892, "*If it were not for the great variability among individuals medicine might as well be a science and not an art*". En d'autres termes, c'est bien la variabilité qui donne tout le piment à notre métier. Un article publié par F. Foissac montre bien l'effet de cette variabilité sur la prise de médicament (Foissac et al., 2011). Il y est question de traiter des enfants atteints du syndrome d'immunodéficience acquise (SIDA) par un cocktail de deux molécules, le Lopinavir et le Ritonavir. Le terme enfant est très intéressant ici car il englobe une large catégorie d'individus, dont le poids peut alors fortement varier. Or le volume de distribution et la clairance d'un médicament ne peut en aucun cas être identique chez un enfant faisant 8 kilos et un autre 50 kilos. Les données de cette étude ont été reprises pour illustrer l'intérêt des covariables dans l'analyse des données (http://www.rechercheclinique pariscentre.fr).

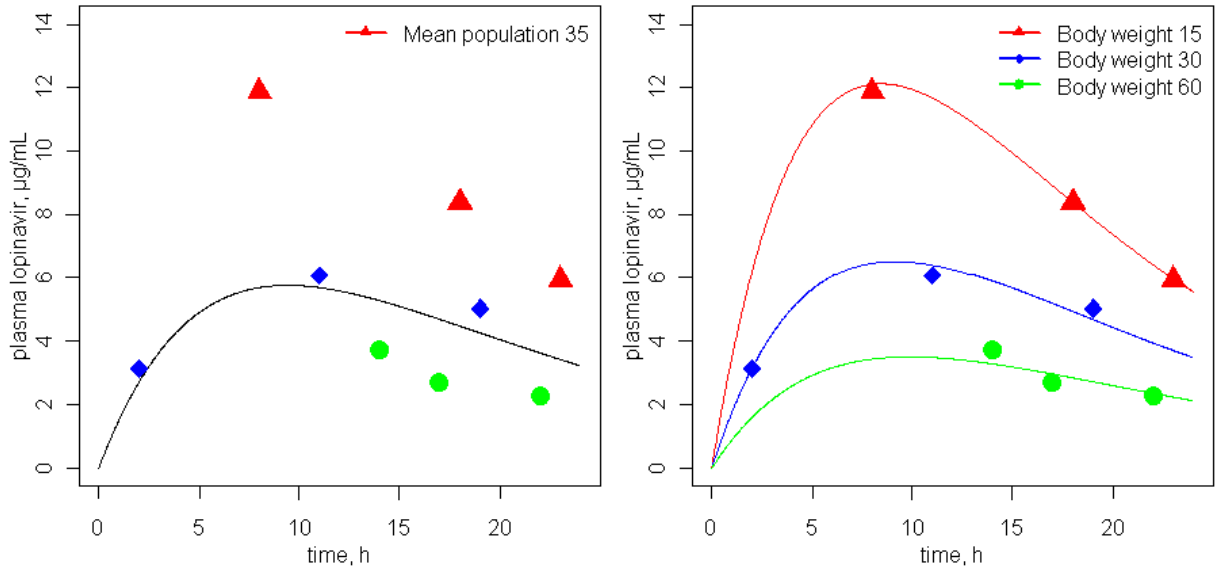


Figure 8. Clairance du Lopinavir chez l'enfant.

Gauche : Concentration de Lopinavir mesurée dans le plasma de trois enfants en fonction du temps. La courbe noire correspond à la courbe de régression non linéaire expliquant au mieux le comportement des échantillons. Droite : Concentration de Lopinavir mesurée dans le plasma d'enfant en fonction du temps. Les courbes de couleurs correspondent aux courbes de régression expliquant au mieux les données pour chacun des enfants. Le poids associé à chacun des enfants indique que plus un enfant est lourd, plus la clairance est rapide. Source : http://www.rechercheclinique pariscentre.fr d'après les données de (Foissac et al., 2011).

Les résultats de la Figure 8 montrent que la prise en compte d'un paramètre caractéristique d'une sous-population d'intérêt change grandement la vision de la posologie du médicament. Sous-estimer l'importance des variations entre individus dans l'administration d'un médicament est une erreur. La variabilité inter-individuelle peut nous renseigner sur l'état d'un système et des éléments qui le composent.

D'ailleurs, une branche de l'immunologie exploite parfaitement ces comportements, celle consistant à l'analyse du répertoire immunitaire. Le répertoire, des cellules LTCD4 par exemple, consiste en la collection des récepteurs des cellules LTCD4 (TCR) pour un individu donné à un instant t. La technique Immunoscope consiste à évaluer les perturbations entre les profils de répertoire de différents individus ou groupes d'individus en étudiant la distribution des longueurs d'une région hypervariable du TCR, la région CDR3. Les profils de longueurs de CDR3 obtenus sont « gaussiens » lorsque l'individu est en bonne santé. Une perturbation (infection, maladie congénitale) peut entraîner une modification de cette distribution impliquant des sur-représentations de certains pics, signifiant une probable expansion clonale. En regardant la

variabilité de la production de la région CDR3, il est possible de déterminer si le système est perturbé ou non (Figure 9).

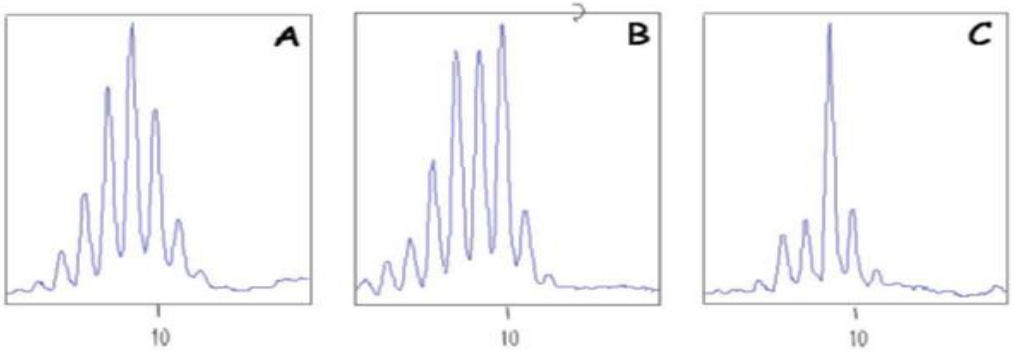


Figure 9. Spectratypes de CDR3.

Différents profils de distribution des longueurs de CDR3 par la méthode Immunoscope. A – profil classique avec distribution gaussienne caractérisant une réponse polyclonale. B – profil présentant plusieurs pics prédominants indiquant une expansion oligoclonale. C – profil avec un pic prédominant caractéristique d'une expansion monoclonale. En mesurant la différence entre les profils perturbés et le profil normal, sous la forme d'une distance, il est possible de calculer un score de perturbation. Source : d'après des résultats obtenus au sein du laboratoire dans le cadre du projet de neuro-paludisme expérimental chez la souris. Avec l'aimable accord de l'auteur, le Dr W. Chaara.

Restons sur le système immunitaire car il présente une caractéristique supplémentaire, en plus de celles citées dans les sections précédentes, générant de la variabilité inter-individuelle, son historicité (voir la définition dans l'encadré « Historicité »). Le système immunitaire possède la capacité de « garder en mémoire » le souvenir des agressions extérieures passées. Cette caractéristique permet au système d'être particulièrement efficace lors d'une prochaine rencontre avec un pathogène déjà rencontré et est par conséquent étroitement lié à la notion de répertoire vue précédemment. C'est le principe qui est exploité par la vaccination : une première injection fait connaître le pathogène à l'organisme qui, cherchant à l'éliminer, va sélectionner des cellules qui répondent particulièrement bien à ce type de pathogène. La finalité de la vaccination étant la création de pools de LT et LB mémoires, reconnaissant un antigène d'intérêt, rapidement mobilisables lors d'une réponse dite secondaire (comme c'est le cas pour les rappels de vaccin). C'est donc une faculté particulièrement intéressante pour la survie de l'organisme mais aussi particulièrement génératrice de variabilité entre les individus car l'histoire immunologique d'un individu est unique. L'exposition aux antigènes n'est absolument pas commune d'un individu à l'autre car elle ne dépend pas seulement de la localisation mais aussi du mode de vie des individus.

L'exemple de l'asthme rend bien compte de l'importance de l'histoire immunologique des individus. L'asthme est une maladie caractérisée par une gêne respiratoire résultante d'une inflammation excessive des bronches. En 2015 et 2016, deux études effectuées en Chine (Feng et al., 2016) et en Europe du nord (Timm et al., 2016), montrent une diminution de la prévalence de la maladie dans les milieux ruraux par rapport aux milieux urbains. La forte exposition à des panels d'antigènes complexes des enfants qui naissent et grandissent en milieu rural par rapport aux enfants des zones urbaines est un facteur avancé par les auteurs pour expliquer ce phénomène. Cependant, tout n'est pas aussi simple et, M. J. Ege montre par exemple que tous les environnements ruraux ne sont pas aussi protecteurs de l'asthme, voire que certains inversent la tendance (Ege et al., 2007). Quoiqu'il en soit, ces études montrent bien que l'environnement joue un rôle important dans l'histoire immunitaire des individus. Pour une même localisation des comportements sociaux différents induisent des expositions différentes à de vastes catégories d'antigènes, induisant par conséquent chez ces individus des réactions immunitaires diverses à un même stimulus ultérieur.

L'intérêt pour la variabilité inter-individuelle du système immunitaire chez l'homme à récemment pris la forme d'une étude à grande échelle (Thomas et al., 2015). Cette étude, nommée Milieu Intérieur, vise à analyser le système immunitaire de 1000 donneurs sains afin d'identifier les composantes expliquant les variations du système immunitaire chez l'homme. Les premiers résultats indiquent que des facteurs biologiques, comme le sexe et l'âge, mais aussi culturels, comme le tabagisme ou le statut relationnel, induisent des changements dans le statut immunitaire des donneurs.

Historicité. L'historicité est une caractéristique d'un système qui a la capacité de stocker les traces des événements antérieurs qui l'ont impacté. La structure du système est alors durablement modifiée en fonction de ces événements. L'avantage de cette caractéristique est d'apprendre des événements pour permettre une meilleure adaptation du système aux événements extérieurs.

VARIABILITÉ INTER-INDIVIDUELLE ET RÉPONSE

La relation entre stimulus et variabilité n'a pas donné lieu à de nombreuses recherches, néanmoins on peut noter qu'il existe un certain nombre de travaux qui décrivent comment un stimulus perturbe l'état stochastique de la production d'ARN dans des systèmes biologiques. Un des points les plus

intéressant développé par les auteurs d'une revue sur le sujet (Lehner and Kaneko, 2011) consiste en une théorie décrite par la formule mathématique suivante :

Formule 1 :

$$\frac{\langle x \rangle_{a+\Delta a} - \langle x \rangle_a}{\Delta a} \propto \langle (\delta x)^2 \rangle$$

La Formule 1 décrit qu'il existe une proportionnalité entre le changement de comportement de l'événement x suite à un changement environnemental a et la fluctuation initiale de l'événement x. Dans le cas qui nous intéresse, on pourra traduire cette formule par le fait qu'il existe une proportionnalité entre les changements d'expression d'un gène noté x soumis à un stimulus a et la variabilité de l'expression du gène x. En d'autres termes, plus l'expression d'un gène varie, plus la réponse attendue suite à un stimulus va être forte. Ce qui se traduit graphiquement par la représentation suivante fournie par B. Lehner et K. Kaneko (Figure 10).

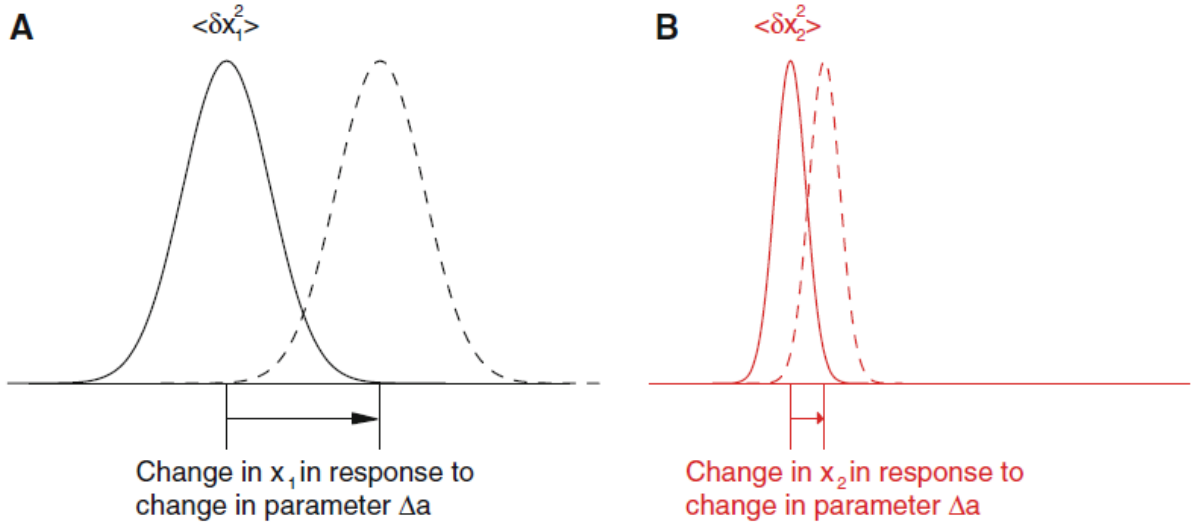


Figure 10. Relation entre fluctuation et réponse par K. Kaneko.
En réponse à un changement a, le gène x_1 possédant une forte variance $\langle (\delta x_1)^2 \rangle$ répond plus fortement à ce changement que le gène x_2, possédant une variance d'expression plus faible. L'intensité de la réponse est représentée sur l'axe des abscisses, la fréquence d'apparition de la réponse est donnée en ordonnée. Source : (Lehner and Kaneko, 2011).

Ainsi, des gènes exprimés avec une forte variance montrent une sensibilité accrue aux changements. B. Lehner et K. Kaneko mettent en évidence ce phénomène dans des expériences *in vivo*. Dans un modèle de levure la réponse de plusieurs gènes à un stimulus est évaluée et mise en relation avec la fluctuation intrinsèque de ces gènes (Figure 11). Les gènes possédant une forte

variabilité intrinsèque sont ceux qui répondent le plus fortement aux variations extérieures. Ces gènes sont aussi susceptibles d'évoluer plus rapidement que les autres (Landry et al., 2007). Les gènes à forte variabilité confèrent donc un avantage pour l'adaptation.

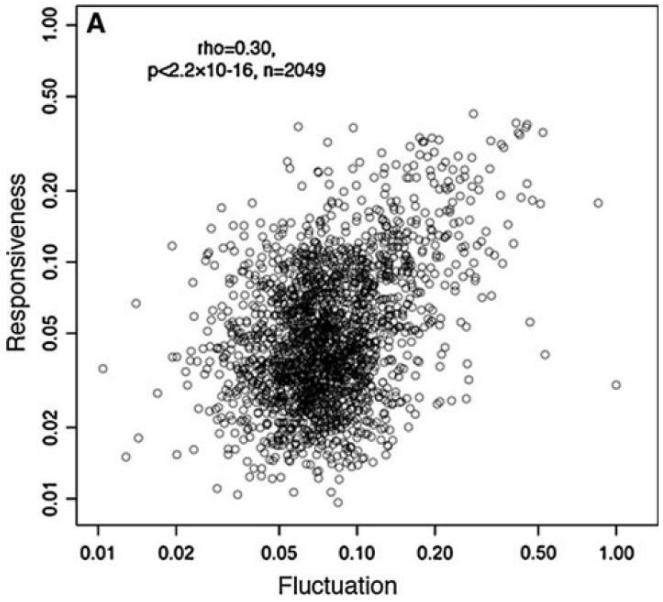


Figure 11. Relation fluctuation d'expression et réponse chez la levure.
Relation entre la fluctuation d'expression de 2049 gènes d'une levure et leur capacité à répondre à un stimulus. La fluctuation est calculée via la mesure décrite dans (Newman et al., 2006), la capacité de réponse par la différence d'expression après le stimulus. Source : (Lehner and Kaneko, 2011)

Le lien qui m'intéresse particulièrement est celui entre un stimulus et la variabilité inter-individuelle. Quelle que soit la variabilité d'expression des gènes, le stimulus va provoquer un changement d'état de leur expression (une augmentation ou une diminution). Selon la Figure 10, les fluctuations de l'expression des gènes avant et après le stimulus sont identiques. La variabilité inter-individuelle serait alors aussi grande avant et après stimulation. Deux expériences faites au laboratoire montrent pourtant qu'il pourrait exister une tendance à la diminution de la variabilité inter-individuelle après une stimulation. On trouvera une description complète des expériences dans la section « Origine des données » de ce manuscrit.

La première d'entre elle consistait à former deux groupes de souris femelles de 7-8 semaines de la lignée C57/BL6. L'un des groupes, composé de sept souris, reçoit une injection de solution saline (PBS), l'autre, composé de cinq souris, reçoit une solution de lipopolysaccharide (LPS), une endotoxine des bactéries gram négatives, souvent utilisée comme contrôle positif pour l'induction

d'une réponse immunitaire. Six heures après l'injection des produits, les rates sont prélevées, dilacérées et les cellules sont traitées pour en extraire les ARN. Le transcriptome est analysés sur des puce WG-6 de la compagnie Illumina tel que décrit dans (Pham et al., 2014) après soustraction du bruit de fond, normalisation par la méthode des quantiles et élimination des sondes communément non détectées dans les deux groupes.

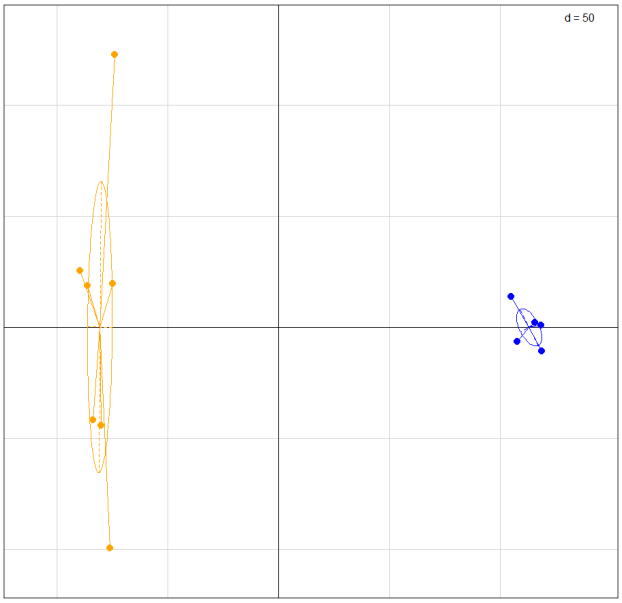


Figure 12. Analyse en composantes principales : Expérience 1.

Analyse en composantes principales (ACP) d'un jeu de données de transcriptome. Expérience : transcriptome de rate de souris ayant reçu une injection de solution saline (en jaune) ou de LPS (en bleu). L'axe des abscisses est le premier axe de l'ACP et représente 47.5% de la variance du jeu de données. L'axe des ordonnées est le second axe de l'ACP et représente 13.2% de la variance du jeu de données. Les ellipses représentent un résumé du nuage de points considérés. La longueur et la largeur de l'ellipse représentent 1,5 fois l'écart-type du groupe d'échantillons d'intérêt sur les axes présentés, le centre de l'ellipse étant le centre de gravité du nuage de points.

Le résultat d'une analyse en composantes principales (ACP) des données nous donne le résultat présenté en Figure 12. Cette méthode de réduction de dimensions, nous permet de projeter des données complexes sous une forme bidimensionnelle comme dans le cas de la figure 12. Les axes représentent des combinaisons linéaires de gènes participant à expliquer une partie de la variance observée à travers échantillons du jeu de données. L'axe des abscisses (axe 1) représente l'axe qui explique le plus de variance (47.5%). L'axe des ordonnées est l'axe perpendiculaire à l'axe 1 représentant le maximum de variance supplémentaire à l'axe 1 (13.2%). Il est possible d'avoir plus

d'axes mais ces deux premiers axes représentent déjà plus de 60% de la variabilité du jeu de données. La Figure 12 indique deux résultats :

- Les échantillons de souris ayant reçu l'injection de LPS sont très éloignés des souris qui ont reçu l'injection de solution saline. Cela est d'ailleurs représenté par l'axe 1, celui qui capte le plus de variance dans le jeu de données. Cela s'évalue par la taille de l'ellipse dont la longueur et la largeur correspondent à 1,5 fois l'écart-type du groupe d'échantillons sur les axes d'intérêts.
- Les échantillons de souris ayant reçu la solution saline montrent une variance plus importante sur l'axe 2 que les souris ayant reçu la solution saline.

La première constatation n'est pas surprenante car le LPS est connu pour induire une très forte réaction du système immunitaire via la voie du TLR4 ((Morris et al., 2014), Figure 13). Le deuxième axe nous intéresse davantage car il implique un changement dans la structuration du groupe de souris. La distance relative entre les individus sur ce type de représentation est directement liée à la ressemblance dans le comportement des individus à travers l'expression de leurs gènes. Les échantillons après la stimulation au LPS se répartissent de manière plus homogène que les échantillons soumis à la solution saline.

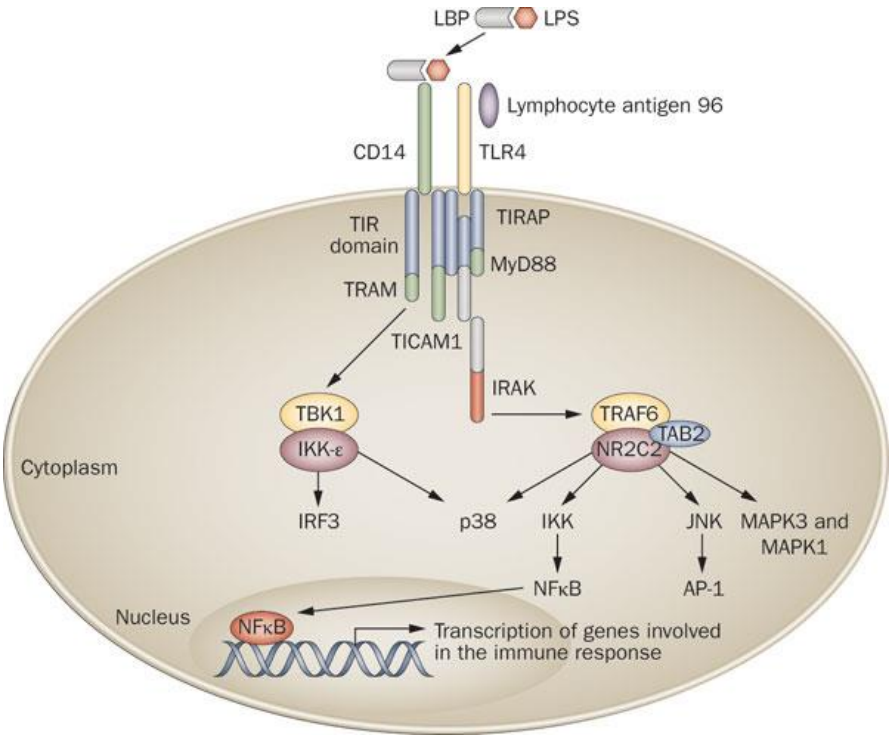


Figure 13. Cascade de réactions induites par le LPS.

Le lipopolysaccharide (LPS) induit, via la cascade de réactions du TLR4, la production d'interféron via la voie IRF3 et de cytokines pro-inflammatoires via la voie NFkB. Source : A. Abu-Shanab & E.M.M. Quiley, 2010.

Dans une autre expérience nous nous intéressons au transcriptome de l'environnement utérin lors de l'implantation du fœtus et sa gestation chez la souris. Pour cela, nous possédons des échantillons d'utérus de souris C57/BL6 (4 à 6 souris) à différentes étapes de gestation : 4, 6, 8, 10, 11 et 12 jours après le coït. Les échantillons servant de témoins négatifs comprennent des utérus de souris non fécondées. Les échantillons utérins sont broyés, traités pour en extraire les ARN et leur transcriptome est analysé sur des puce WG-6 fabriquées par Illumina tel que décrit dans (Nehar-Belaid et al., 2016). Le même procédé de représentation de l'expérience précédente est utilisé ici (Figure 14).

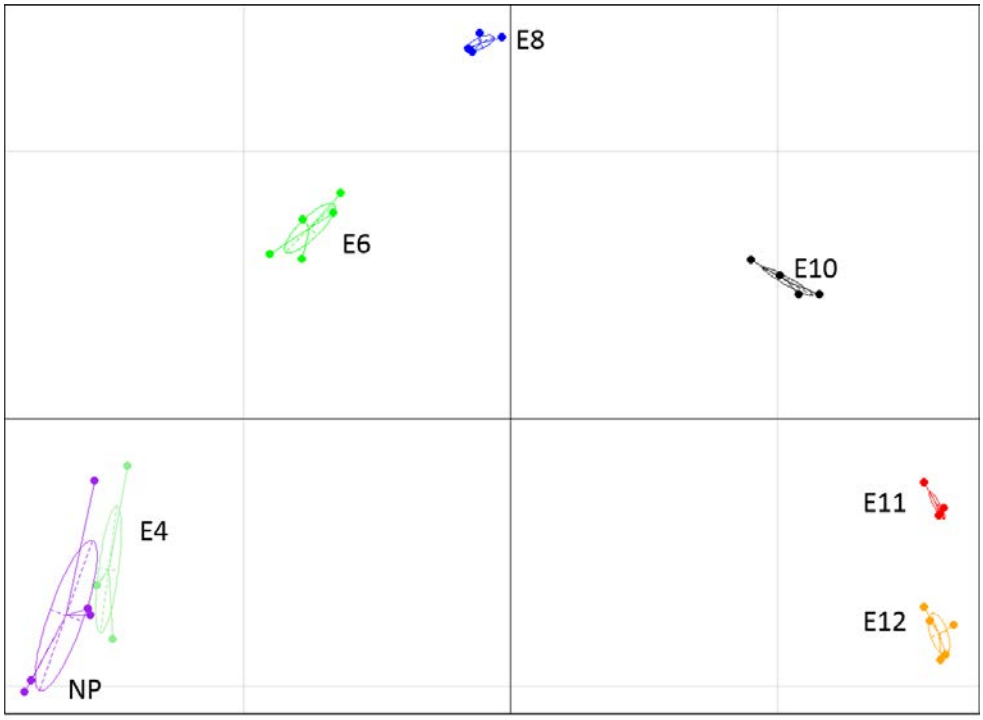


Figure 14. Analyse en composantes principales : Expérience 2.

Analyse en composantes principales (ACP) d'un jeu de données de transcriptome. Expérience : transcriptome d'utérus de souris sans fécondation (NP, en violet) et de souris 4 (vert foncé), 6 (vert clair), 8 (bleu), 10 (noir), 11 (rouge) et 12 (orange) après le coït. L'axe des abscisses est le premier axe de l'ACP et représente 49.4% de la variance du jeu de données. L'axe des ordonnées est le second axe de l'ACP et représente 18.4% de la variance du jeu de données. Les ellipses représentent un résumé du nuage de points considérés. La longueur et la largeur de l'ellipse représentent 1,5 fois l'écart-type du groupe d'échantillons d'intérêt sur les axes présentés, le centre de l'ellipse étant le centre de gravité du nuage de points.

L'ACP distingue parfaitement les différents temps de la cinétique et la taille de l'ellipse associée à chaque groupe rend compte de la distance entre les échantillons sur les deux premiers axes. Une fois encore, la dispersion des échantillons est plus forte chez les souris témoins que chez les souris fécondées. Elle diminue de manière dramatique à partir du jour 6. L'impression déjà perçue dans l'expérience LPS semble se confirmer ici, dans une expérience qui n'a rien de commun avec la précédente, en termes de type tissulaire et de nature de la stimulation. Nous étudions par la suite une troisième expérience qui consiste cette fois à analyser le transcriptome de LTreg après stimulation par l'IL-2, une molécule connue pour activer les LTreg. Les cellules sont issues du sang de 9 donneurs sains. Chaque échantillon est divisé en deux lots, l'un traité à l'IL-2, l'autre pas. Les différents lots sont alors traités pour en extraire les ARN et ces derniers sont analysés via des puces Gene ST1.0 de chez Affymetrix. Cette fois-ci l'ACP ne montre pas de large variation de

la dispersion des points entre les deux échantillons (Figure 15). Les données étant appariées par la construction de l'expérience, les points semblent simplement s'être décalés vers le haut de l'axe 2. En s'intéressant à d'autres axes, et notamment le deuxième (22,3%) combiné avec le troisième (10,6%), on constate une diminution de la taille de l'ellipse associée aux échantillons traités à l'IL-2.

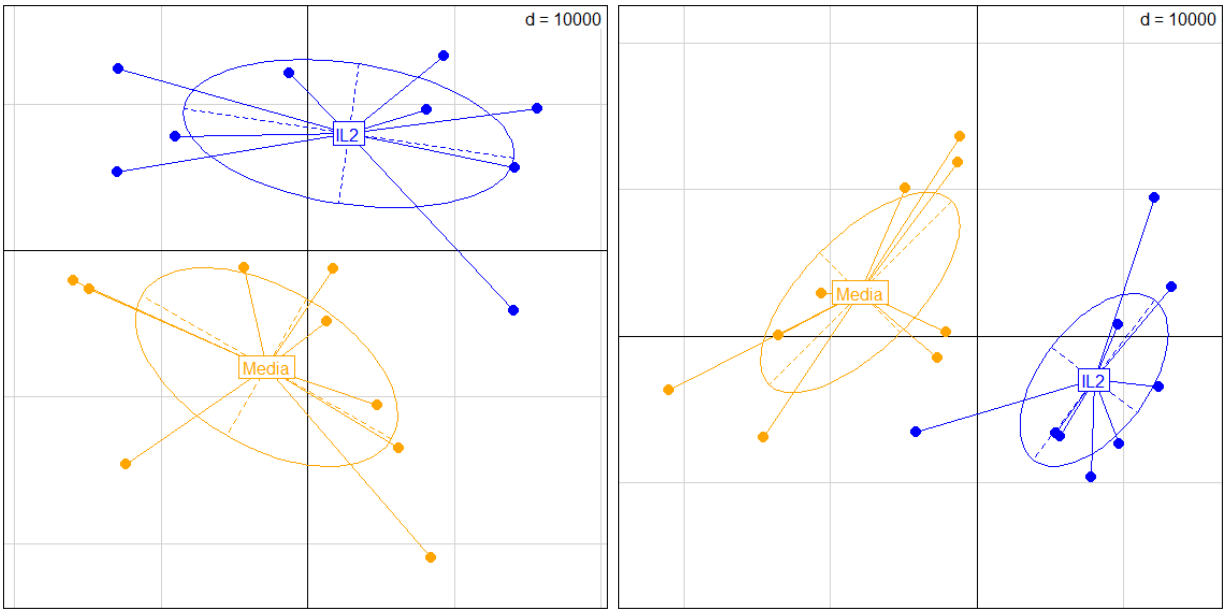


Figure 15. Analyse en composantes principales : Expérience 3.

Analyse en composantes principales (ACP) d'un jeu de données de transcriptome. Expérience : transcriptome de cellules LTreg triées du sang, de 9 donneurs sains. Les cellules sont séparées en deux lots, l'un sans traitement (en orange), l'autre traité avec de l'IL-2 (en bleu). Gauche : premier et second axe de l'ACP (23.9% et 22.3% de la variance du jeu de données). Droite : second et troisième axe de l'ACP (22.3% et 10.6% de la variance du jeu de données). Les ellipses représentent un résumé du nuage de points considérés. La longueur et la largeur de l'ellipse représentent 1,5 fois l'écart-type du groupe d'échantillons d'intérêt sur les axes présentés, le centre de l'ellipse étant le centre de gravité du nuage de points.

De la même manière, les résultats de l'expérience Milieu Intérieur, montrent qu'il existe un lien de cause à effet entre les perturbations du système immunitaire et des facteurs biologiques (âge, sexe) et socio-culturels (tabagisme, statut personnel). Cela indique que ces facteurs induisent chacun des changements similaires chez les donneurs, sans quoi il n'y aurait pas de corrélation. Les profils des donneurs répondant à ces critères sont alors plus proches les uns des autres qu'avec le reste de la cohorte (Thomas et al., 2015).

Mes compétences en bioinformatique, développées au cours de mes années au laboratoire, s'articulent essentiellement autour de l'analyse du transcriptome. C'est la raison pour laquelle les

derniers exemples développés dans ce chapitre, concernent des études du transcriptome. Cette thèse se concentre donc l'étude de cet objet biologique, dont on trouvera une description détaillée au chapitre suivant. Les expériences décrites précédemment suggèrent, au moins dans les systèmes expérimentaux étudiés, un lien entre la stimulation et la diminution de la variance au sein d'un groupe, à l'échelle du transcriptome. Ce phénomène nous a incités à approfondir l'analyse de cette variabilité inter-individuelle. Le projet de ma thèse a pour objectif de mesurer et représenter cette variabilité intra-groupes, mais aussi, de la comparer à travers différents groupes. L'idée initiale de ce projet va dans le sens d'apporter un regard nouveau sur les données de transcriptome afin d'enrichir le panel d'outils disponibles pour son analyse. L'originalité de ce projet s'établit alors dans l'utilisation des indices de diversité pour apprécier la variabilité inter-individuelle des données transcriptomiques, ce qui, à notre connaissance n'a jamais été publié à ce jour.

Je commence par la description des approches méthodologiques utilisé en routine au laboratoire afin d'expliciter notre intérêt initial pour la variabilité inter-individuelle et les méthodes connues que nous utilisons pour en tenir compte dans nos analyses. Je décris par la suite les indices de diversité et leur application au jeu de données de transcriptome. J'introduis enfin la notion d'indice de spécificité développée par O. Martinez (Martínez and Reyes-Valdés, 2008), son apport vis-à-vis de l'outil mathématique qu'est le coefficient de variation et enfin les résultats de son application aux données du transcriptome dans le cadre d'une analyse de la variabilité inter-individuelle.

APPROCHES MÉTHODOLOGIQUES

LE TRANSCRIPTOME

Les transcrits sont nombreux, plusieurs millions, divers en tailles (de 70 à 100 mille nucléotides) et en fonctions (messagers, de transfert, d'interférence, etc.). Le transcriptome correspond à la collection de ces différents transcrits dans un système biologique particulier (organe, tissu, population cellulaire, etc.) et à un instant t. Ils sont le produit de l'étape de transcription des gènes qui consiste à la copie du message génétique porté par un gène vers un produit mobile et support à la traduction, le transcrit (Figure 16). Le message est donc transcrit du langage du gène, l'ADN, en langage du transcrit, l'ARN. Leur analyse informe sur l'état d'activité des gènes dans le contexte observé.

La transcription chez les eucaryotes comporte plusieurs étapes. Tout d'abord, dans le noyau, la chromatine possède une conformation inactive, car compactée grâce à des complexes protéiques, les histones. Sous cette forme, la machine transcriptionnelle ne peut pas agir, la chromatine doit donc se trouver dans une conformation active, c'est-à-dire décompactée, pour permettre l'accès à l'information génétique. Cette étape de relaxation de la chromatine est celle qui explique les phénomènes de pulsations décrites dans l'introduction. La chromatine se relâche, libérant les gènes pour la transcription, puis se resserre, empêchant la transcription des gènes. Le temps nécessaire pour la compaction de la chromatine, mêlé à l'efficacité de la transcription, détermine en partie le niveau d'activation des gènes concernés.

Une fois la chromatine décompactée, un complexe protéique associant facteurs de transcription, facteurs de régulation et ARN polymérase se fixe sur l'ADN au niveau de la région dite promotrice du gène. Cette région contient des séquences nucléotidiques particulières (comme la boîte TATA) qui permettent la fixation des différents acteurs de la transcription. La fixation de l'ARN polymérase engendre l'ouverture de la double hélice d'ADN libérant ainsi les brins pour l'initiation de la transcription (Figure 16. La transcription.). L'étape suivante permet l'élongation du transcrit par polymérisations successives de nouveaux nucléotides correspondant à l'information fournit par le brin d'ADN (A→U, C→G, G→C et T→A). L'ARN polymérase sert alors de catalyseur pour

cette réaction chimique, permettant ainsi la polymérisation de 30 à 60 nucléotides à la seconde. La dernière étape correspond à la terminaison de la transcription et fait, là aussi, intervenir des séquences particulières dans la région terminale du gène ainsi que des facteurs protéiques de terminaison de la transcription. Le produit final de cette étape est un transcrit dit pré-messager car il va subir une étape de maturation conduisant à sa transformation en ARN messager et à sa délocalisation en dehors du noyau. Au cours de la maturation, le transcrit est coiffé d'un complexes protéiques dans sa partie 5' qui participe à la stabilisation du transcrit, sa migration en dehors du noyau et l'initiation de la traduction. La région 3', quant à elle, est prolongée d'une séquence polyadénylée qui participe à la stabilisation du transcrit et à sa dégradation. Avant d'être externalisé, le transcrit est nettoyé de ses séquences introniques (séquences se situant entre deux exons) au cours de l'épissage. Cette dernière étape se déroule sous le contrôle d'un complexe protéine/ARN, le spliceosome. L'épissage est par ailleurs l'étape qui permet de produire, à partir de transcrits pré-messager identique, des transcrits messagers différents engendrant la génération de protéines aux conformations diverses ayant potentiellement des fonctions différentes.

La quantification des transcrits messagers est alors directement corrélée à l'activité du gène mais non nécessairement à la quantité de protéine produite (Vogel and Marcotte, 2012). L'analyse transcriptomique renseigne donc plus sur l'état d'activité du système que sur un profil phénotypique précis, et cela correspond parfaitement à ce que nous souhaitons caractériser dans cette thèse. Les techniques pour quantifier les transcrits sont diverses mais reposent systématiquement sur le principe d'une amplification des transcrits associée à un marquage et une mesure de ce marquage. La première technique est la réaction de polymérisation en chaîne (PCR) consistant à utiliser les systèmes biologiques pour répliquer des fragments d'ADN *in vitro*. La version quantitative de cette technique permet de quantifier les brins produits après plusieurs cycles de réplication (voir Encadré « Réaction de polymérisation en chaîne »). Comparé aux résultats de séquences contrôles, elle donne une très bonne estimation de la quantité de brin d'ARN dans l'échantillon initial. Les puces à ADN profitent de cette méthode d'amplification du signale pour la quantifier non plus pour un brin d'ADN mais plusieurs dizaines de millier en même temps (voir Encadré « Les puces à ADN »). Une autre étape en encore franchie par le séquençage haut débit qui fournit des informations (quantification et séquence) sur plusieurs millions de transcrits en même temps.

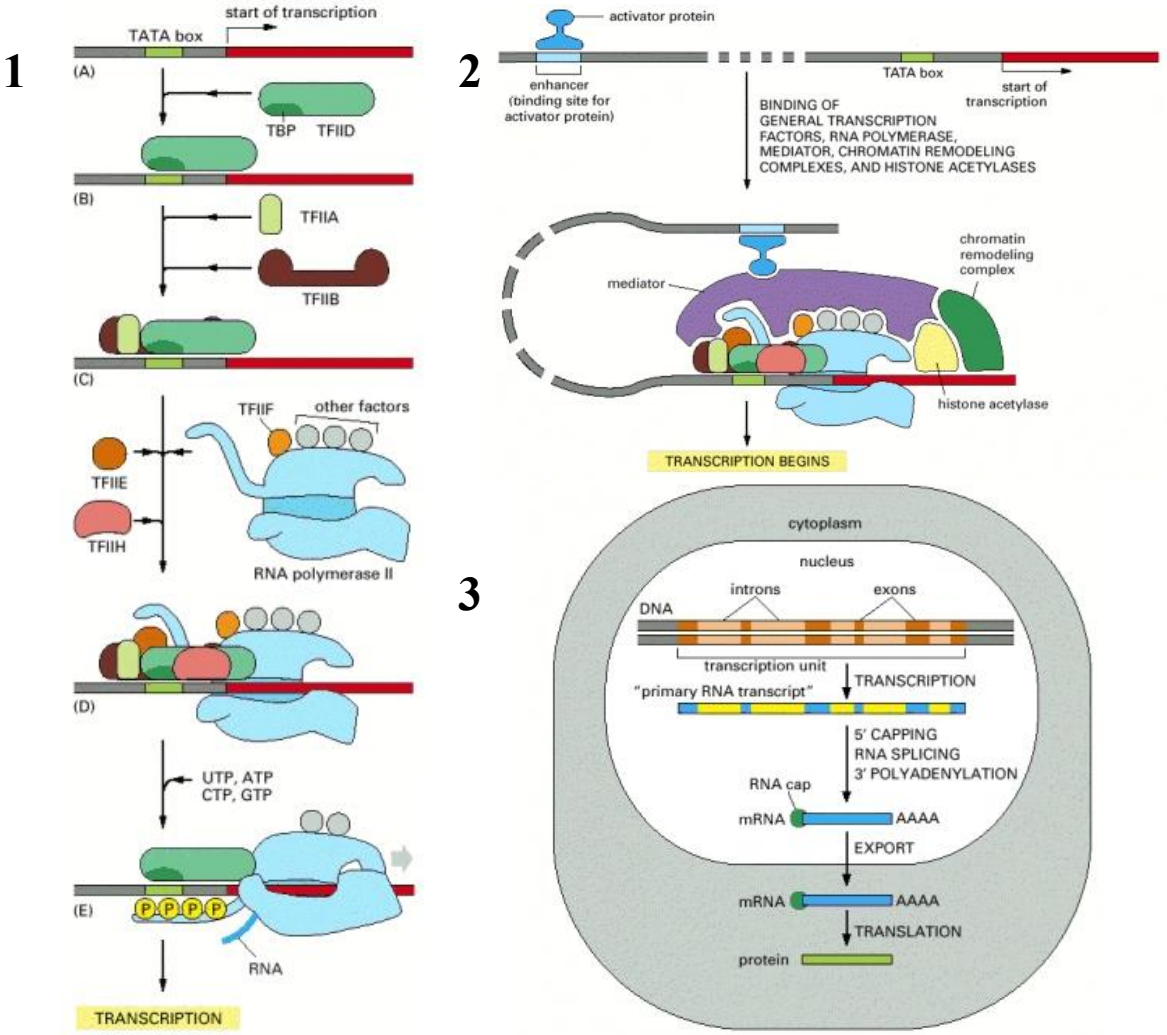


Figure 16. La transcription.

Le complexe protéique nécessaire à la transcription se compose de protéines qui se fixent sur la région promotrice du gène (1), de l'ARN polymérase (1) et de protéines qui se fixent sur des régions activatrices de l'ADN (2). Une fois produit, le brin d'ARN subit une phase de maturation dans le noyau. Cette étape s'effectue tout d'abord par l'ajout de protéines chaperonnes de part et d'autre du brin. Une séquence AAUAAA est reconnue en 3' du brin par diverses protéines qui s'y fixent. Le brin est aussi coupé en région 3' par des facteurs spécifique. En 5' et 3', des nucléotides sont ajoutés pour protéger le brin d'ARN et permettre sa délocalisation vers le cytoplasme. L'étape de maturation suivante consiste en l'élimination des parties non codantes, les introns. La coiffe en 5' composée d'une seule guanine est reconnue par les pores du noyau permettant l'export du brin d'ARN vers le milieu cytoplasmique (3). Source : Bruce Alberts, et al. Molecular Biology of the Cell. 4th edition 2002, Garland Science.

La réaction de polymérisation en chaîne (PCR). Publiée pour la première fois au milieu des années 80 (Mullis and Faloona, 1987), la technique PCR à révolutionnée la biologie moléculaire. Le principe est la génération de copies de fragments d'ADN par l'utilisation des complexes biologique dédiés à cette étape dans les cellules, on parle d'amplification. La préparation initiale consiste en une solution contenant les brins d'ADN à répliquer, des nucléotides et le système de réplication de l'ADN, l'ADN polymérase. Les différentes étapes de la réplication se font par le contrôle de la température de la préparation. Forte au début (95°C), elle permet la séparation de l'ADN double brin en deux fragments simple brin. S'ensuit une phase d'hybridation des amorces spécifiques des fragments que l'on souhaite amplifier (température : aux alentours de 60°C). L'étape d'élongation se produit à 72°C ce qui permet aux polymérases de synthétiser le nouveau brin d'ADN, complémentaire du fragment d'origine. Le retour à la température de 95°C permet de relancer un nouveau cycle de synthèse. En théorie, après vingt cycles de PCR une séquence d'ADN double brin génère 2^{20} copies. En pratique le nombre est moindre car la technique n'a pas un taux de réussite de 100%. Cette technologie n'a pas d'intérêt particulier pour l'analyse du transcriptome car elle est confinée à l'analyse d'un faible nombre de gènes à la fois. Toutefois, elle est encore indispensable pour confirmer des variations vues à l'aide de technologies dite haut-débits car elle est plus sensible que ces dernières.

Les puces à ADN. Les puces à ADN sont une technique d'analyse de fragments nucléotidiques utilisée en génomique pour l'analyse des polymorphismes nucléotidiques, de la méthylation et en transcriptomique pour la semi-quantification des ARN et l'étude des variations d'épissage. Le dispositif général est constitué d'un support (en verre, silicium ou plastique) sur lequel sont collées plusieurs dizaines de milliers de courtes séquences nucléotidiques d'intérêts appelées sondes (25-60 mers). Ces sondes, uniques mais représentées en plusieurs copies sur le support, sont, par exemple, des séquences spécifiques d'un ARNm, d'une région particulière de cet ARNm, ou encore d'une région du génome.

D'un point de vue industriel, la mise en place des sondes se fait par photolithographie (Affymetrix), par impression *in situ* (Agilent) ou encore par l'utilisation de billes (Illumina). Dans tous les cas, le principe reste l'agglomération de plusieurs copies d'une même sonde dans un espace restreint.

Le matériel biologique (ADN ou ARN) est extrait, traité (les ARN sont généralement transformés en ADN complémentaires, beaucoup plus stables et amplifiables par des techniques dérivées de la PCR). Lors de cette dernière étape, les fragments produits sont marqués par des molécules fluorescentes. Les fragments préparés sont ensuite incubés sur la plaque en verre afin que les fragments s'accrochent sur leurs sondes spécifiques par hybridation. L'excitation des molécules fluorescentes donne une image de l'intensité de fluorescence pour chaque groupement de sondes et donc de la présence/absence d'hybridation (Figure 17). Le système n'étant pas capable de dénombrer précisément les fragments hybridés sur la puce, la quantification ne peut être que semi-quantitative et la valeur d'intensité d'une sonde n'a de sens que si elle est mise en perspective des autres valeurs de la puce. Le fichier de sortie consiste en une matrice contenant les valeurs d'expression déduites de chaque gène pour l'échantillon donné. La grande flexibilité des puces à ADN permet d'analyser l'activité de l'ensemble des gènes connus et de discriminer un grand nombre de variations d'épissage.

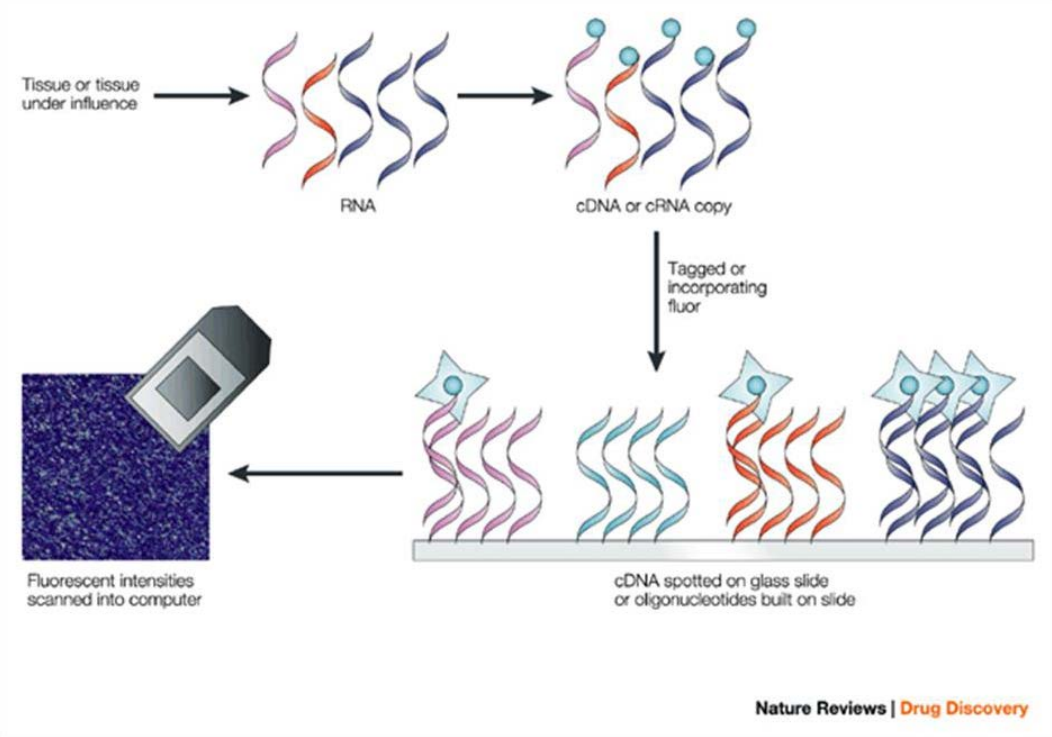


Figure 17. Les puces à ADN.

Principes des puces à ADN. L'ARN est extrait d'un tissu puis amplifié et marqué. Les fragments vont s'hybrider avec des séquences nucléotidiques attachées sur une plaque de verre. La mesure de la fluorescence de chaque groupe de séquences donne une estimation de la quantité de fragments hybridés. Source : A. Butte, 2002.

La grande différence entre les puces à ADN et le séquençage (RNAseq) est la quantité d'information obtenue. Les puces vont fournir des données sur des transcrits connus comme les ARN messagers, intéressant car ils servent de support à la traduction pour la production de protéine. Le RNAseq va fournir des informations sur l'ensemble du transcriptome incluant donc les ARN non codants dont les rôles sont très divers (aide à la transcription [tRNA], régulation transcriptomique [miRNA], maturation des transcrits [snRNA], etc.). Cette technologie a donc ouvert un nouveau champ d'investigation dans la compréhension du système et fournit encore aujourd'hui de nouvelles découvertes grâce à l'amélioration des technologies et de la profondeur de séquençage (voir Encadré « Le séquençage massif » et Figure 18). Le choix entre les deux technologies est aujourd'hui encore grandement impacté par la différence de coût, Le RNAseq coûtant environ deux fois et demie plus cher que les puces à ADN.

Le séquençage massif. Le séquençage massif est une technique d'analyse de fragments nucléotidiques utilisée en génomique pour l'analyse des polymorphismes nucléotidiques, de la méthylation et la reconstruction génomique, et en transcriptomique pour la quantification des ARN et l'étude des variations d'épissage. Le matériel biologique (ADN ou ARN) est extrait, traité (fragmentation, retro-transcription si nécessaire, amplification), les fragments se voient alors flanqués d'adaptateurs permettant le séquençage. Lors de cette dernière étape, chaque cycle de séquençage incorpore un nucléotide marqué, le système est lavé et scanné pour la détection du nucléotide pour chaque séquence. Le fichier en sortie se traduit sous la forme de fichier au format FASTQ, contenant la liste des fragments séquencés associés aux scores de qualité des nucléotides individuels de chaque fragment. L'abondance d'un transcrit se traduit par le nombre de séquences produites pour celui-ci. En effet, plus un transcrit est abondant dans la solution initiale, plus, en théorie, il sera abondant dans l'échantillon amplifié. Par définition, le séquençage massif ne cible pas de séquence particulière et permet donc en théorie de fournir le comptage de l'ensemble des transcrit d'un échantillon. La profondeur de séquençage, correspondant à la densité de séquence produite pour chaque kilobase, joue un rôle important dans la qualité des données, plus la profondeur sera forte, plus les transcrits rares auront des chances d'être séquencés.

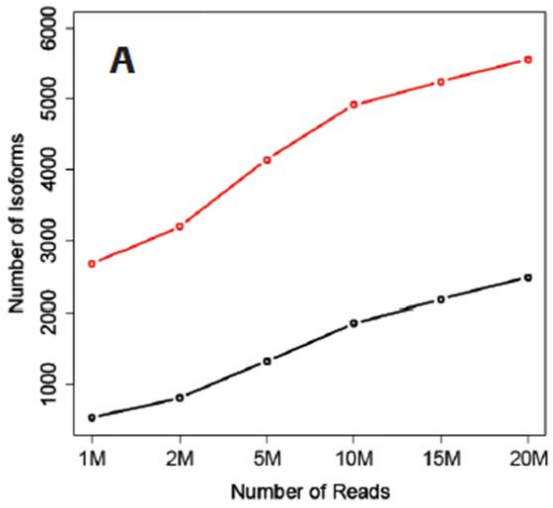


Figure 18. Impact de la profondeur de séquençage et de la technologie en RNAseq.
Nombre d'isoformes de transcrit trouvés lors de séquençages haut-débits en fonction du nombre de séquences produites (reads) et en fonction de la technologie utilisée (Roche SeqCap RNA en rouge, autre en noir). (Source : Roche®). Source : Roche®.

ORIGINES DES DONNÉES

Deux aspects me sont apparus très importants dans le choix des données à utiliser pour l'objectif de ma thèse :

- Les données doivent être de qualité. Nous avons vu précédemment qu'il existe potentiellement une grande part de variance dans les jeux de données de transcriptome qui s'explique par du bruit technique.
- Les jeux de données doivent inclure un groupe d'échantillons sous l'effet d'une stimulation et un groupe d'échantillons témoins, c'est-à-dire sans stimulation ou sous l'effet d'une stimulation permettant d'éliminer un biais de manipulation, comme l'effet d'une injection par exemple.

J'ai choisi des jeux de données produits au laboratoire, répondant à ces critères et faisant par ailleurs référence à des contextes biologiques très différents. Ces jeux de données sont détaillés dans les sections suivantes.

LPS

Actuellement, les vaccins sont les outils les plus efficaces pour prévenir les maladies infectieuses : ils ont un rôle prophylactique. La vaccination fait entrer en jeu les systèmes immunitaires inné et adaptatif. Brièvement, une vaccination induit une réponse dite primaire, au sens que l'organisme est confronté à l'antigène pour la première fois. Les cellules dendritiques (DC) sont les premiers acteurs immunitaires à agir après une vaccination par leur rôle de cellules présentatrices d'antigène. Présentant l'antigène à leur surface, elles migrent vers les ganglions lymphatiques où elles présentent l'antigène aux LTCD4 et aux LTCD8 naïfs. C'est alors l'étape de réponse adaptative qui voit les LT proliférer et se différencier en LT effecteurs et LT mémoires. Les LTCD4 vont soutenir les LB dans leur réponse anticorps (réponse immunitaire humorale) et les LTCD8 dans leur différentiation en LT cytotoxiques (réponse immunitaire cellulaire).

Lors d'une réponse secondaire liée à une infection par le pathogène ciblé par le vaccin, la réponse humorale conduit à la production d'anticorps par les LB. Les anticorps vont se fixer de manière spécifique à la surface du pathogène et le neutraliser. Ils participent aussi à la destruction du pathogène par l'intermédiaire du complément et de la réponse cytotoxique dépendante des anticorps. La réponse immunitaire cellulaire va, quant à elle, cibler les cellules infectées par le pathogène en sécrétant des cytokines antivirales et en ayant une action cytolytique. La plupart des vaccins activent la réponse humorale et cellulaire mais la variété de structure des vecteurs induit des différences dans l'ampleur de ces activations. Ces activations sont aussi très influencées par le mode d'administration du vaccin (Ellis et al., 2016) et par l'histoire immunologique de l'individu (voir Introduction).

Pour être efficace, un vaccin doit répondre à des exigences biologiques et structurelles particulières (Six et al., 2012):

- Activation efficace des cellules présentatrices d'antigène
- Activation importante des LT et des LB pour induire une grande quantité de lymphocytes avec un phénotype mémoire
- Une activation des LB (réponse humorale)
- Une activation des LT (réponse cellulaire)
- La production de LT mémoires pour des épitopes différents du pathogène pour permettre une meilleure adaptation de la réponse

- La persistance d'une mémoire immunitaire sur le long-terme

Le projet CompuVac fut alors développé pour proposer un cadre standardisé de développement et d'évaluation de nouvelles plates-formes de vecteurs. Ces vecteurs sont utilisés comme des véhicules exprimant, ou exposant, à leur surface des antigènes d'intérêts. Ils agissent aussi comme des adjuvants, c'est-à-dire des catalyseurs, de la réponse immunitaire. La standardisation du programme CompuVac passe par l'utilisation de l'antigène de référence gp33-41. Il s'agit d'un peptide dérivé de la glycoprotéine (gp) du virus LCMV (virus de la chorioméningite lymphocytaire), composé des résidus 33 à 41. Cette molécule est reconnue par les LTCD8 via le complexe majeur d'histocompatibilité de classe I.

Le but est de déterminer si un vecteur possède les qualités requises pour une vaccination : déclencher une réaction immunitaire et délivrer le message antigénique. Au cours du programme, une cinquantaine de vecteurs a été testée ; ces vecteurs regroupent en quatre grandes classes :

- Les vecteurs viraux
- Les vecteurs bactériens
- Les vecteurs ADN
- Les *virus-like particules* (VLP)

Ces vecteurs ont été testés chez la souris pour un certain nombre de paramètres :

- Expansion clonale de LT antigène-spécifiques
- Production d'interféron gamma
- Le jour de survenue du pic de réponse clonale
- La présence du phénotype mémoire
- L'activité cytotoxique
- La production d'anticorps
- L'activité neutralisante de ces anticorps

Ces paramètres fournissent des renseignements précieux sur le potentiel de telle ou telle construction (Figure 19). La plate-forme se veut alors un outil d'aide à la décision dans l'orientation du développement des vecteurs mais aussi dans le choix des vecteurs à utiliser dans une condition particulière (vaccination ou thérapie génique) (Blazewicz et al., 2012).

C'est dans ce contexte que le laboratoire a produit un grand nombre d'expériences visant à utiliser les modifications du transcriptome chez des souris quelques heures seulement après la vaccination comme prédicteur de l'efficacité d'une vaccination (Dérian et al., 2016). Sur la base du

transcriptome des cellules dendritiques triées de la rate de souris vaccinées seulement six heures auparavant, nous avons construit un modèle prédictif de l'efficacité de la réponse immunitaire. Notre modèle est par ailleurs suffisamment sensible pour prédire des jeux de données provenant de familles de vecteurs qui lui sont inconnus, mais aussi de jeux de données de transcriptome de rate totale ou de PBMC. Enfin, il prédit avec une grande efficacité des jeux de données d'origine humaine (Zak et al., 2012). Nous démontrons ainsi toute la pertinence de ce type d'approche pour la prédiction de l'efficacité de vecteurs vaccinaux.

Au cours de ce projet nous avions aussi produit un jeu de données composés de souris auxquelles nous avions injecté du LPS ; c'est le jeu de données que j'ai utilisé dans ma thèse. Il est composé de douze souris femelles de sept semaines de la lignée C57BL/6. Sept souris ont reçu une injection de solution saline (PBS), les cinq autres une solution de LPS. Il s'agit d'une expérience servant de contrôle positif car le LPS, qui n'est autre que le composant principal de la membrane externe de certaines bactéries, dites à gram négatif, induit une réaction du système immunitaire (voir Figure 13). Le transcriptome est produit à partir de cellules de la rate dilacérée des souris six heures seulement après l'injection du produit. La préparation des ARN et des puces à ADN Illumina sont similaires aux protocoles décris dans le papier Dérian *et al.* en Annexe #1. Par la suite, je ferai référence à ces données comme le jeu de données LPS.

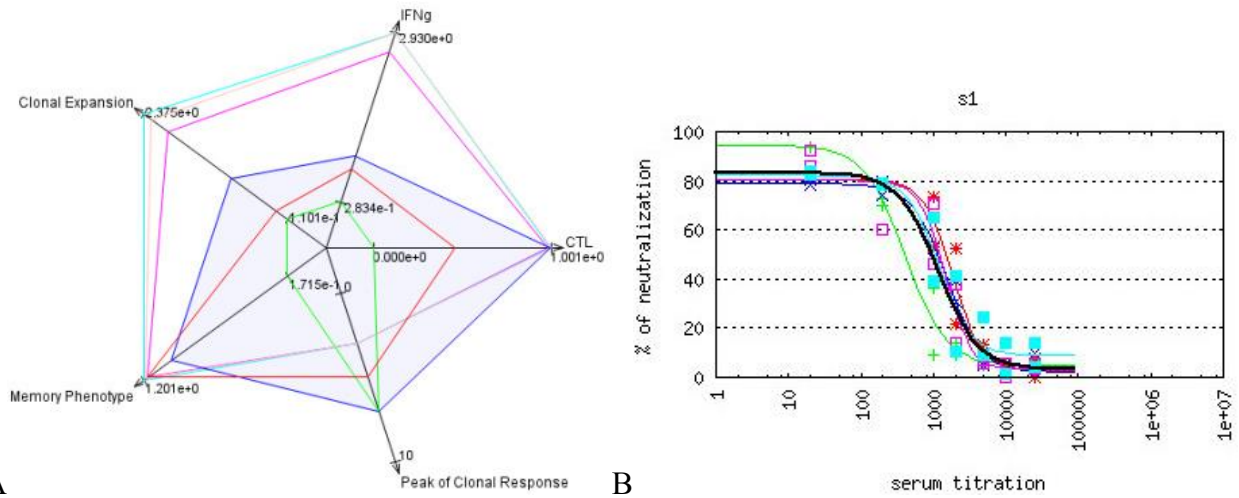


Figure 19. Modules de GeVaDSs.

Représentations graphiques des résultats obtenus pour les modules d'analyse de GeVaDSs. A) Représentation radar des mesures de différents paramètres liés aux LT et à la réponse vis-à-vis du vecteur testé. B) Courbe de neutralisation pour un vecteur testé chez différents individus. Source : (Blazewicz et al., 2012).

TOLÉRANCE FŒTO-MATERNELLE

Parmi les cellules qui participent au développement, les trophoblastes sont un sujet d'étude très intéressant pour qui s'intéresse aux cancers. Issues de la différenciation des cellules du cytotrophoblaste, elles constituent ce qui sera plus tard la partie fœtale du placenta. Elles possèdent des caractéristiques très similaires aux cellules cancéreuses (Ferretti et al., 2007; Holtan et al., 2009) :

- Auto-suffisantes pour les signaux de croissance
- Insensibilité aux signaux anti-croissance
- Échappement à l'apoptose
- Un potentiel de réplication illimité
- Une forte activité angiogénique
- Une invasion des tissus de l'organisme
- Une capacité à détourner la machinerie immunitaire de l'organisme (échappement immunitaire).

La ressemblance avec les cellules cancéreuses est telle que les trophoblastes ont été considérés comme des cellules pseudo-malignes (Strickland and Richards, 1992), voire des métastases physiologiques (Genbacev et al., 1997).

Comme pour le cancer, il existe localement une modulation de la réponse immunitaire de l'organisme afin d'empêcher ce dernier de provoquer la destruction du futur fœtus. Dans le cas du cancer, l'immuno-modulation est (Munn and Mellor, 2016):

- Acquise, car même un antigène étranger est toléré s'il est placé sur des cellules tumorales
- Active, car la tolérance ne peut pas être mise en défaut par un très bon antigène ou des adjuvants
- Dominante, car des LT effecteurs, activés et spécifiques des antigènes tumoraux, transférés chez l'organisme sont eux aussi sujets cette régulation.

Ces caractéristiques expliquent pourquoi la tumeur échappe au système immunitaire alors même que l'environnement tumoral est riche en cellules immunitaire. L'endomètre utérin (ou décidua) est lui aussi très riche en cellules du système immunitaire inné (cellules NK, cellules dendritiques et macrophages). Elles représentent environ 40% des cellules de l'endomètre (Moffett-King, 2002).

Le phénotype des cellules NK utérines, les plus abondantes, est différent de celui des cellules NK du sang. Ces cellules n'expriment pas le récepteur CD16 nécessaire à la réponse dépendante de l'antigène. Elles expriment en revanche la galectin-1 décrite pour induire les cellules dendritiques tolérogènes (Koopman et al., 2003). Elles ont donc un rôle d'immuno-régulateur, plus que cytotoxique. Leur déplétion, dans des modèles de souris IL15-KO, ne permet pas d'induire le rejet du fœtus (Barber and Pollard, 2003). Les macrophages participent également à l'instauration d'un milieu tolérogène par la sécrétion de IDO et d'IL-10 (Abumaree et al., 2006). IDO est, par ailleurs, décrit pour être sécrétée par certaines cellules cancéreuses dans l'environnement tumoral, faisant de lui une cible thérapeutique (Munn and Mellor, 2016). IDO agit notamment sur une troisième population importante dans la tolérance de l'organisme pour le fœtus, les LTreg. Les LTreg sont des LTCD4 exprimant fortement le CD25 et Foxp3. Ce dernier marqueur est au centre du programme régulateur des LTreg (Hori et al., 2003).

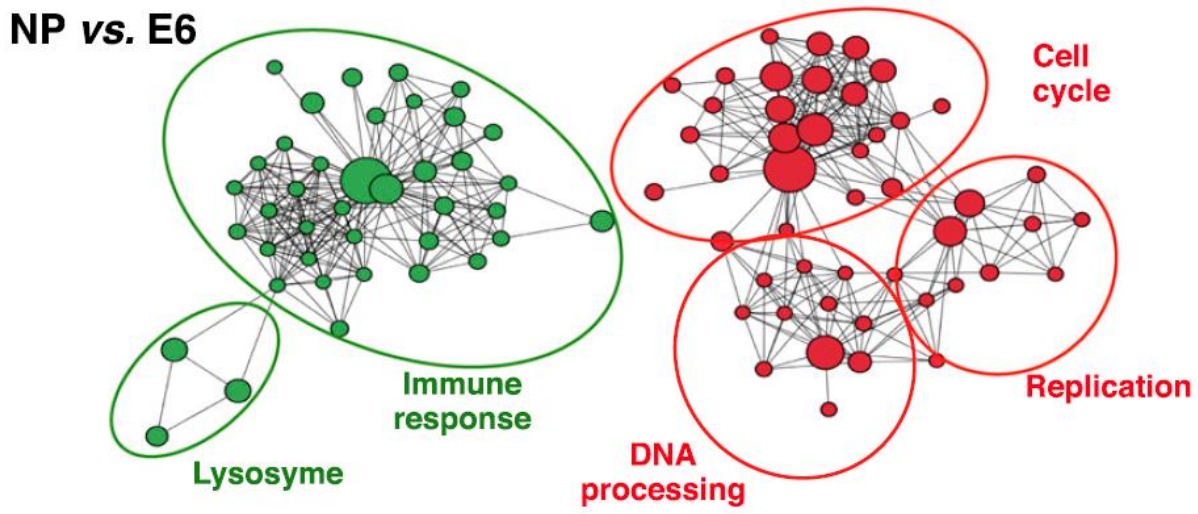


Figure 20. Enrichissement de signatures dans l'environnement fœtal.

Les réseaux de signatures sont créés à partir des données du jeu de données tolérance fœto-maternelle : Les groupes expérimentaux E6 (A) et E12 (B) sont comparés au jeu de données contrôle NP. Des listes ordonnées de gènes sont constituées à partir du test t modifié (eBayes) comparant E6 (ou E12) à NP. Une collection de signatures moléculaires issues de Gene Ontology est testée sur les listes ordonnées par l'outil GSEA. Les signatures sont alors triées pour leur significativité d'enrichissement. Il reste donc les signatures statistiquement régulées (FDR<0.01) dans une condition par rapport au contrôle. Les réseaux sont obtenus par Cytoscape à l'aide du module Enrichment Map. Les nœuds sont des signatures (vertes si sous-exprimées et rouges si sur-exprimées) dont la taille est proportionnelle au nombre de gènes dans la signature. Les liens sont définis sur la base d'un test de Jaccard. Source : (Nehar-Belaid et al., 2016).

Tout comme pour les tumeurs (Darrasse-Jèze et al., 2009), les LTreg sont recrutés dès les premiers jours d'implantation du fœtus dans les ganglions utérins (Sasaki et al., 2004) ainsi qu'au niveau de l'environnement fœtal (Chen et al., 2013). Des cas de fausses-couches sont d'ailleurs associées à une nette diminution des LTreg par rapport à des grossesses saines (Jin et al., 2009) et leur élimination dans des souris gestantes provoque un rejet du fœtus (Aluvihare et al., 2004). De la même manière, la déplétion de LTreg chez des souris développant des tumeurs induit la régression de ces dernières (Sakaguchi et al., 2001).

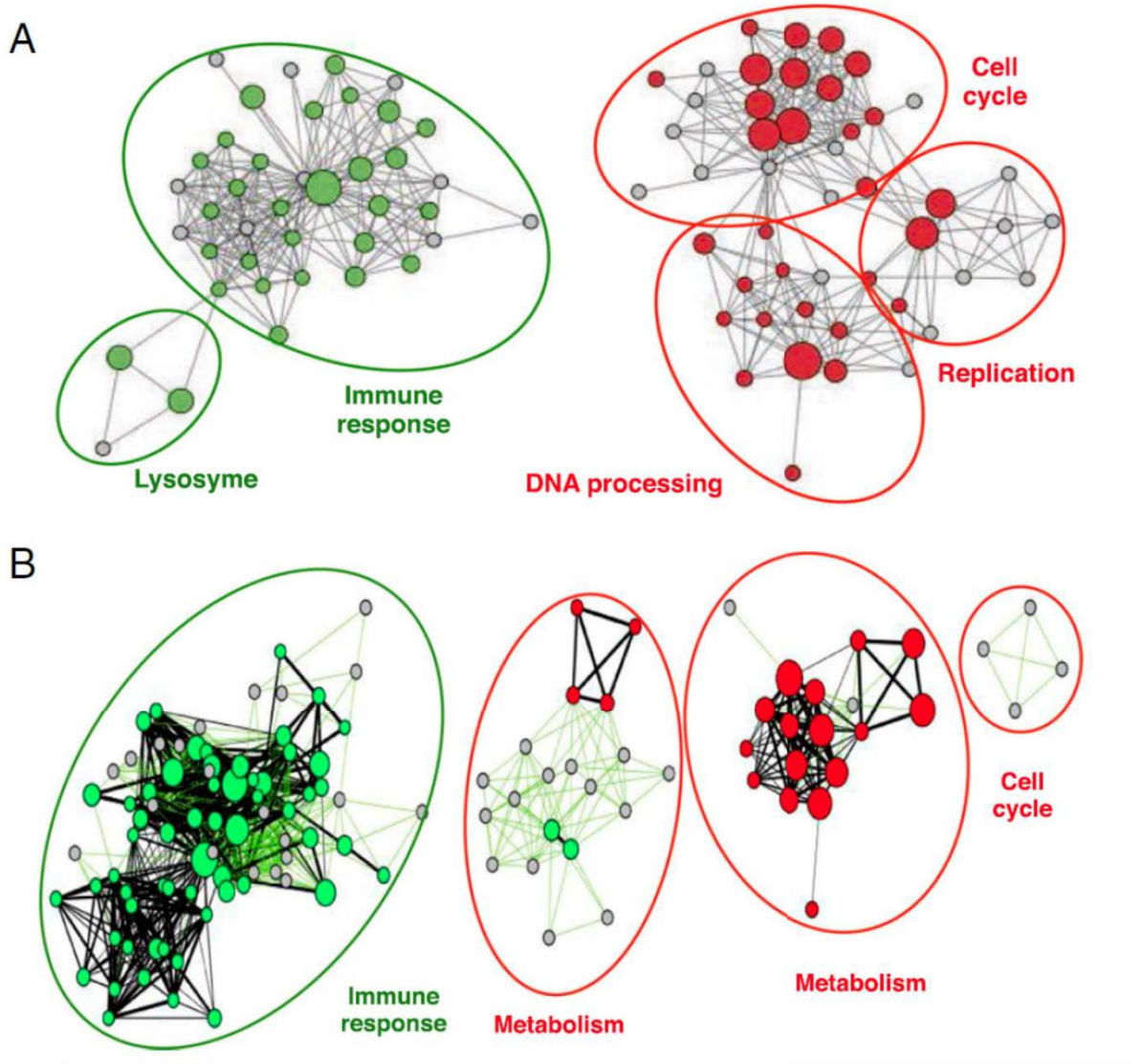


Figure 21. Relation entre environnements fœtal et tumoral.

A (respectivement B) : Superposition des signatures significativement enrichies (FDR<0.05) dans l'environnement tumoral 4 jours (14 jours) après injection sous-cutanée d'une tumeur de type mélanome B16F10 sur le réseau de signatures calculé à partir des données de souris gestantes 6 jours (12 jours) après le coït. Source : (Nehar-Belaid et al., 2016).

Grâce aux approches développées au laboratoire (Pham et al., 2014), nous avons tout d'abord décrits les processus moléculaires mis en jeu lors de l'installation de la tolérance fœto-maternelle (Figure 20). Dès le quatrième jour après le coït, l'environnement fœtal voit des signatures moléculaires liées à la réplication de l'ADN et au cycle cellulaire être fortement positivement régulés. À l'inverse, des signatures liées au système immunitaires sont très fortement négativement régulées. La Figure 20 montre le réseau de ces signatures au sixième jour après le coït.

Nous avons ensuite montré qu'il existe effectivement des mécanismes moléculaires communs mis en jeu durant l'installation de la tolérance fœto-maternelle et celle de la tumeur. La Figure 21 montre les signatures communes aux deux environnements et leur régulation.

Que ce soit précocement ou tardivement, les réseaux moléculaires des deux environnements sont très proches et indiquent tous les deux une induction de l'activité immunitaire régulatrice, induite et orchestrée par les LTreg en étroite association avec les DC, et une augmentation des processus de prolifération et des métabolismes cellulaires.

Pour démontrer cela, notre laboratoire a construit des jeux de données autour des microenvironnements tumoraux et fœtaux. Le jeu qui nous intéresse pour cette thèse est constitué de vingt-six échantillons d'utérus de souris C57BL/6 de 6 à 8 semaines pris quatre, six, huit, dix, onze et douze jours après le coït. Les utérus de cinq souris C57BL/6 non-gestantes servent de contrôle négatif pour cette expérience. À noter que l'implantation fœtale s'effectue entre le quatrième et le sixième jour (Aghion and Poirier, 2000). La préparation des ARN et des puces à ADN est décrite dans la section Matériel et Méthode de l'article Nehar-Belaid *et al.* auquel j'ai contribué (Annexe #2). Par la suite, je ferai référence à ces données comme le jeu de données Tolérance fœto-maternelle.

LYMPHOCYTES T RÉGULATEURS ET IL-2

Avant les travaux effectués au milieu des années 1990, notamment de ceux de B. Sadlack (Sadlack et al., 1993, 1994, 1995), l'IL-2 était connue pour être nécessaire à l'activation les LT. Les LT sont activés par la stimulation TCR et des molécules de co-stimulation (CD80, CD86, CD40…). Cela engendre la production de l'IL-2 et de son récepteur IL-2R. L'interaction entre IL-2 et IL-2R participe à l'induction de l'expansion clonale des LT, faisant de l'IL-2 un acteur majeur de la réponse immunitaire (Cheng et al., 2011). B. Sadlack montre alors que des souris, dont les gènes exprimant l'IL-2 ou son récepteur sont inactivés, présentent des pathologies autoimmunes

associées à des proliférations anarchiques des LT. Dans la même période, S. Sakaguchi découvre que la population de LT régulateurs, qu'il avait découverte dix ans auparavant (Sakaguchi et al., 1996), est caractérisée par le cluster de différenciation CD25 : le récepteur de l'IL-2 de haute affinité. Ce cluster de différentiation est en fait la chaîne α du récepteur de l'interleukine-2 (IL-2), l'une des trois chaînes du récepteur de l'IL-2 avec CD122 (chaîne β) et CD132 (chaîne γc) (Sakaguchi et al., 2001).

Les LT effecteurs expriment aussi CD25 à leur surface et plus encore lorsqu'ils sont activés et l'IL-2 est importante pour leur différenciation terminale, leur expansion clonale, leur survie et l'instauration du phénotype mémoire (Cheng et al., 2011). L'IL-2 est aussi importante pour le développement des LTreg : concomitante à STAT5, elle induit la différenciation des LTreg dans le thymus aboutissant à l'expression de FOXP3 dans la cellule (Fontenot et al., 2003). L'IL-2 est aussi nécessaire au maintien du phénotype $FOXP3^+$ des LTreg, et donc de leur fonction régulatrice. Les différences entre LT effecteurs et LTreg concernant l'action de l'IL-2 et de son récepteur sont résumés en Table 1 et Figure 22.

Table 1 : Système IL-2/IL-2R dans les LTreg et les LT effecteurs (Cheng *et al*, 2011)

Property	Treg	Teff
IL-2 production	No	Yes
Repression of IL-2	Foxp3/Runx/NF-AT	T-bet, Blimp-1
High affinity IL-2R expression	High, constitutive	Low, transient
Upregulation of IL-2Rα	Yes	Yes
IL-2R signaling	STAT5	MAPK; PI3K/Akt; STAT5
Activity during T cell development	Yes	No
Peripheral homeostasis in vivo	Yes	No
Growth	Yes	Yes
Survival	Yes	Yes
Regulation of function	Suppression	TH1, TH2

L'activation de l'activité régulatrice des LTreg par l'IL-2 a donc amené les chercheurs à se demander s'il n'était pas possible d'activer sélectivement les LTreg sans moduler les LT effecteurs,

et ainsi favoriser la régulation de la réponse immunitaire. Le pari est donc de parvenir à un traitement de pathologies grâce à de l'IL-2 faible dose (Klatzmann and Abbas, 2015). L'intérêt d'un tel schéma est évident dans les cas de pathologies autoimmunes et/ou autoinflammatoires. L'effet recherché des traitements dans ces pathologies est la réduction de la réponse immunitaire contre les cellules du soi. En activant les LTreg, ceux-ci vont inhiber l'activation des LT effecteurs et par conséquent diminuer le processus de destruction de l'organe.

Dans ce contexte, l'équipe de T. R. Malek a produit un jeu de données de transcriptome visant à mieux comprendre les mécanismes moléculaires sous-jacents à la stimulation des LTreg par l'IL-2. Ce jeu de données comprend neuf échantillons de LTreg triés prélevés sur des donneurs sains. Ces échantillons ont chacun été divisés en deux échantillons et cultivés 24 heures *in vitro*. Un des échantillons de la paire est cultivé avec 100 u/ml d'IL-2, l'autre non. Les ARN sont extraits, purifiés et amplifiés avant d'être déposés sur des puces Affymetrix Gene ST1.0.

Par la suite, je ferai référence à ces données comme le jeu de données Treg.

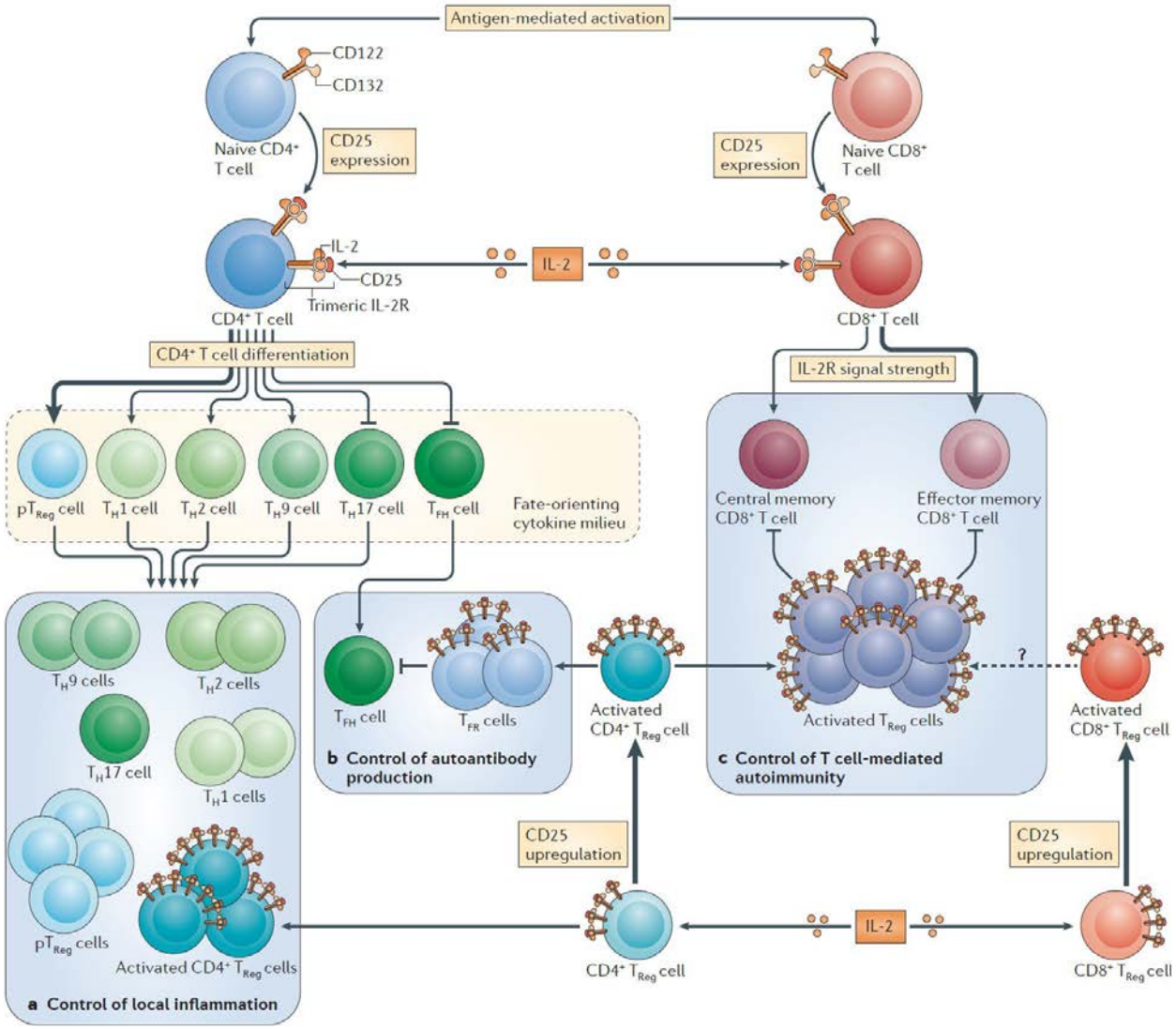


Figure 22. Rôles de l'IL-2 sur les lymphocytes.

Effets pléïotropes de l'IL-2 dans le contrôle de l'autoimmunité. Les LTCD4 et les LTCD8 naïfs expriment de manière transitoire le cluster de différentiation CD25. Le cocktail de cytokines présentes dans le milieu détermine la différentiation des LTCD4 en LT auxiliaires de type 1 (T_H1), T_H2, T_H9 T_H17, T_{FH} et LTreg. L'orientation de la différenciation est influencée par la concentration du signal IL-2. Très forte, elle oriente la différenciation vers les LTreg et les LT effecteurs mémoires. Les LTreg expriment CD25 de manière constitutive et répondent fortement à l'IL-2 par une augmentation de l'expression du CD25 conduisant à un phénotype encore plus régulateur. Certaines de ces cellules migrent vers les ganglions lymphatiques des centres germinatifs (T_{FR}). Ces changements conduisent à orienter la réponse immunitaire vers la régulation plutôt que l'inflammation, notamment par : l'induction de la différenciation des LTCD4 en LTreg plutôt qu'en cellules T_H17(a) ; le contrôle de la production d'anticorps en favorisant la différenciation des T_{FR} plutôt que des T_{FH} (b) ; le contrôle des LTCD8 effecteurs par l'augmentation en nombre et en qualité des LTreg (c). Source : (Klatzmann and Abbas, 2015).

MÉTHODES D'ANALYSE CLASSIQUES

LES DONNÉES

Les méthodes d'analyse des données transcriptomiques sont directement liées à la technologie utilisée. Si notre laboratoire mise maintenant sur le RNAseq pour l'étude transcriptomique des essais actuels et futurs, il n'en fut pas toujours le cas et nous possédons une collection d'expériences faites sur puce à ADN. Par ailleurs, la littérature est particulièrement riche en données de puces à ADN comparativement à celles de RNAseq notamment en ce qui concerne des expériences mettant en scène de grandes cohortes. L'atout majeur du RNAseq est la découverte de nouveaux transcrits codants ou non car en ce qui concerne la quantification des transcrits, cette technologie apporte finalement assez peu par rapport aux puces à ADN. La sensibilité est meilleure, certes, mais elle n'élimine pas le biais majeur de la préparation des échantillons, l'étape d'amplification. Les transcrits de faibles abondances ont tendance à ne pas sortir lors des analyses car les transcrits de fortes abondances captent l'ensemble de l'information. Il est aussi possible, sinon probable, que deux transcrits de même abondance initiale, donne des abondances différentes car il suffit d'un cycle où l'amplification ne se passe pas de manière optimale pour un des transcrits pour générer des variations importantes dans l'abondance finale. J'attends avec impatience l'avènement des systèmes de séquençage sans amplification, et leur prédit un impact sur la biologie moléculaire aussi important que la PCR en son temps. La thèse ici développée se base sur des données transcriptomiques issues des puces à ADN et je vais donc me focaliser sur les méthodes d'analyses de ces dernières.

Il est important de rappeler ici que la bonne qualité des données est la condition *sine qua none* pour la réussite d'un projet. Chaque étape de préparation des échantillons à analyser est susceptible d'être impactée par des erreurs de précision et/ou d'exactitude. Il suffit alors de penser au nombre d'étapes nécessaires pour extraire les transcrits d'un échantillon, plus d'une quinzaine, pour comprendre que l'ampleur des erreurs peut avoir un impact important dans la mesure finale. Les protocoles sont standardisés et il est aussi possible d'automatiser les différentes étapes pour limiter le facteur humain. Mais même ainsi, il y aura toujours du bruit technique dans les données. Savoir si le transcriptome d'un échantillon est exploitable ou non est relativement aisé car chaque technologie possède ses contrôles internes permettant de se faire une idée de la qualité de

l'hybridation. Une première étape va donc consister à analyser les informations de ces contrôles internes. Les contrôles internes se répartissent en deux catégories : les contrôles positifs et négatifs. Les contrôles positifs sont des sondes qui capturent des ARN fortement exprimés quel que soit le contexte ; il s'agit de transcrits issus de gènes dits de ménage. Les contrôles négatifs correspondent à des sondes qui ne capturent pas d'information sur les transcrits de l'échantillon. Ils doivent par conséquent ne mesurer que le bruit de fond de la puce. La deuxième étape consiste à analyser le comportement global des puces. Nous pouvons produire, par exemple, une matrice de corrélation entre les échantillons. Deux échantillons, même provenant de groupe expérimentaux différents, doivent garder une corrélation assez forte (>0.9). Cela vient du fait que la part des transcrits qui varient entre deux groupes expérimentaux est faible comparativement à l'ensemble des transcrits mesurés. Par conséquent, identifier une puce avec une mauvaise corrélation indique un problème technique sur la puce, généralement rédhibitoire pour l'analyse de cette puce. La troisième étape consiste à exporter l'information sur la manière de produire les données. Ces informations sont une source de renseignements sur les facteurs produisant du bruit technique dans les données. Reprenons les données présentées en introduction (Figure 2) : nous avions vu que le design des puces à ADN de la technologie Illumina (6 échantillon par lame) explique 29.6% de la variance du jeu de données. Connaissant ce biais il est possible, lorsque la répartition des échantillons sur les lames est faite correctement, de le corriger par des méthodes mathématiques telle que celles utilisées dans la librairie ComBat disponible sous R (Johnson et al., 2007; Müller et al., 2016). La Figure 23 montre qu'une fois les données traitées, ce sont bien les facteurs biologiques qui expliquent le plus de variance du jeu de données, le résiduel reste quant à lui très fort et correspond à la part de variabilité non expliquée par les facteurs testés donc potentiellement générée par d'autres facteurs techniques.

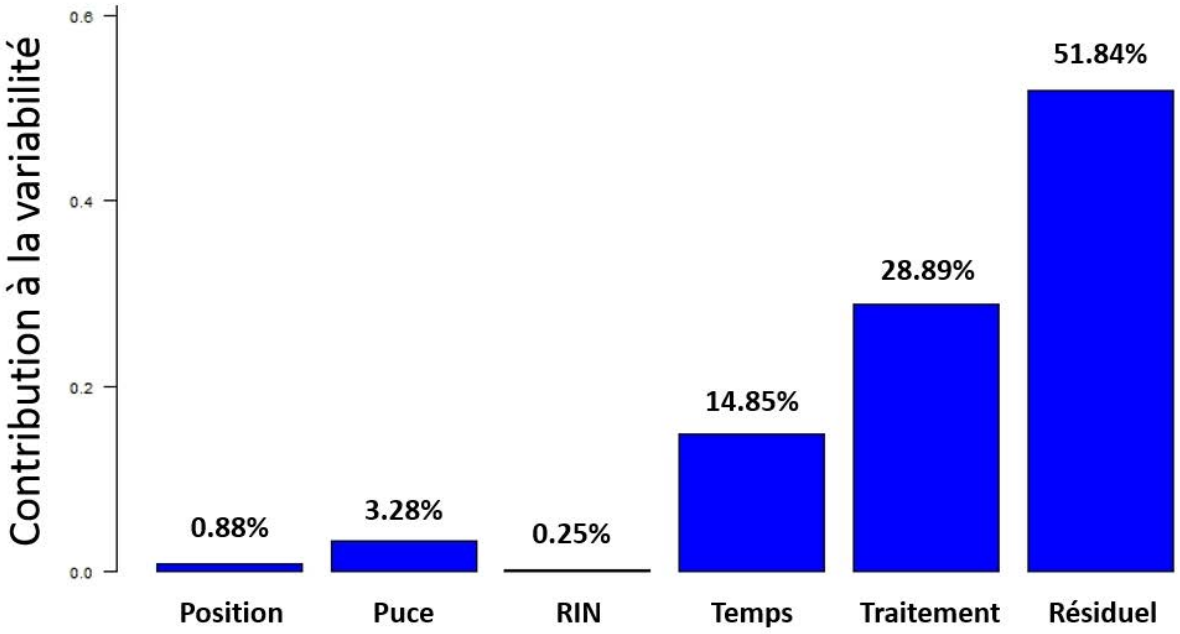


Figure 23. Impact de la correction des biais techniques sur un jeu de transcriptome.
Contribution de différentes sources de variabilités à la variance observée dans les données présentées en Figure 2 après correction par l'algorithme ComBat. La correction est effectuée en précisant le facteur à corriger, ici le facteur Puce, et en fixant les facteurs biologiques. Résultats obtenus au laboratoire dans le cadre du projet CompuVac.

Une fois nettoyée, la quatrième étape a pour but de rendre les données comparables les unes avec les autres, il s'agit de l'étape de normalisation. Les méthodes de normalisation des puces à ADN sont légions et leurs auteurs ont tous de bonnes raisons de penser que la leur est la meilleure. Devant l'ampleur des possibilités, je me suis fié aux comparatifs effectués par Schmid *et al.* (Schmid et al., 2010) où les auteurs tests différentes méthodes de normalisation. Il en ressort que la normalisation par quantile couplée à une transformation des données par le log2 donne des résultats parmi les plus satisfaisants pour les données Illumina. J'ai toujours utilisé cette combinaison depuis car comme on le constate dans ce même papier, d'une normalisation à l'autre, les résultats varient grandement ce qui pose un problème de cohérence pour la comparaison de résultats d'analyse de plusieurs jeux de données différents. La méthode de normalisation par quantile part du principe simple, et vrai en théorie, que chaque puce possède la même quantité d'information de fluorescence d'un échantillon à l'autre. Elle prend aussi en compte le fait que la majorité des transcrits ne sont pas impactés par les perturbations biologiques et que, par conséquent, la distribution des données

doit être globalement identique d'un échantillon à l'autre. La procédure est basée sur plusieurs étapes :

- Ordonner chaque échantillon de la plus forte expression à la plus faible
- Calculer la médiane d'expression de chaque rang de valeurs
- Remplacer les valeurs initiales d'un rang par sa médiane
- Réordonner les valeurs comme à l'origine

De ce fait, la valeur d'expression du gène le plus fortement exprimé (et respectivement, le moins exprimé) dans chaque échantillon est désormais la même. La distribution des expressions des gènes et la quantité d'information sont alors parfaitement identiques d'un échantillon à l'autre comme le montre les distributions présentées dans la Figure 24 sous forme de boîtes à moustaches.

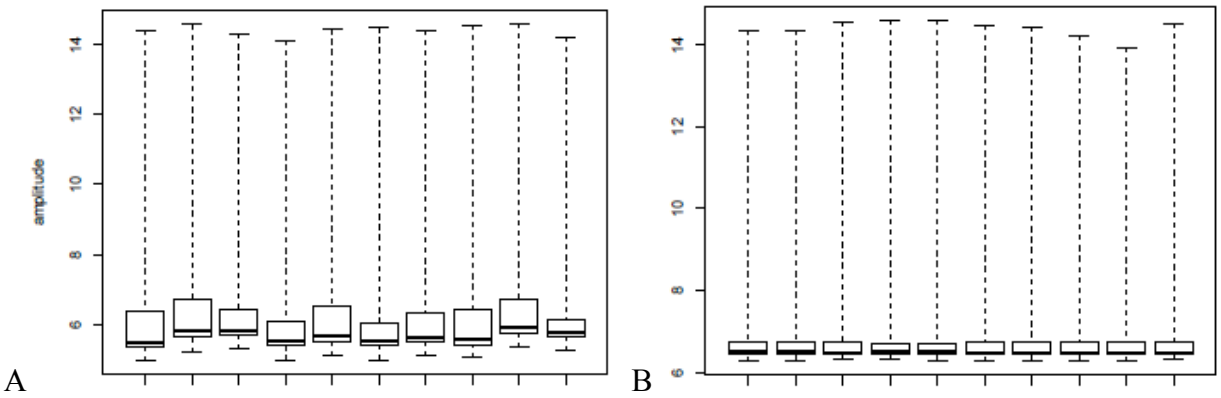


Figure 24. Distribution des données avant et après normalisation.

Distribution de valeurs d'expression log2-transformées de 10 échantillons avant (A) et après (B) normalisation par la méthode des quantiles. Données produites au laboratoire par l'algorithme *quantroSim*.

L'étape suivante consiste à continuer le nettoyage des données en enlevant la part des données dont la détection n'est pas fiable. Chaque technologie possède ses propres moyens de calculer un indice de confiance de la détection de chaque transcrit. Sachant que chaque sonde est systématiquement représentée plusieurs fois sur la puce, la qualité de la détection se fait par :

- La comparaison de l'intensité de fluorescence du niveau de la sonde par rapport au bruit de fond mesuré aux alentours de la sonde. Si la fluorescence n'est pas significativement différente du bruit de fond, alors la sonde est considérée comme n'ayant pas hybridé de transcrit marqué.

- La stabilité de la mesure au travers des différentes sondes identiques dispersées sur la puce. Si la variabilité est trop importante alors la sonde est considérée comme n'étant pas fiable dans la mesure de l'abondance de son transcrit.

Les protocoles d'analyses de données fournissent alors des p-values, correspondant au risque de se tromper en disant que la valeur observée est vraie. Une p-value de 0.001 indique donc que le risque de se tromper est d'une chance sur mille, une valeur bien moindre que le nombre de réplicas de sondes testés (entre 20 et 50 selon les fabricants). L'élimination des sondes mal détectées peut prendre plusieurs formes et va dépendre, selon moi, de deux facteurs majeurs, l'espèce observée et la question biologique posée. En biologie, nous utilisons beaucoup de modèles biologiques (animaux ou végétaux) qui sont par définition très contrôlés pour leur production et leur environnement de vie. Dans ces cas précis, il est attendu, lors d'une expérience impliquant plusieurs individus de la même fratrie de souris par exemple, que les individus soient stables les uns par rapport aux autres. Il est alors tout à fait concevable de considérer une sonde comme étant bien détectée pour un jeu de données si et seulement si elle possède une p-values de détection correcte dans au moins les deux-tiers des échantillons d'un groupe expérimental. Cela vaut bien sûr pour chaque groupe expérimental de l'expérience. En revanche, travailler sur l'humain par exemple implique une variabilité très importante et une sonde peut alors être considérée comme intéressante si elle est bien détectée dans un seul échantillon. Cette logique rejoint le fait que si l'on s'intéresse à une réaction globale pour un groupe d'individus, on cherchera à assurer la qualité des détections dans l'ensemble des échantillons, en revanche, si l'on s'intéresse, comme moi à la variabilité inter-individuelle, il est préférable de garder toutes les sondes pourvu qu'elles soient bien détectées dans un échantillon au moins.

La dernière étape de la préparation des données consiste à rassembler en une seule mesure les valeurs de sondes qui ciblent le même transcrit, sur la base de l'annotation biologique fournie par le fabriquant. Quand ce cas apparaît, nous calculons la médiane de l'expression des sondes concernées. La matrice de valeurs d'expression finale comprend alors une valeur pour un transcrit à travers les échantillons de l'expérience.

ANALYSE DES DONNÉES

Les données nettoyées sont un préalable important à l'analyse, les variations d'expression de gènes ressortent avec d'autant plus de finesse et de justesse que les données sont moins bruitées. Dans le cadre des analyses classiques effectuées en transcriptomique, les descripteurs et les tests statistiques sont des méthodes relativement faciles à mettre en œuvre. Les analyses statistiques effectuées sur les trois jeux de données, décrits dans la section « Origine des données » par un test t modifié (Phipson et al., 2016), nous donnent des nombres de gènes statistiquement différemment exprimés de l'ordre du millier pour le jeu de données Treg à plusieurs milliers pour les jeux de données LPS (>5000) et Tolérance fœto-maternelle (entre 2500 et 7000 suivant les groupes expérimentaux comparés). Il est fréquent d'exprimer l'expression d'un gène dans un groupe d'individus par la moyenne arithmétique de son expression et son écart-type. La moyenne arithmétique est un estimateur sans biais de l'espérance du groupe d'échantillons observés, telle que :

Équation 2 :

$$\mathbb{E}[g] = x_1.p_1 + x_2.p_2 + \cdots + x_n.p_n$$

Équation 3 :

$$\bar{x}_g = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Où $\mathbb{E}[g]$ est l'espérance du gène g, soit la somme des n valeurs x que peut prendre ce gène multipliées par leur probabilité p. L'estimateur sans biais de $\mathbb{E}[g]$ est $\bar{x}_g$, c'est-à-dire la moyenne arithmétique de l'expression du gène g au sein d'un groupe d'échantillons de taille n. L'écart-type peut alors lui aussi être estimé en utilisant $\bar{x}_g$, tel que :

Équation 4:

$$s_g = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_g)^2}$$

L'écart-type s_g correspond alors à une mesure de la dispersion des valeurs mesurées pour le gène g par rapport à la moyenne de ces valeurs $\bar{x}_g$. Cette notion de dispersion est intéressante car elle correspond à la définition que je donnais de la variabilité au début de ce manuscrit. L'écart-type est en effet une manière d'appréhender la variabilité d'un groupe d'échantillons en prenant comme

point de repère la moyenne. Il est à relier aussi à la notion de précision, vue dans l'Introduction, pour la mesure des erreurs techniques.

La moyenne et l'écart-type présentent toute fois des limites d'ordre intellectuel lorsqu'il s'agit de les appliquer à des données biologiques. La moyenne est un bon descripteur si et seulement si les données sont distribuées selon une loi normale parfaitement centrée ($\mathcal{N}(\mu, \sigma)$, où μ et σ sont la moyenne et l'écart-type de l'échantillonnage, respectivement estimé par $\bar{x}$ et s vus plus haut). Toute perturbation à cette distribution biaise le calcul de moyenne. Or biologie, il est fréquent d'avoir des distributions excentrées, ou asymétriques. L'estimation sans biais de l'asymétrie d'une distribution normale est donnée par l'équation suivante :

Équation 5:

$$\hat{S} = \frac{n^2}{(n-1)(n-2)} \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{(\sigma^2)^{3/2}}$$

Où n est le nombre d'échantillon, $\bar{x}$ et σ^2 sont respectivement la moyenne et la variance de l'échantillonnage. Une valeur positive indique une distribution décalée vers la gauche, une distribution négative indique une distribution décalée vers la droite, tandis qu'une valeur négative correspond à une distribution normale sans asymétrie. L'asymétrie positive rend l'utilisateur de la moyenne erronée pour la description des données car elle va être d'autant plus fortement biaisée que la queue est étalée vers la droite car elle contient des valeurs extrêmes. Dans ce cas précis, il est préférable d'utiliser la médiane comme descripteur des données. La médiane sépare les données en deux groupes de taille identique, quelles que soient les valeurs associées aux groupes. Ainsi, la médiane n'est pas influencée par le poids de certaines valeurs extrêmes mais ne s'intéresse qu'à la structure de l'ensemble de données.

L'écart-type est lui aussi très influencé par les valeurs mesurées et par voie de conséquence par la moyenne calculée. Plus les valeurs sont fortes, plus l'écart entre les valeurs et la moyenne seront fortes sans pour autant représenter une dispersion plus importante. Pour contrer ce biais, il est possible d'utiliser l'estimateur du coefficient de variation $\hat{c}_v$ qui se défini comme l'écart-type pondéré par la moyenne.

Équation 6:

$$\hat{c}_v = \frac{s}{\bar{x}}$$

Nous retrouvons s et $\bar{x}$ respectivement l'écart-type et la moyenne de l'échantillonnage observé. Dans ce cas, un échantillonnage ayant pour moyenne 10 et un écart-type de 2 montre la même dispersion des données qu'un échantillonnage ayant pour moyenne 1000 et un écart-type de 200.
Nous avons là quatre outils classiques pour décrire les données, la moyenne, la médiane, l'écart-type et le coefficient de variation.
Si ces outils restent indispensables, car utilisé en statistique notamment, il existe dans l'arsenal du scientifique bon nombre d'autres techniques d'analyses qui aide à la compréhension de la structure du jeu de données. Plutôt que de passer en revue ces méthodes je vais dans les pages qui viennent décrire comment, en utilisant certain de ces outils, nous avons développé une méthodologie efficace, qui tient compte de la variabilité inter-individuelle, pour l'évaluation et la classification de vecteurs destinés à la vaccination (Dérian et al., 2016).
Ma mission dans le projet CompuVac, décrit plus haut, fut de développer une stratégie de classification de vecteurs vaccinaux sur la base du transcriptome des DC six heures après l'injection du vecteur chez la souris. L'intérêt pour les cellules dendritiques n'est pas anodin puisqu'elles sont les premières actrices de la réponse adaptative par leur activité de présentation des antigènes. L'évaluation de l'efficacité des vecteurs à induire une bonne réponse immunitaire se fait par la mesure de l'expansion clonale des lymphocytes T spécifiques du gp33-41 chez la souris 5, 7 et 10 jours après l'injection du vecteur.

Extraire l'information

L'analyse supervisée (voir définition sans l'encadré « Analyse supervisée et non supervisée ») de données transcriptomiques, effectuées sur des puces à ADN du fabriquant Codelink, donnait des profils d'expression très contrastés entre les vecteurs induisant des réponses immunitaires fortes et ceux induisant des réponses nulles. Je parle ici de plusieurs milliers de gènes individuellement statistiquement régulés dans le premier cas et d'aucun dans le second. Le nombre de gènes régulés étant extrême dans les deux cas, la stratégie à développer consistait alors en un *workflow* incluant une étape d'analyse non supervisée visant à rechercher l'information sous-jacente aux valeurs d'expression normalisées. Le choix de la méthode s'est porté sur l'Analyse en Composantes Indépendantes (Chiappetta et al., 2004), une méthode statistique de réduction de dimensions. L'ACI recherche à travers les données des axes, appelés des sources, qui contiennent l'information décrivant un sous-ensemble des données. L'ACI rend compte de la contribution de chaque gène

pour cette source, répondant ainsi à une particularité de la mesure du transcrit. En effet, la valeur mesurée en puce à ADN, ou en séquençage massif, est l'accumulation de l'expression du gène d'intérêt dans l'ensemble des contextes pour lequel il agit :

- Différentes populations cellulaires
- Différentes voies de signalisation au sein d'une population cellulaire donnée.

Ainsi l'expression d'un gène peut se traduit par une combinaison linéaire.

Équation 7:

$$x_i = \alpha A + \beta B + \cdots$$

Où A et B seraient des contextes biologiques différents, α et β représentent alors les coefficients illustrant la part pris des différents contextes dans l'expression totale.

L'ACI produit en sortie d'analyse deux matrices nommées A et S telles que :

Équation 8:

$$X = A.S$$

Où X est la matrice d'expression originale de dimension $g \times m$, où m est le nombre d'échantillons et g est le nombre de gènes présents dans le jeu de données. A est une matrice, dite de mélange, de dimension $n \times m$, où n vaut le nombre de sources trouvées par l'organisme, elle contient les poids associant les différents échantillons à chaque source. S est une matrice de dimension $m \times g$, contenant les poids associant les différents gènes à chaque source. Les sources répondent à deux critères importants :

- Elles sont mutuellement indépendantes
- Elles sont le moins gaussiennes possible

Ce dernier point est spécifique à l'algorithme que nous utilisons pour la recherche des sources. Il est décrit dans la librairie *fastICA* (Hyvärinen, 1999) disponible pour le logiciel R (www.r-project.org, voir l'Encadré « R project »). L'auteur de cette librairie propose, entre autres, rechercher des sources dont la distribution est de la forme :

Équation 9:

$$\Phi(x) = \frac{1}{\alpha}\log(\cosh(\alpha x)), \alpha \in [1,2]$$

Le choix de cette distribution se base sur le fait que les variations biologiques n'impliquent qu'un faible nombre de gènes parmi les gènes analysés. Dans ce cas, la plupart des gènes sont concentrés

autours d'une moyenne, seuls quelques gènes se retrouve aux extrémités, ce qui correspond à une distribution de type super-gaussienne (Lee and Batzoglou, 2003).

ICA est décrit pour proposer des composantes voyant les gènes interagir de manière synergétique c'est-à-dire de concert pour créer un effet global. La matrice *S* nous intéresse particulièrement ici car c'est elle qui nous indique ces groupes de gènes. La distribution des poids associant les gènes à une source montre que la plupart des gènes ne portent aucune contribution à la source tandis que quelques gènes on des poids très forts marquant ainsi leur forte contribution pour le contexte décrit par la source. En sélectionnant ces gènes sur la base de la déviation standard de la distribution des poids de la matrice *S* tel que décrit dans (Chiappetta et al., 2004), deux groupes de gènes sont obtenus pour chaque source. L'analyse de chaque jeu de données va alors produire une collection de signatures, au sens de groupes de gènes caractérisant, ou signant donc, une structure particulière du jeu de données.

La méthode proposée par *fastICA* a cependant une contrainte, elle ne donne pas nécessairement le même résultat d'un calcul à l'autre. En effet, la découverte des sources se base sur l'estimation d'une matrice W, telle que :

Équation 10:

$$W \approx A^{-1}$$

Analyse supervisée et non supervisée. En matière d'analyse de données, il est possible d'appréhender les données de deux manières différentes. La première consiste à donner du poids à des informations sur les données, en définissant par exemple qu'un ensemble d'échantillons issus du jeu de données possèdent des caractéristiques similaires et se doivent donc d'être regroupés en une classe particulière. Définissant plusieurs classes de la sorte, il est possible d'extraire l'information qui explique la différence entre ces différentes classes, par un test statistique par exemple. Il s'agit de l'analyse supervisée au sens où la définition des classes est faite sous la supervision de l'expérimentateur. *A contrario*, une méthode d'analyse non supervisée consiste à laisser l'algorithme rechercher lui-même le regroupement des échantillons sur la base seule des valeurs fournies par le jeu de données et d'extraire l'information qui explique ces regroupements.

Cette matrice est initiée par des valeurs aléatoires, l'algorithme modifiant itérativement ces valeurs pour maximiser les paramètres. L'initiation de la matrice est donc différente à chaque itération de l'algorithme, engendrant des différences dans les résultats. Afin de pallier à ce problème, nous procédons à plusieurs itérations de l'algorithme et moyennons les sources semblables. Deux sources sont dites semblables quand leur corrélation est supérieure ou égale à 0,9. Un indice de crédibilité est aussi calculé, correspondant, pour chaque source, au nombre de sources nécessaires pour l'obtenir. Un indice fort indique donc une source dont on retrouve des sources semblables dans un grand nombre d'itérations. Les groupes de gènes sont alors extraits de ces sources moyennes et nous les définissons comme des signatures potentielles (Pham et al., 2014).

Information et contexte biologique

Reste à savoir si ces signatures potentielles ont un intérêt pour la différentiation des vecteurs induisant des réponses immunitaires différentes. Il est possible de regarder le comportement des gènes d'une signature pour la comparaison d'intérêt en utilisant la méthode d'analyse des groupes de gènes (*GSA : Gene Set Analysis*). Dans notre cas nous utilisons l'algorithme développé par Subramanian *et al.* (Subramanian et al., 2005) et disponible sur le site du *Broad Insitut*, GSEA (*Gene Set Enrichement Analysis*). GSEA analyse le comportement des gènes d'une signature par rapport à une liste ordonnée de gènes. L'ordonnancement est déterminé par la question posée. Notre question tournant autour de la réponse des individus à l'injection d'un vecteur, nous avons ordonné les gènes selon la statistique associée à un t-test modifié (McCarthy and Smyth, 2009) effectué sur les données d'expressions de souris injectées avec soit une solution saline (PBS : tampon phosphate salin), créant ainsi un groupe témoin, soit avec un vecteur d'intérêt. Les gènes sont ainsi ordonnés du gène le plus sur-régulé dans le cas de la vaccination par rapport au témoin, jusqu'au gène le plus sous-régulé, et ce, quelle que soit la valeur statistique de la différence. Il est alors possible de calculer un score Y représentant le comportement global des gènes d'une signature pour cette liste ordonnée d'intérêt.

Le score Y de départ vaut $Y_0 = 0$. L'algorithme parcours la liste ordonnée de gènes et applique un gain au score Y_{i-1} si le $i^{ème}$ gène de la liste est présent dans la signature d'intérêt. Inversement, le score Y_{i-1} se voit pénalisé si le $i^{ème}$ gène n'est pas présent dans la signature. La valeur du gain dépend de la position du gène dans la liste ordonnée et de la valeur de statistique qui lui est associé. Le score final Y vaut alors la somme des pénalités et des gains fournit à Y_0. Ce qui va nous

intéresser ici, n'est pas le score Y mais comment celui-ci évolue au cours du parcours de la liste ordonnée. L'évolution du score est visible au travers d'une courbe d'accumulation du score illustrant les gains et les pénalités. Le point qui définira le score d'enrichissement ES est son *extremum*, c'est-à-dire le point pour lequel la courbe atteint son maximum ou réciproquement son minimum. Le score ES indique l'existence, ou non, d'un biais dans la composition d'une signature qui se traduit par la présence de gènes majoritairement sur-régulés (réciproquement sous-régulés). La significativité statistique du score est contrôlée par la génération, pour une signature donnée, de 1000 signatures de même taille mais contenant des gènes pris aléatoirement dans la liste ordonnée. Pour chacune d'elles un score ES est calculé, la significativité est alors donnée par la position de ES dans la distribution des ES (ES_{null}).

Le score ES et la distribution ES_{null} sont normalisés pour pallier les effets dus aux nombre de signatures dans l'analyse mais aussi pour tenir compte de la possible redondance d'information dans les différentes signatures analysées.

Équation 11:

$$NES = \frac{ES}{moyenne\ des\ ES_{null}},\ ES\ et\ ES_{null}\ étant\ de\ même\ signe$$

Les ES_{null} sont normalisés de la même manière (NES_{null}). On peut alors définir le taux de faux positifs (FDR q-value) pour un groupe de signatures S ayant des NES positifs par :

Équation 12:

$$q\text{-}value = \frac{card\{S|NES_{null} \geq 0\ et\ NES_{null} \geq NES\}}{card\{S|NES \geq 0\ et\ \geq NES\}}$$

Pour les signatures de S ayant des NES négatifs, il suffit de remplacer ≥ par ≤ dans l'Équation 12:.

Les signatures obtenues par ICA sont donc ainsi évaluées pour la question biologique d'intérêt. Les signatures statistiquement significatives seront donc fortement liées aux modifications survenues suite à l'injection du vecteur chez la souris et auront les scores NES les plus extrêmes.

Tenir compte de la variabilité inter-individuelle

Évaluer statistiquement l'implication d'une signature dans un processus biologique particulier est une chose, tenir compte de la variabilité inter-individuelle en est une autre. Les signatures sont certes intéressantes mais l'évaluation de leur intérêt est basée sur la liste préalablement ordonnée

des gènes du jeu de données. L'ordre est fixé par une statistique qui évalue la différence des moyennes d'expression en tenant compte d'une variance calculée sur l'ensemble du jeu de données (McCarthy and Smyth, 2009). Mais, nous l'avons précédemment, une moyenne n'est pas nécessairement représentative de la structure d'un jeu de donnée. Nous contournons donc ce sujet sensible par l'introduction d'une étape supplémentaire dans la méthodologie générale. Elle consiste à répéter les phases de contextualisation biologique des signatures, non plus par l'utilisation d'une liste ordonnée de gènes uniquement, mais sur un ensemble de listes ordonnées gènes, définies après une étape de *bootstrap*. Le *bootstrap* est une méthode qui permet d'induire les caractéristiques d'une série d'échantillons par l'analyse de sous-ensemble de ces échantillons, on parle alors de technique d'inférence statistique (Efron, 1979). Elle permet dans notre cas de tester des combinaisons d'échantillons pour revoir l'ordonnancement de la liste de gènes du jeu de données. Pour cela, nous générons un ensemble de sélection aléatoire avec remise d'échantillon de même taille que le nombre d'éléments initiaux du groupe expérimental. Les différentes sélections générées sont ensuite associées pour former des paires de conditions expérimentales de type témoins *versus* vaccinés à partir desquelles nous établissons de nouvelles listes ordonnées. Nous obtenons ainsi un ensemble de listes qui prend en compte la variabilité biologique au sein des groupes comparés. Dans notre étude, nous définissons cent listes ordonnées pour un même jeu de données qui servent de support à la contextualisation des signatures extraites par ACI. La signification d'une signature pour caractériser l'injection d'un vecteur est alors fournie par cent valeurs de NES contre une seule auparavant. De cette manière nous évaluons la stabilité de l'information portée par la signature pour caractériser l'événement biologique d'intérêt.

Classer l'information

L'étape suivant consiste à comparer l'information portée par les signatures pour les ordonner, c'est à dire, chercher parmi les signatures, celles qui classifient aux mieux la condition contrôle et la condition vaccinée. En d'autres termes, nous voulons évaluer la capacité d'une signature à prédire le changement d'état transcriptionnel entre les deux conditions. Là aussi, il existe plusieurs méthodes pour répondre à cette question et celle que nous avons retenue ici est celle développée par l'algorithme dit de la forêt aléatoire (*random forest*) (Liaw and Wiener, 2002).

L'étude CompuVac comporte dix-neuf jeux de données de type groupe contrôle *versus* groupe vacciné. Chaque jeu de données fourni son lot de signatures, une dizaine environ, pour un total de

210 signatures. Ces signatures sont systématiquement testées sur les cents listes ordonnées produites pour chacun des dix-neuf jeux de données. Le bilan est donc de 1900 NES par signature au travers de l'ensemble des listes ordonnées. D'une certaine manière, la méthode de la forêt aléatoire procède comme pour le *bootstrap*, elle ne se contente pas d'évaluer le caractère prédictif d'une signature pour un ensemble de jeu de données, elle établit aussi la robustesse de la prédiction en effectuant un sous échantillonnage des données.

Mais revenons un peu en arrière, l'algorithme se base sur le principe de la construction d'arbres de décision. Nous définissons deux classes, une contenant les groupes vaccinés avec des vecteurs induisant des réponses immunitaires fortes et l'autre rassemblant les vecteurs induisant une faible réaction immunitaire. L'appartenance à telle ou telle classe se base sur les mesures d'activation des lymphocytes T gp33-44-spécifiques effectuées 5, 7 et 10 jours après l'injection des vecteurs dans une série de souris. Les classes étant définies, il s'agit maintenant de laisser l'algorithme sélectionner les signatures dont le comportement à travers les listes ordonnées de gènes est le plus à même de différencier les deux classes de vecteurs. Chaque signature est alors testée pour sa capacité prédictive et la signature qui produit le moins d'erreurs dans la classification est sélectionnée. Les données sont séparées en deux groupes, l'un dont les données sont caractérisées par une signature possédant des NES supérieurs à un certain seuil, l'autre ayant des NES inférieurs à ce seuil. À moins d'avoir une signature qui ne produit aucune erreur, ces deux groupes étant constitués de données des deux classes de vecteur. L'étape suivante consiste alors à chercher de nouvelles signatures qui vont discriminer les deux classes dans ses nouveaux sous-ensembles, et ainsi de suite. Au final, un arbre de décision est construit pour lequel les feuilles sont des groupes purs, c'est-à-dire constitués de données d'une seule et même classe. L'inconvénient majeur de cette technique est qu'elle est particulièrement liée à au jeu de données testé engendrant un biais désigné par le terme anglais *overfitting,* l'arbre risque de n'être bon que pour la prédiction des données du jeu qui a servi à son entraînement, et cela ne nous intéresse évidemment pas.

Entre alors en jeu la particularité de la forêt aléatoire, et qui lui vaut d'ailleurs son nom, celle de ne pas se contenter de créer un arbre mais plusieurs (une forêt donc) à partir d'une sélection aléatoire d'une partie des données du jeu de données d'origine. Le processus décrit plus haut va se produire pour chacun des sous-ensembles sélectionnés, générant un arbre de décision à chaque fois. La détermination du caractère prédictif des signatures se fait non plus sur la base d'un arbre seul mais à travers les arbres. Une signature qui se voit systématiquement placée en haut des arbres crées,

possède de toute évidence une forte capacité à discriminer les deux classes, et ce quelle que soit la composition initiale du sous-ensemble. Elle n'est donc pas ou peu influencée par une partie limitée des données qui lui conférerait sa capacité prédictive. C'est une fois de plus une manière de prendre en compte la variabilité des données dans l'évaluation de la caractéristique de prédiction d'une signature et donc de minimiser l'*overfitting*. L'erreur de prédiction de l'arbre est calculée après prédiction de la partie du jeu de données qui n'a pas été utilisée, le *out of bag* (OOB). Bien entendu, le résultat final de cette méthode n'est plus un arbre qu'il est possible de visualiser, mais une forêt d'arbre, donc sans représentation graphique. Afin de déterminer le classement des signatures, il est possible de calculer un indice d'importance, au sens d'importance pour la bonne prédiction des données. Cet indice évalue la différence de qualité de prédiction de chaque arbre lorsqu'on permute les prédicteurs, à savoir les signatures. Une signature en haut de l'arbre peut donc se retrouver en bas et vice-versa. Si l'emplacement des signatures est important pour la prédiction alors ce changement doit avoir une répercussion importante elle aussi sur la précision de la classification. Pour chaque permutation un score, correspondant à la différence d'efficacité de prédiction des arbres avant et après permutation est calculé et associé à la signature qui a été permuté. Les scores sont moyennés à travers la forêt et normalisés pour chaque signature en pondérant la moyenne par l'écart-type des scores qui lui sont attribués.

L'analyse des données CompuVac nous a permis de créer un modèle prédictif de la réponse immunitaire suite à une vaccination, chez la souris, et ce, à partir des transcriptome de cellules dendritiques triées de la rate seulement six heures après l'injection du vaccin. Ce modèle a confirmé son efficacité en prédisant correctement des vaccinations faites avec des vaccins efficaces mais de souches vectorielles inconnues du modèle. Il l'est aussi pour la prédiction faite à partir de jeu de transcriptome de rate totale et même de sang périphérique, des systèmes biologiques plus complexe que les cellules triées. Enfin, il prédit aussi efficacement un jeu de données humain tirés de PBMC, confirmant les résultats publiés quant à l'activation du système immunitaire après une vaccination (Zak et al., 2012). En faisant en sorte de tenir compte de la variabilité inter-individuelle à tous les niveaux de construction de notre modèle, nous rendons celui-ci prompt à détecter les variations les plus fines dans des jeux de données beaucoup plus complexes que ceux servant à son entraînement. La Figure 25 montre, par ailleurs la différence de classification des vecteurs obtenue sur la base des NES issus des données originales et après *bootstrap* ; L'impact du *bootstrap* est

particulièrement évident en donnant une classification nette de trois groupes de vecteurs dans le second cas quand il est difficile de dégager des groupes dans le premier cas.

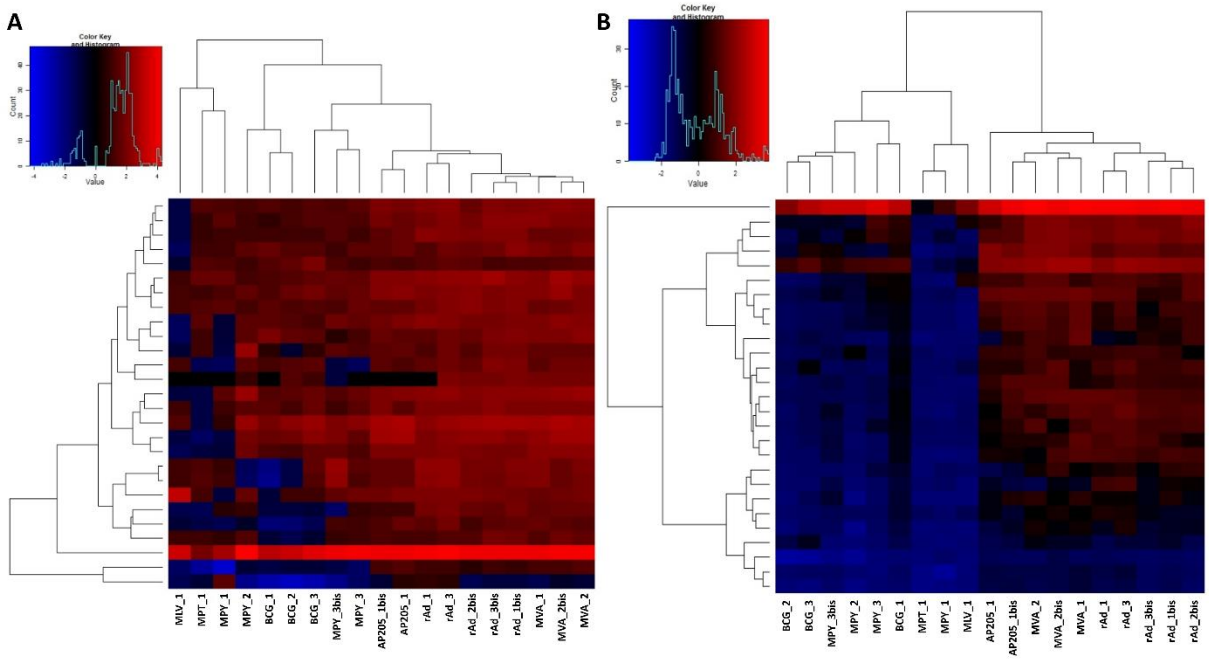


Figure 25. Effet du bootstrap sur la classification des vecteurs.

Heatmap des valeurs de NES de 27 signatures sélectionnées obtenues sur le jeu de données initial (gauche) et à partir de la médiane des NES obtenus suite au bootstrap (droite). Source : Derian et al, 2016

Les méthodes statistiques exposées ci-dessus sont efficaces pour décrire les données et tenir compte de la variabilité inter-individuelle ; elles ne rendent néanmoins pas compte du comportement de cette dernière. Il est un domaine scientifique où l'intérêt des distributions d'échantillons biologiques est au cœur même du métier, c'est l'écologie. Voyons maintenant comment on peut utiliser les outils développés par les écologues pour analyser les données transcriptomiques.

LES INDICES DE DIVERSITÉ

L'écologie est la science étudie des écosystèmes, c'est-à-dire la structure temporo-spatiale des individus associés à leur environnement (voir la définition dans l'Encadré « Écosystème »). Les écologues étudient non seulement la répartition des groupes d'individus dans un espace mais aussi les interactions qui existent entre eux.

Les indices de diversité sont un moyen d'appréhender la structure de répartition des groupes d'individus dans un espace donné. Nous désignons les groupes d'individus par le terme d'espèce, chaque espèce étant constituée d'un ensemble homogène d'individus pour une caractéristique donnée, cela peut-être génétiquement parlant mais tout aussi bien géographiquement parlant. La mesure de la diversité est alors une mesure de la moyenne de la rareté des espèces dans un environnement donné (Moine, 2002). Une communauté est un ensemble d'espèces dans un environnement donné et est caractérisé par le nombre d'espèces et le vecteur d'abondance associé. Le vecteur d'abondance est défini comme le vecteur des fréquences des espèces, plus une espèce est abondante, plus sa fréquence relative au sein des espèces est importante. L'étude de ces deux paramètres d'une communauté, nombre d'espèce et abondance de ces dernières, donne les règles suivantes :

- Une communauté est d'autant plus diverse qu'elle contient un grand nombre d'espèces.
- Une communauté est d'autant plus diverse que la distribution d'abondance des espèces est proche d'une distribution équilibrée, c'est-à-dire que l'abondance des espèces est similaire.

Il en découle que pour deux communautés possédant des distributions d'abondance égale, la diversité sera d'autant plus forte que le nombre d'espèces sera grand.

Il existe trois grands types de diversité à observer, la diversité α, la diversité β et la diversité γ (voir la définition dans l'Encadré « Diversité α, β et γ »).

Écosystème. *Système formé par un environnement (biotope) et par l'ensemble des espèces (biocénose) qui y vivent, s'y nourrissent et s'y reproduisent (Larousse)*. C'est finalement tout ou partie d'un ensemble biologique cohérent en terme d'environnement et d'interaction. Les différents éléments de l'écosystème, c'est-à-dire les espèces qui le composent, participent au maintien d'un équilibre global du système. Un écosystème est sensible par essence car tous les éléments qui le composent entretiennent des liens de cause à effet très étroits. Une perturbation sur l'un de ces éléments aura des répercussions sur l'ensemble du système.

MESURES D'ENTROPIE ET VRAIE DIVERSITÉ

Les indices de diversité sont des outils mathématiques pour mesurer la diversité des échantillons observés. Les indices qui nous intéressent ici font partie d'une famille d'indices non paramétriques. Ils prennent la forme de fonctions monotones de la forme $\sum_{i=1}^{S} p_i^q$ ou de limites cette fonction (Jost, 2006). L'utilisation des indices tels quels est problématique car ils sont difficilement interprétables en l'état, certains n'étant d'ailleurs pas des mesures de la diversité du système mais de son entropie, comme l'indice de Shannon défini comme suit :

Équation 13:

$$H \equiv -\sum_{i=1}^{S} p_i \ln p_i$$

Où S est le nombre d'événements observés dans l'échantillon et p_i la fréquence relative de l'événement i. L'entropie de Shannon est très utilisée en théorie de l'information pour mesurer la quantité d'information de toutes sortes de sources (physique, informatique, …). On la retrouve d'ailleurs dans l'algorithme *fastICA* décrit plus haut pour l'estimation du caractère non-gaussien des sources. L'entropie de Shannon d'une source est ainsi comparée à l'entropie de Shannon d'une source gaussienne de même moyenne. Plus la différence est grande, plus les sources sont dissemblables en termes d'organisation, et par conséquent, plus la source testée est différente d'une gaussienne. Dans le cas de *fastICA*, tel que nous l'utilisons, plus la différence d'entropie est grande, plus la source prend la forme d'une super-gaussienne.

L. Jost explique dans son article que si cette famille d'indice de diversité peut être généralisée par une entropie, il serait trompeur de considérer l'entropie comme la diversité car le comportement

mathématique de l'entropie n'est pas en adéquation avec l'intuition biologique du résultat d'une expérience.

L'expérience suivante est tirée de (Jost, 2006)) : imaginons deux communautés de cinquante espèces d'oiseaux n'ayant pas d'espèce en commun mais dont l'abondance des espèces est identique. La diversité totale des deux communautés rassemblées est donc de cent espèces. Nous souhaitons connaître la proportion de la diversité totale (les cent espèces) contenue dans la moyenne des communautés. Regardons les résultats proposés par trois indices connus :

La richesse, qui correspond au nombre d'espèce présent dans les communautés nous donne très logiquement 0.5, c'est-à-dire qu'une communauté représente cinquante pourcents de la diversité totale. L'indice de Shannon (Équation 13) quant à lui nous donne un résultat de 0.85 tandis-que l'indice de Gini-Simpson ($x = 1 - \sum_{i=1}^{S} p_i^2$) nous donne 0.99. Les deux derniers indices nous donnent donc des résultats proches de 1, c'est-à-dire indiquant de très fortes similitudes entre les deux communautés, alors même qu'elles sont parfaitement différentes.

Jost propose donc de convertir les indices de diversité en ce qu'il appelle la vraie diversité, c'est-à-dire celle dépendante de p_i. La diversité est alors interprétée en termes de nombre d'espèces dans l'échantillon. La transformation des indices se généralise par l'équation suivante :

Équation 14:

$$^{q}D \equiv \left(\sum_{i=1}^{S} p_i^q\right)^{1/(1-q)}$$

À partir de cette formule, voyons maintenant quelques cas d'intérêt pour différentes valeurs de q :

- Si $q = 0$, alors $^{0}D = \sum_{i=1}^{S} p_i^0$ () qui est la richesse, donc le nombre d'espèce dans l'échantillon.
- ^{q}D n'est pas défini pour $q = 1$, mais sa limite vaut $^{1}D = exp\left(-\sum_{i=1}^{S} p_i \ln p_i\right) = exp(H)$, où H est l'entropie de Shannon.
- Lorsque que $q = 2$, $^{2}D = 1/\left(\sum_{i=1}^{S} p_i^2\right)$ soit l'inverse de la concentration de Simpson (Table 2).

Les indices utilisés plus haut sont donc tous présents ici et transformés pour être comparables les uns avec les autres. L'utilisation de cette transformation sur l'exemple des deux communautés d'oiseaux donne cette fois-ci des résultats similaires pour les trois indices utilisés.

Table 2. Les indices de diversité (source : Jost *et al*).

Index x:	Diversity in terms of x:	Diversity in terms of p_i:
Species richness $x \equiv \sum_{i=1}^{S} p_i^0$	x	$\sum_{i=1}^{S} p_i^0$
Shannon entropy $x \equiv -\sum_{i=1}^{S} p_i \ln p_i$	exp(x)	$\exp\left(-\sum_{i=1}^{S} p_i \ln p_i\right)$
Simpson concentration $x \equiv \sum_{i=1}^{S} p_i^2$	1/x	$1/\sum_{i=1}^{S} p_i^2$
Gini–Simpson index $x \equiv 1-\sum_{i=1}^{S} p_i^2$	$1/(1-x)$	$1/\sum_{i=1}^{S} p_i^2$
HCDT entropy $x \equiv \left(1-\sum_{i=1}^{s} p_i^q\right)/(q-1)$	$[(1-(q-1)x)]^{1/(1-q)}$	$\left(\sum_{i=1}^{S} p_i^q\right)^{1/(1-q)}$
Renyi entropy $x \equiv \left(-\ln \sum_{i=1}^{s} p_i^q\right)/(q-1)$	exp(x)	$\left(\sum_{i=1}^{S} p_i^q\right)^{1/(1-q)}$

Faire varier l'ordre q de l'Équation 14 est une façon de calculer la diversité à différent degré de sensibilité pour les espèces rares. Quand q vaut zéro, l'ensemble des espèces, quelques soient leur abondance, est pris en compte. Quand q vaut un, les fréquences relatives p_i sont multipliées par leurs logarithmes, augmentant ainsi le poids des espèces à très faibles abondance. Enfin, quand q vaut deux, les p_i sont multipliés par eux-mêmes limitant cette fois le poids des espèces ayant une abondance faible.

Diversité α, β et γ. Les différentes nominations de la diversité (α, β et γ) impliquent le niveau d'organisation que l'on souhaite observer. La diversité α est la diversité d'un système écologique cohérent, il y est donc question de nombre d'espèces et de leur abondance. La diversité β se concentre sur la modification de la diversité α quand on passe d'un écosystème à un autre, on y étudie donc les différences entre les distributions d'espèces. Enfin, la diversité γ est l'étude de la diversité de systèmes écologiques différents, les outils sont les mêmes que pour l'étude de la diversité α mais appliqués des composantes organisationnelles plus importante. Prenons l'exemple d'un lac de montagne, la diversité α consiste alors à étudier la diversité en termes de composition d'espèce de ce lac. La diversité β comparera la diversité de ce lac avec celui d'à côté quand la diversité γ englobera les deux lacs dans un seul et même système, celui des lacs de montagne de la région A. Mais finalement ce dernier mode organisationnel peut lui aussi être comparé à un autre mode englobant les lacs de la région B, etc.

APPLICATION AU TRANSCRIPTOME : UN CAS BIBLIOGRAPHIQUE

Étudier la diversité des transcrits d'un échantillon biologique comme on étudie la diversité des espèces d'un étang ou d'une prairie est une idée des plus intéressante car le transcriptome présente des caractéristiques des écosystèmes. Il est constitué de transcrits, assimilable à des espèces différentes, en étroites relations les uns avec les autres par des liens directs ou indirects. C'est aussi un système dynamique et résilient, les phases de relaxations succèdent aux phases d'activités intenses. Pour ces raisons, il est tentant de considérer le transcriptome comme un écosystème décrivant l'activité des gènes d'un échantillon donné. Un bel exemple de cette vision est le papier de O. Martinez et M. H. Reyes-Valdés publié en 2008 (Martínez and Reyes-Valdés, 2008). Les auteurs y décrivent comment ils ont utilisé l'indice de Shannon pour analyser la diversité du transcriptome à travers plusieurs tissus. L'indice de Shannon est une mesure d'entropie issue de la théorie de l'information, elle consiste à quantifier une quantité d'information dans une source. Plus la source contient d'information, plus la mesure d'entropie va être forte, ce qui se traduit pour le transcriptome par le déséquilibre des abondances des transcrits. En d'autres termes, si tous les transcrits sont exprimés aux mêmes niveaux, et ont par conséquence des abondances identiques, alors l'entropie mesurée est maximale car l'information portée par le système est maximale. À l'inverse, un tissu qui ne verrait qu'un seul gène s'exprimer, aurait pour mesure de son entropie nulle. L'entropie de Shannon est donc directement liée au vecteur d'abondance du système observé, ici un tissu. L'entropie de Shannon pour un tissu j vaut alors :

Équation 15:

$$H_j = -\sum_{i=1}^{g} p_{ij} \log_2(p_{ij})$$

Où p_{ij} est la fréquence relative du gène i dans le tissu j. H_j représente donc la quantité d'information portée par le transcriptome du tissu et peut être comparée à la quantité portée par les autres tissus.

Les auteurs définissent un autre indice, non utilisé par les écologues, l'indice de spécificité d'un gène. Il s'agit ici d'évaluer comment un gène se comporte, non plus dans un tissu, mais à travers les tissus. Pour cela il faut d'abord poser p_i, la moyenne des abondances d'un gène à travers les tissus :

Équation 16:

$$p_i = \frac{1}{t}\left(\sum_{j=1}^{t} p_{ij}\right)$$

La spécificité d'un gène prend alors la forme d'une entropie de Shannon adaptée pour prendre en compte l'expression des gènes à travers les tissus et pondérée par le nombre de tissus observés :

Équation 17:

$$S_i = \frac{1}{t}\left(\sum_{j=1}^{t} \frac{p_{ij}}{p_i} \log_2\left(\frac{p_{ij}}{p_i}\right)\right)$$

La valeur S_i est maximale, et vaut $\log_2(t)$, quand un gène est exprimé de manière exclusive dans un tissu, alors qu'elle est nulle si le gène est exprimé de manière équitable. Il est possible de ramener cette information au niveau de chaque tissu en multipliant le S_i d'un gène par son abondance dans un tissu et ainsi définir la spécialisation du tissu comme étant la somme de ces produits :

Équation 18:

$$\delta_j = \sum_{i=1}^{g} p_{ij} S_i$$

Un tissu est alors d'autant plus spécialisé qu'il contient des gènes avec une forte spécificité.

Afin de fixer les idées sur le comportement des indices, les auteurs propose l'exemple simple suivant : l'analyse de l'expression de quatre gènes dans quatre tissus différents (Figure 26).

Regardons la diversité tout d'abord, un seul gène sur les quatre s'exprime dans l'échantillon a, tandis que trois gènes sont exprimés de manière équivalente dans l'échantillon d. L'échantillon a possède une diversité nulle tandis que l'échantillon d est proche de la diversité maximale (respectivement à gauche et à droite sur l'axe des abscisses). Les deux autres échantillons ont des diversités intermédiaires. Concernant la spécialisation, le gène 1 n'est exprimé que dans le tissu a et le gène 3 est exprimé de manière équivalente dans trois tissus. Le gène 1 est donc hautement spécialisé, et comme le tissu a n'exprime que ce gène, sa spécialisation est maximale et vaut $\log_2(4) = 2$ (en haut sur l'axe des ordonnées). À l'inverse, le gène 3 est très faiblement spécialisé et participe donc peu à la spécialisation des tissus b, c et d.

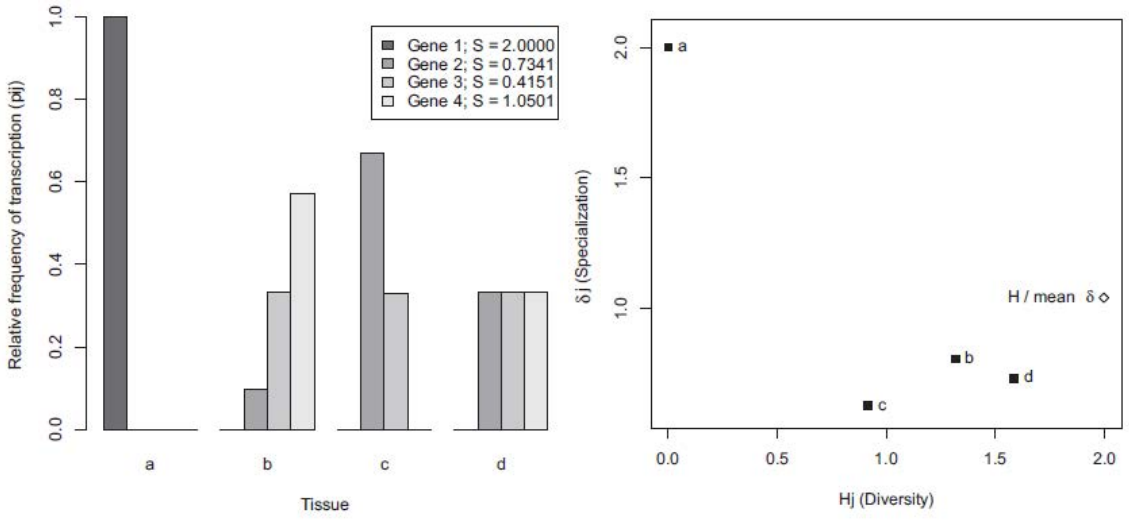


Figure 26. Diversité et spécificité.

Calcul la diversité (entropie de Shannon) et de spécificité sur des jeux de données théoriques : 4 tissus exprimant 4 gènes. Les gènes sont exprimés de manières différentes et à des intensités différentes selon les organes (e.g. le gène 1 est exprimé exclusivement dans le tissu A). Source : Martinez et al, 2008.

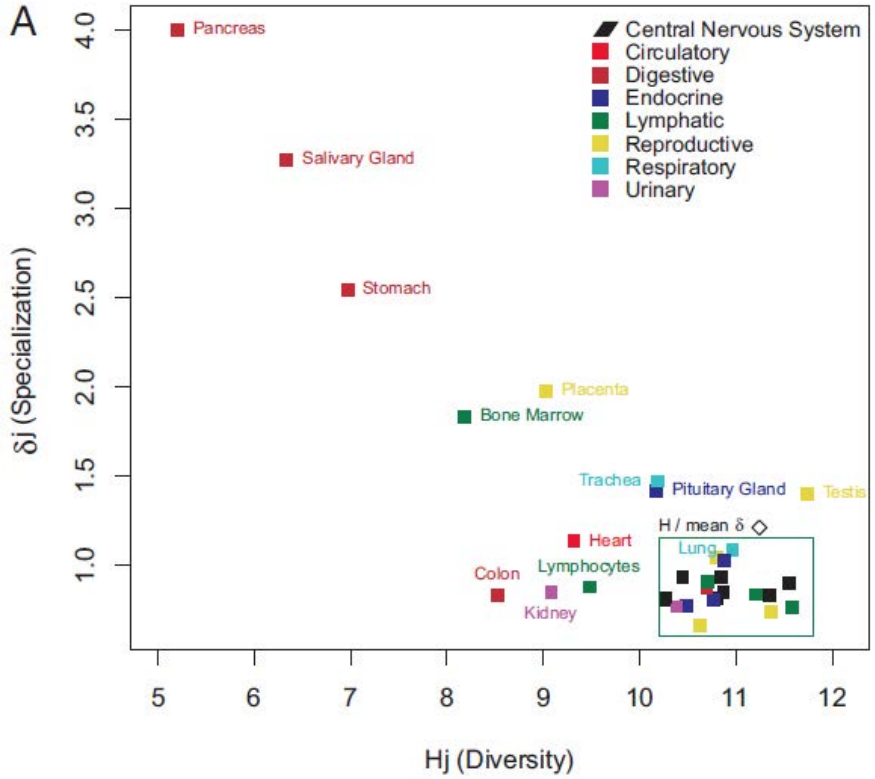


Figure 27. Diversité et spécialisation des organes humains au niveau transcriptomique.

La diversité de chaque organe est calculée par l'entropie de Shannon. La spécialisation est la somme des spécificités des gènes au sein de chaque organe. Le losange noir représente la diversité du système comprenant l'ensemble des organes et la moyenne des spécialisations de ces organes. Source : Martinez et al, 2008

Calculées sur des données de transcriptome, les résultats montrés en Figure 27 indiquent que le pancréas est l'organe le moins divers et le plus spécialisé de l'organisme, suivi par les glandes salivaires et l'estomac.

Nous avons vu précédemment qu'il est possible de calculer la variation de l'expression d'un gène par des outils statistiques classiques comme l'écart-type, voire même mieux, par le coefficient de variation car ce dernier n'est pas sensible à l'intensité d'expression. Qu'apporte alors l'indice de spécificité par rapport au coefficient de variation ?

SPÉCIFICITÉ ET COEFFICIENT DE VARIATION

Le coefficient de variation donne une variation autour de la moyenne et augmente au fur et à mesure que les points s'éloignent de cette moyenne. L'indice de spécificité va en revanche maximiser quand un gène est exprimé dans un seul échantillon et ce quel que soit l'amplitude de l'expression de ce dernier. Il se trouve que dans les données de transcriptome produit par puce à ADN, il n'est pas possible de maximiser la spécificité car il est toujours associé une valeur d'expression à un gène, si petite soit-elle. Dans ce cas la spécificité d'un gène est directement liée à l'amplitude d'expression qui le différencie des autres échantillons. Les valeurs faibles induisent en quelque sorte du bruit qui applique une certaine incertitude à la valeur de spécificité. En résumé, plus un gène est différemment exprimé dans un échantillon par rapport aux autres, plus il verra sa spécificité augmenter. Il verra d'ailleurs son coefficient de variation augmenter lui aussi. Afin de comprendre la relation qui unie ces deux calculs, j'ai simulé un jeu de données de transcriptome en utilisant la librairie *quantroSim* (Hicks and Irizarry, 2014) disponible pour le logiciel R. L'algorithme procède en deux étapes :

Tout d'abord il simule l'expression de gènes telle qu'elle est mesurée sur des puces à ADN. Il respecte ainsi des règles statistiques pour la distribution des données. L'expression des gènes est définie selon distribution de Poisson ($p(k) = \frac{\lambda^k}{k!} e^{-\lambda}$) dont la distribution des logarithmes des valeurs suit une loi normale $\mathcal{N}(\mu, \sigma)$. Une quantité d'ARN pour un gène donné est tirée de manière aléatoire d'une distribution de Poisson ayant pour moyenne d'occurrence une valeur λ comprise entre 0,01 et 4662,66.

```
Library(quantroSim)
set.seed(999)
geneTruth <- simulateGExTruth(nGenes = 25000, nGroups = 1, pDiff = 0)
```

La deuxième étape consiste à utiliser les valeurs d'expression générées à la première étape pour simuler des jeux de données de fabricants et donc décomposer les valeurs d'expression de gènes en valeurs de détection de sondes pour n échantillons.

```
sim <- simulateGEx(geneTruth, GEx.platform = "GExArrays", nSamps = 10)
```

La représentation par les boîtes à moustaches donne un aperçu de la distribution des données générées pour une simulation de dix échantillons de 25000 gènes (Figure 28). Les boîtes à moustaches sont typiques des données de transcriptome où beaucoup de gènes se rassemblent autour d'une valeur assez faible et quelques gènes possèdent des valeurs d'expression très fortes.

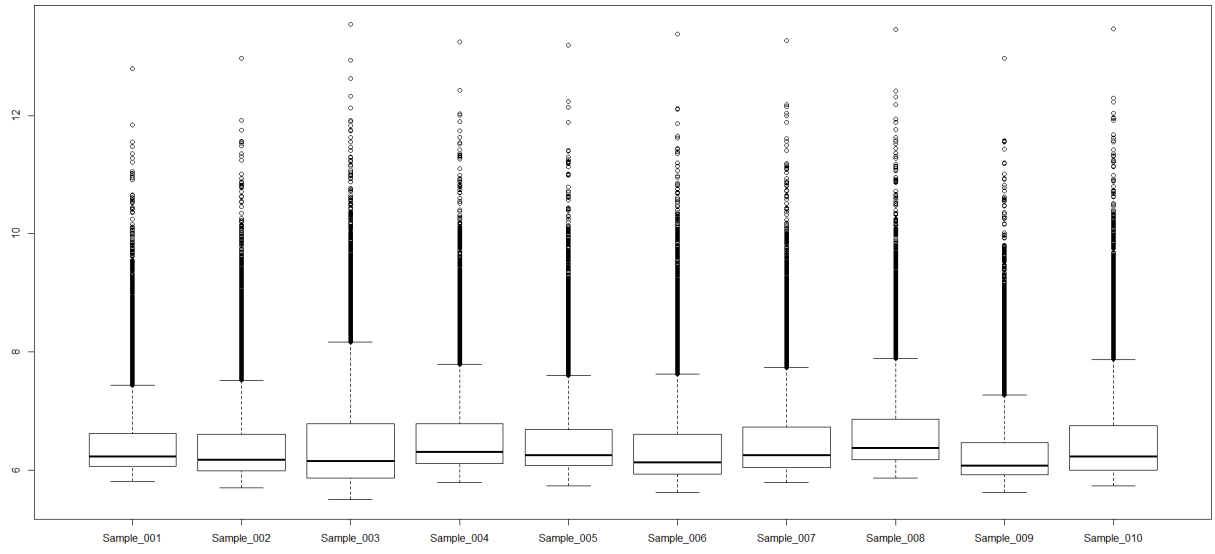


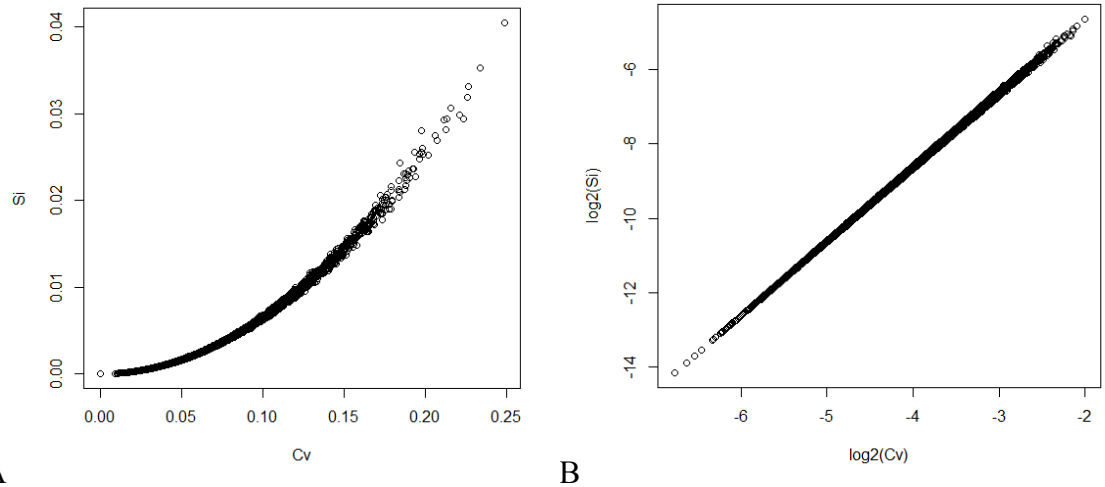
Figure 28. Distributions des données simulées par *quantroSim*.

Simulation de dix jeux échantillons de 25000 gènes par l'algorithme *quantroSim*. Distribution des expressions par la représentation des boîtes à moustaches. Les boîtes représentent la distribution de 50% des valeurs et la barre noire indique la médiane. Les traits en pointillés de chaque côté des boîtes représentent la distribution des valeurs les plus fortes quand elle est au-dessus et des valeurs les plus basses quand elle est en-dessous. Les cercles noirs représentent les valeurs atypiques.

Le jeu de données est normalisé par la méthode des quantiles et les gènes sont analysés pour connaître leur spécificité et leur coefficient de variation. Fait important, les données de transcriptome sont en général log-transformées afin de contrôler la variance et donner une allure

gaussienne aux données, prérequis à certains tests statistiques. Dans notre cas, cette transformation ne me semble pas judicieuse car elle a pour effet de minimiser l'impact des valeurs les plus fortes et renforcer l'impact des valeurs faibles. Elle diminue donc l'ampleur de la variation observée, ce qui va à l'encontre de ce que nous voulons étudier.

En représentant la spécificité en fonction du coefficient de variation, nous constatons qu'il existe une relation mathématique entre les deux indices de type puissance. Cela est confirmé en procédant à la même représentation avec les indices log-transformés. En effet, les données sont alors parfaitement alignées sur une droite (Figure 29).



A B

Figure 29. Relation entre spécificité et coefficient de variation #1.

Représentation des indices de spécificité par rapport au coefficient de variation pour un jeu de données simulé de 10 échantillons contenant 25000 gènes (A). Même représentation pour les valeurs transformées par log2 des deux indices.

Je cherche alors la courbe qui explique au mieux la relation entre les deux indices pour pouvoir en déduire ses caractéristiques. Pour cela j'utilise la fonction *nls()*, pour *nonlinear least squares*, de la librairie *stats* et applique aux données une fonction de type puissance telle que :

Équation 19:

$$Si = aCv^b$$

La fonction va alors chercher les valeurs de a et b qui expliquent au mieux les données. Le résultat est représenté en Figure 30.

```
model<-nls(Si~a*Cv^b,start=list(a=0.5,b=2))
```

L'algorithme nous donne alors les valeurs de $a = 0{,}63$ et $b = 1{,}99$. Par la suite, j'applique à ce jeu de données quelques modifications afin de tester le comportement des deux indices.

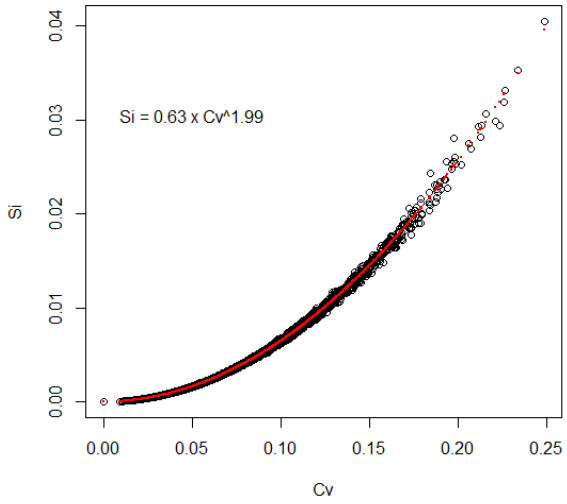


Figure 30. Relation entre spécificité et coefficient de variation #2.
Représentation des indices de spécificité par rapport aux coefficients de variation pour un jeu de données simulé de 10 échantillons contenant 25000 gènes. Les points rouges correspondent à la prédiction des valeurs de spécificité des gènes compte tenu de leur coefficient de variation et l'équation de la courbe de régression définie par *nls()*, notée sur le graphique.

Pour chaque échantillon, je sélectionne aléatoirement 100 gènes dont je modifie l'intensité d'expression d'un facteur trois. Les gènes sélectionnés ne sont pas redondant entre les échantillons. Ainsi nous avons pour chaque échantillon un ensemble de cent gènes qui voient leur expression être trois plus importante que pour les neuf autres échantillons. Il s'agit bien évidement de simuler des cas où la spécificité des gènes serait très forte. L'opération est répétée pour des paires d'échantillon cette fois. Partant du jeu de données originel, cinq paires d'échantillons voient donc l'expression de cent de leurs gènes être trois fois plus important que pour les huit autres échantillons. Suivant le même mode opératoire, je modifie l'expression de cent gènes pour trois triplets, deux quadruplets et un quintuplet. Lorsque cela est nécessaire, pour l'étape des triplets et des quadruplets, les échantillons résiduels (respectivement un et deux échantillons) sont eux aussi modifiés afin de garder une homogénéité des distributions d'un échantillon à l'autre. Le jeu de données est normalisé par la méthode des quantiles et chaque gène modifié est analysé pour en connaître sa spécificité et son coefficient de variation. Les courbes de régression associées à la représentation des coefficients de variation par rapport aux spécificités sont affichées en rouge sur les graphiques de la Figure 31.

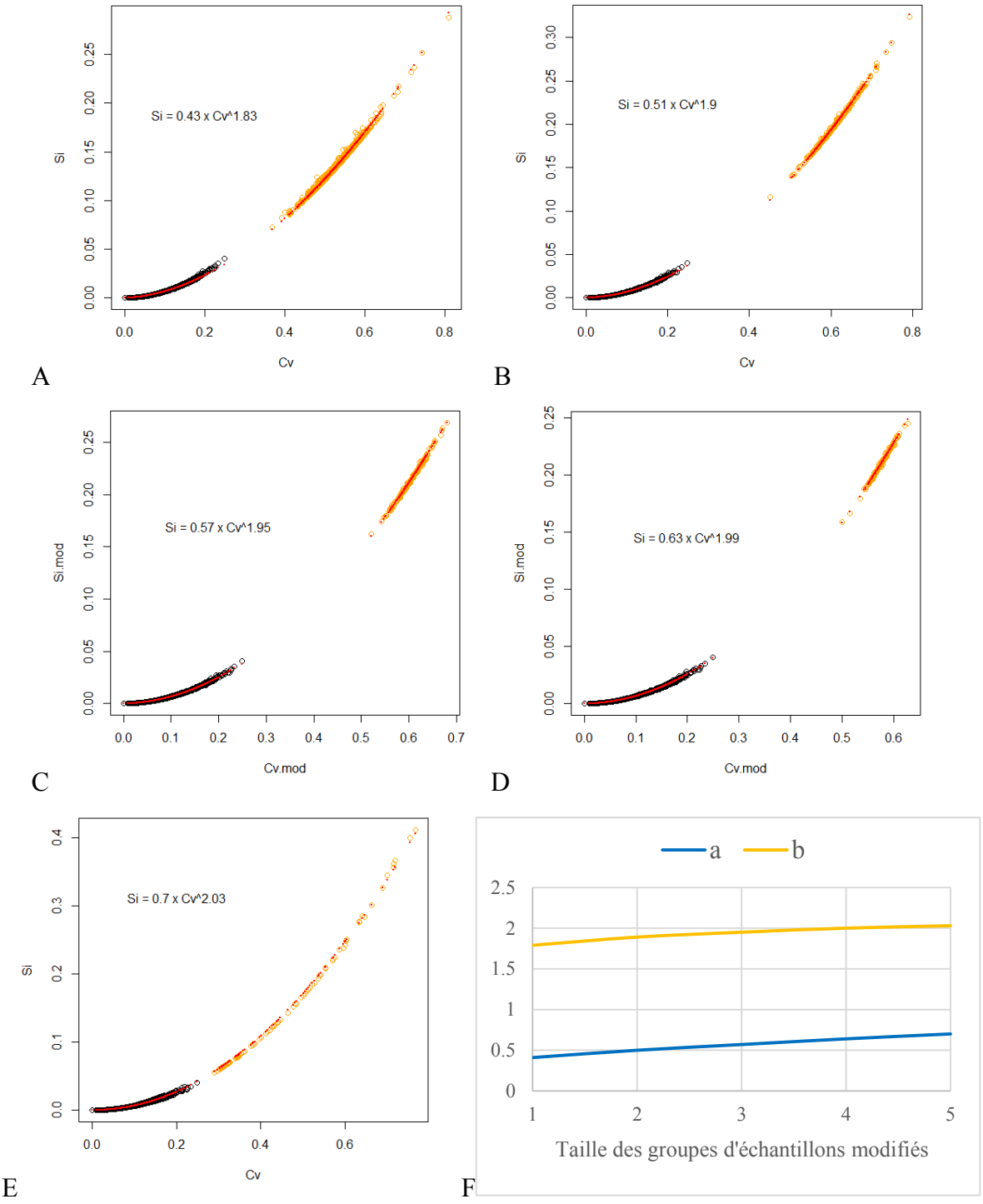


Figure 31. Relation entre spécificité et coefficient de variation #3.

Représentation des indices de spécificité par rapport aux coefficients de variation pour des jeux de données simulés et modifiés d'un facteur 3 pour l'expression de 100 gène dans un échantillon (A), des paires d'échantillons (B), des triplets d'échantillons (C), des quadruplets d'échantillons (D) et des quintuplets (E). En rouge, prédiction des indices de spécificité pour les coefficients de variation selon l'équation définie par *nls()*. Évolution des coefficients de la courbe de régression en fonction du nombre d'échantillons modifiés.

Quelle que soit la modification effectuée, sur un simple échantillon ou des paires ou des triplets, etc., la relation entre les deux indices est parfaitement stable et prend bien pour forme une équation du type $Si = aCv^b$. Les coefficients a et b évoluent faiblement mais de manière croissante avec le nombre d'échantillons impliqués dans la transformation ($a \in [0{,}51\ ;\ 0{,}68], b \in [1{,}9\ ;\ 2{,}02]$).
Regroupés sur le même graphique, les différentes transformations montrent bien comment l'indice de spécificité change l'ordonnancement des gènes par rapport au coefficient de variation (Figure 32). Pour un coefficient de variation identique, deux gènes peuvent avoir des spécificités différentes. L'indice de spécificité apporte donc une information supplémentaire à l'outil standard en matière d'analyse de variabilité.

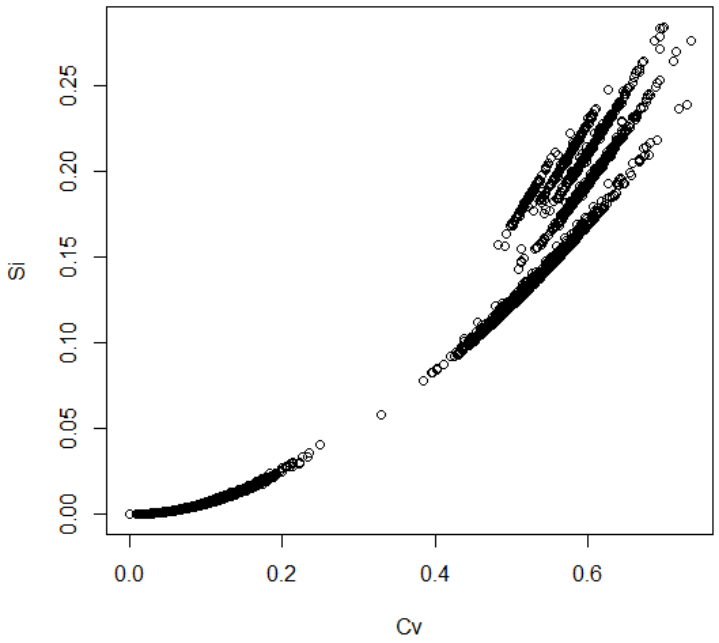


Figure 32. Relation entre spécificité et coefficient de variation #4.
Représentation des indices de spécificité par rapport aux coefficients de variation pour des jeux de données simulés et modifiés d'un facteur 3 selon plusieurs modalités : des échantillons uniques, des paires d'échantillons, des triplets, des quadruplets et des quintuplets.

Afin de nous assurer que l'indice de spécificité garde un lien étroit avec l'amplitude de la variation, j'ai répété la même expérience en procédant cette fois ci à des modifications d'expression d'un facteur cinq.

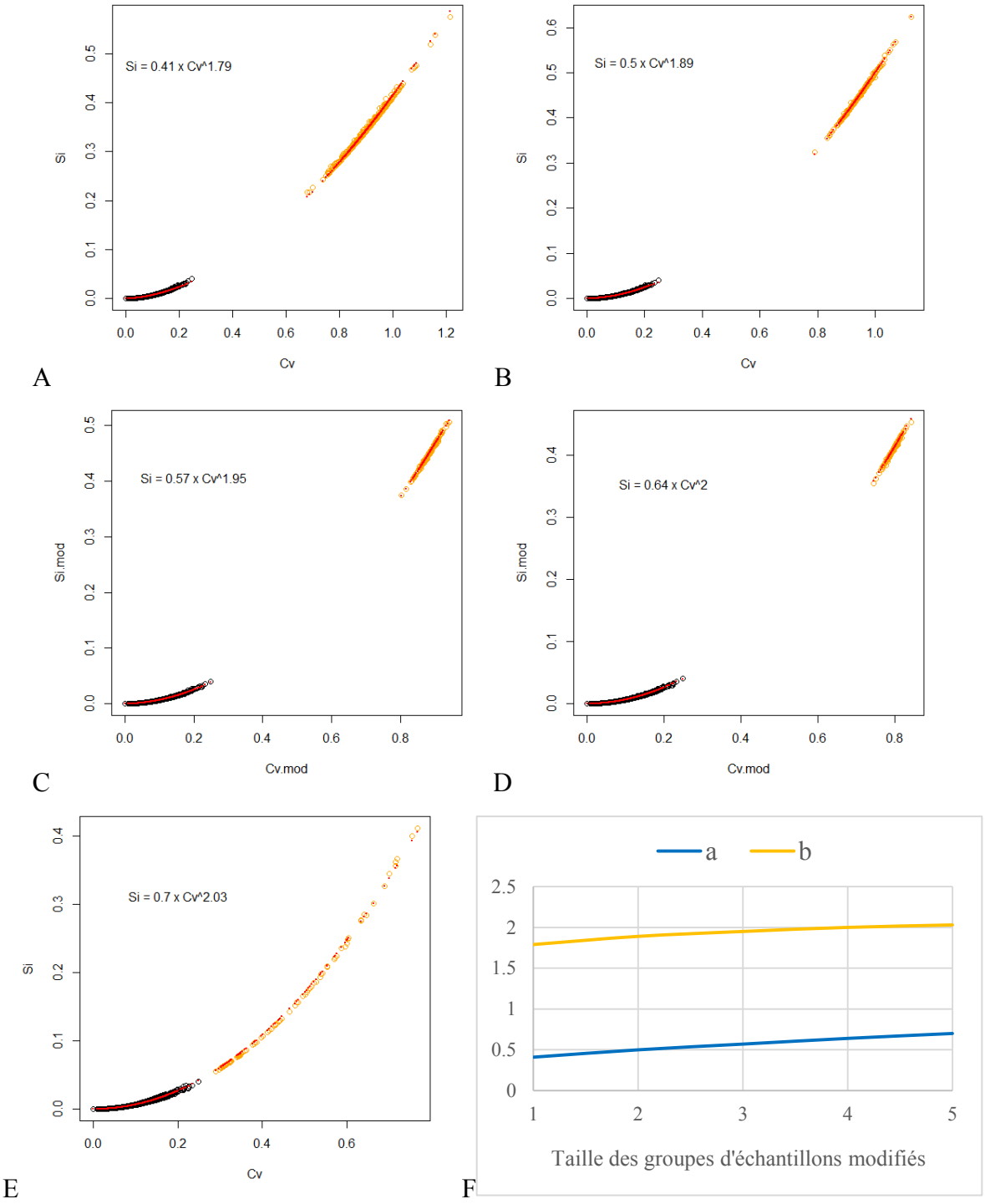


Figure 33. Relation entre spécificité et coefficient de variation #5.

Représentation des indices de spécificité par rapport aux coefficients de variation pour des jeux de données simulés et modifiés d'un facteur 5 pour l'expression de 100 gène dans un échantillon (A), des paires d'échantillons (B), des triplets d'échantillons (C), des quadruplets d'échantillons (D) et des quintuplets (E). En rouge, prédiction des indices de spécificité pour les coefficients de variation selon l'équation définie par *nls()*. Évolution des coefficients de la courbe de régression en fonction du nombre d'échantillons modifiés.

Les résultats présentés en Figure 33 indiquent que les paramètres a et b sont extrêmements similaires à ceux obtenus par une modification d'un facteur trois. Les seules modifications notables concernent les valeurs des indices qui se trouvent grandement augmentés pour les deux indices. L'amplitude de la modification est donc bien conservée par l'indice de spécificité.

Comme nous venons de le voir, les indices de diversité fournissent un ensemble intéressant de mesures permettant d'évaluer la diversité d'un jeu de données de transcriptome, à l'échelle individuelle mais aussi d'un groupe d'échantillons. Selon l'ordre de l'indice ($q \in [0, +\infty[$), nous pouvons apprécier cette diversité en tenant plus ou moins compte des transcrits de faible abondance. Cela offre à l'analyse de transcriptome une approche originale qui complète les analyses classiques.

La mesure de la spécificité est, quant à elle, particulièrement intéressante pour la compréhension de la structure d'un groupe d'échantillons. En plus des mesures de variance ou de coefficient de variation, elle permet de mettre en évidence des gènes aux comportements singuliers (expression individuelle ou dans des petits groupes d'échantillons) qui ne sont pas mis en avant par l'analyse du coefficient de variation (Figure 32).

R project : R est une plate-forme dédiée à l'analyse mathématique et statistique des données. Une communauté très active participe à l'élaboration de librairie, c'est-à-dire des codes informatiques assemblant les fonctions nécessaires pour effectuer une tâche précise. La librairie *quantroSim* est par exemple dédiée à la génération de données simulées de transcriptome et de méthylome.

TRAITEMENT DES DONNÉES

Utiliser les indices de diversité pour analyser le transcriptome c'est considérer que la mesure que nous faisons de l'abondance de transcrit est similaire à celle effectuée par les écologues dans un environnement donné. Or il y a une différence notable lorsque nous utilisons les puces à ADN : Nous connaissons la richesse maximale du système. En écologie, lorsqu'un chercheur effectue des prélèvements sur le terrain, il ne sait pas combien d'espèces différentes sont contenues dans son échantillon. Plus encore, il ne sait pas si le nombre obtenu est le nombre maximal d'espèces pour ce terrain. Il existe d'ailleurs tout un champ statistique pour estimer la richesse d'un environnement en fonction des prélèvements effectués (Marcon, 2015).

Avec les puces à ADN, nous n'avons pas ce problème, puisque nous savons combien d'espèces nous avons au maximum : c'est le nombre de sondes uniques présentent sur la puce. Ce nombre n'est pas représentatif de l'ensemble des transcrits qu'il est possible de trouver dans un échantillon mais on ne peut espérer voir quelque chose qu'on ne mesure pas. Dans ce cas, le nombre d'espèces maximal est fixé par la technologie que j'utilise pour mesurer l'abondance des transcrits.

La deuxième différence vient du fait qu'un écologue ne mesure pas ce qu'il ne capture pas, alors que les puces à ADN le font. En effet, les puces à ADN fournissent des valeurs, si faibles soient elles, pour l'ensemble des transcrits. Nous nous sommes alors posé la question de savoir s'il n'existe pas un moyen de se rapprocher des conditions de mesure des écologues. L'idée qui nous est venue à l'esprit est d'utiliser les scores de détection des sondes comme un facteur de décision sur la présence ou non d'un transcrit. La première intention est de considérer qu'un transcrit est absent si sa p-value de détection est inférieure à un certain seuil et par conséquent de ramener la valeur de son expression à zéro. L'idée est séduisante car elle ramène bien à des conditions proches de celles des écologues mais elle implique des choix difficilement défendables :

- Quel est le seuil de significativité qui détermine si une sonde a détecté ou non un transcrit ?
- Une sonde avec une mauvaise p-value de détection vaut-elle pour autant zéro ?

Nous avons vu que considérer qu'une sonde détecte de l'information dépend de deux paramètres :

- Si la mesure d'intensité est significativement supérieure au bruit de fond.
- Si la mesure est stable à travers une collection de mesures de sondes identiques.

Dans le premier cas, si la mesure n'est pas supérieure au bruit de fond, cela signifie que la valeur d'intensité mesurée est très faible ; son poids dans l'abondance totale de l'échantillon sera alors elle aussi très faible. Le risque que nous encourons ici est de garder une quantité trop importante de ces petites abondances, ce qui pourrait au final représenter une quantité non négligeable et donc fausser les résultats. Prenons l'exemple du jeu de donné Treg.

Pour ce jeu de données, j'ai appliqué la méthode que j'utilise systématiquement dans ce cas : j'élimine les sondes dont la p-value de détection est systématiquement supérieure à 0,001 dans tous les échantillons. Ainsi, la matrice des valeurs d'expression finale contient 86719 sondes non uniques. En procédant de la même manière mais pour chaque échantillon individuellement, j'obtiens en moyenne 63006 sondes. La différence du nombre de transcrits est donc de 27,3%. La quantité d'information portée par les sondes dont la p-value de détection est supérieure à 0,001 est quant à elle de 14.9%. C'est un chiffre particulièrement élevé et il reste à savoir quelle est la qualité de cette information. Pour cela j'ai calculé la part d'information représentée par les sondes dont la p-value de détection est supérieure à 0,05 ; elle est de 0,0006%. Cela indique que le filtre que j'applique aux données ne conserve au final que 0,0006% d'information provenant de sondes dont la mesure est statistiquement fausse dans 5% des cas. Sans passer par un filtre extrêmement drastique, enlevant systématiquement les sondes avec des p-values inférieures à 0,001, nous obtenons comme jeu de données final, une matrice extrêmement propre en matière de qualité d'information.

Néanmoins, nous pouvons aussi considérer que la p-value de détection est une mesure de l'incertitude de la mesure d'expression. Dans cas, nous proposons de pénaliser les valeurs d'expression par ces p-values. Pour cela j'ai créé la matrice de pénalisation PN à partir de la matrice de p-values de détection DE de la manière suivante :

Équation 20:

$$PN = 1 - DE$$

La matrice d'expression normalisée, filtrée et pénalisée E_{PN} est alors calculée à partir de la matrice d'expression normalisée et filtrée E telle que :

Équation 21:

$$E_{PN} = E \times PN$$

Plus les p-values de détection vont être fortes plus la valeur d'expression va être pénalisée. Même si dans ce jeu de données, cela ne représente pas une quantité d'information très importante (la part

d'information représentée par les sondes dont la p-value de détection est supérieure à 0,05 passe alors de 0,0005%), la méthodologie reste très pertinente car elle va fortement diminuer l'expression d'une sonde très mal détectée (si une p-value vaut 0,5 par exemple) et donner alors très peu de poids à cette expression dans l'abondance totale. Nous avons appliqué cette stratégie de traitement des données à l'ensemble des jeux de données que nous avons analysés.

Une étude des facteurs de variabilités des jeux de données produits au laboratoire par PVCA nous indique qu'elles sont de très bonne qualité. En effet, même avant la normalisation les données, LPS présentent un effet biologique qui explique près de 60% de la variabilité du jeu de données. Seule la qualité des ARN (RIN) présente un taux important (près de 24%) (Figure 34). La normalisation va complètement corriger ce biais pour le faire descendre à moins de 5%. La part de la variabilité expliquée par le traitement est alors augmentée à plus de 78%. Un constat similaire est fait pour le jeu de données tolérance fœto-maternelle. Cette fois-ci, avant la normalisation la part expliquée par la cinétique est déjà de plus de 80% (Figure 35). Après normalisation elle passe à plus de 96%.

Après traitement, les jeux de données se composent d'un nombre de transcrits différent :

- LPS : 9787.
- Tolérance fœto-maternelle : 12375
- Treg : 16502

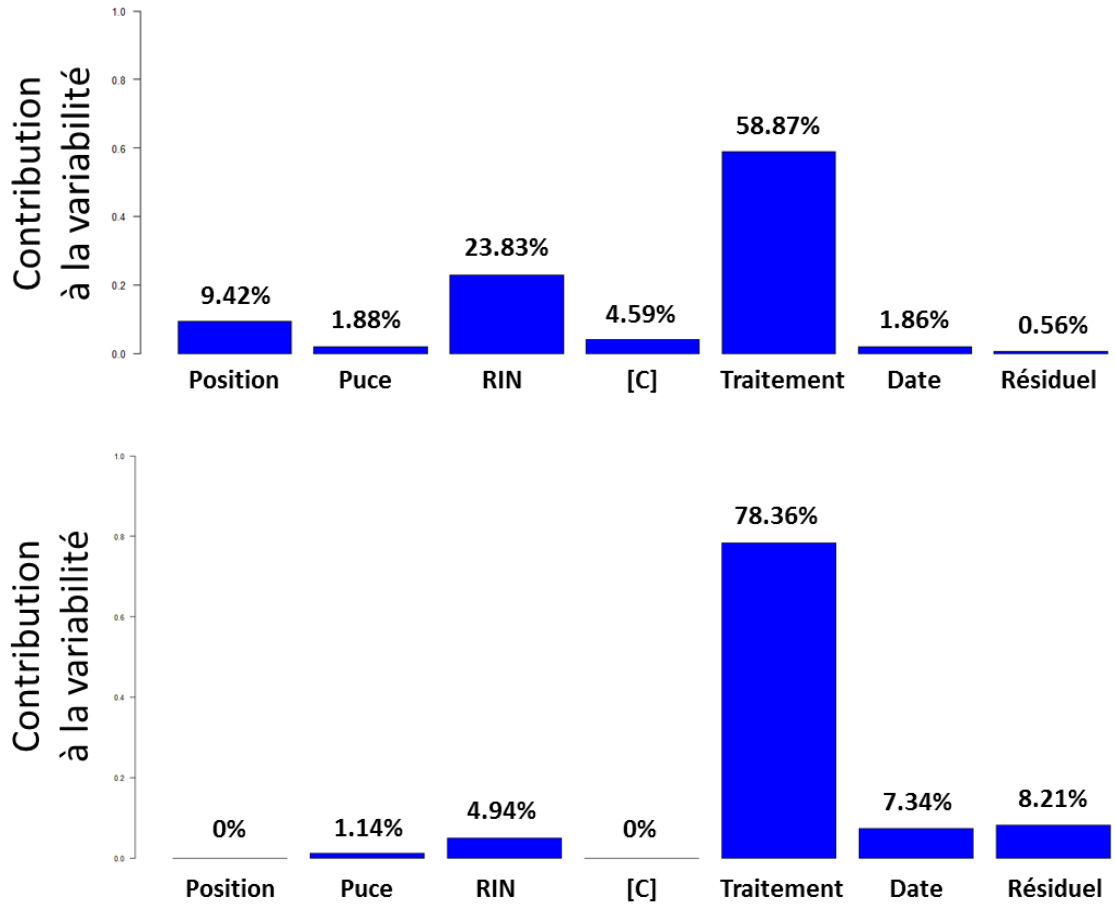


Figure 34. PVCA sur données LPS avant et après normalisation.

Représentation de la part de variabilité, du jeu de donnée de LPS, expliquée par différents facteurs biologiques et techniques. Les transcriptomes ayant été effectués sur puces du fabriquant Illumina, il existe un lot de puce (Puce) contenant chacune six positions (Position). Les ARN des souris ayant reçu une injection de PBS ou de LPS (Traitement) sont extraits à partir des cellules de rates totales prélevées à des dates connues (Date) et dosés par Nanodrop ([C]). La qualité des ARN est mesurée sur un BioAnalyseur (Agilent) ; le score de qualité est déduit de la détection des ARN ribosomaux (RIN).

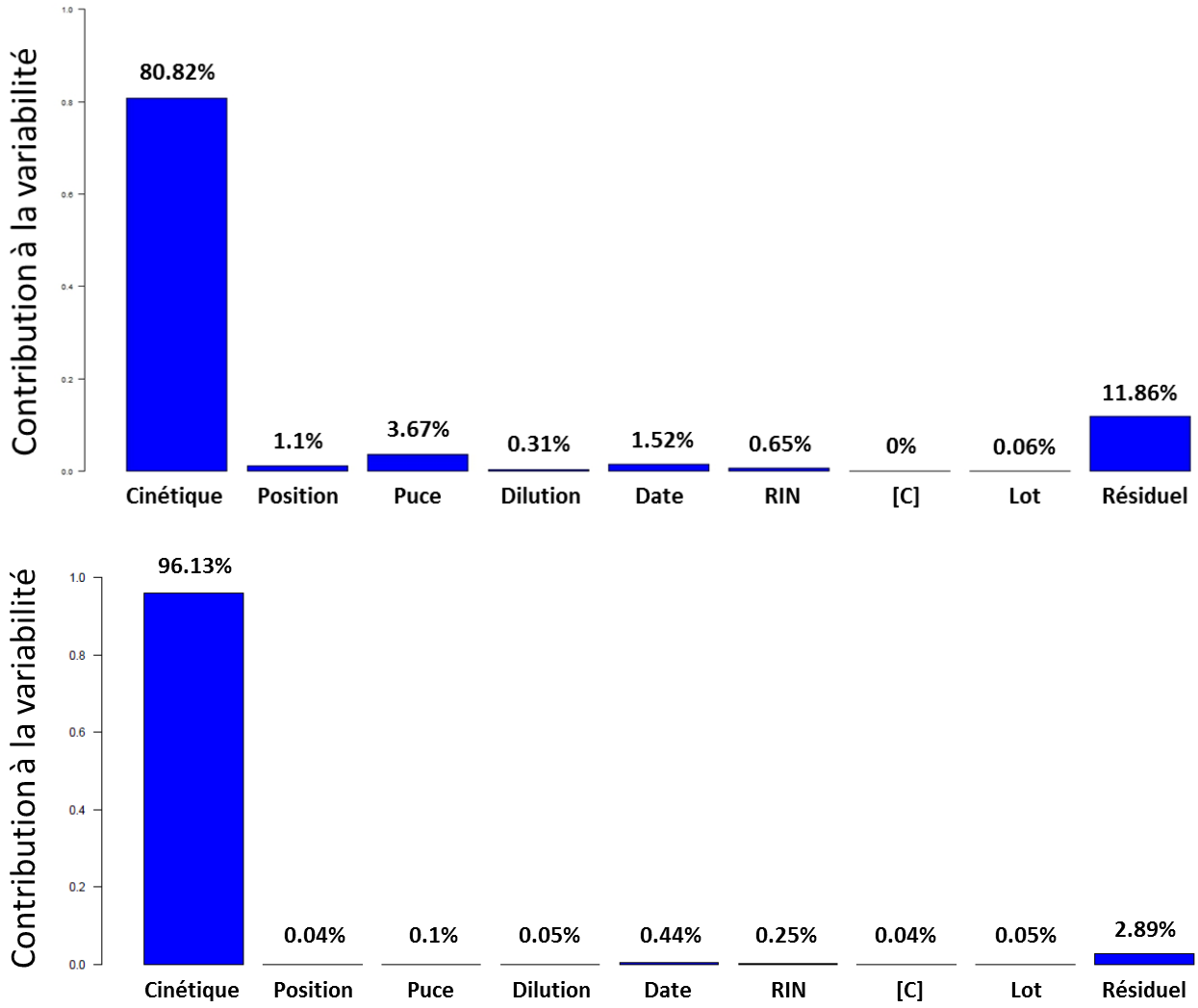


Figure 35. PVCA sur données Tolérance fœto-maternelle avant et après normalisation.

Représentation de la part de variabilité, du jeu de donnée de Tolérance fœto-maternelle, expliquée par différents facteurs biologiques et techniques. Les transcriptomes ayant été effectués sur puces du fabriquant Illumina, il existe un lot de puce (Puce) contenant chacune six positions (Position). Le nombre d'échantillons implique l'utilisation de différents lots de trois puces (Lot). Les échantillons d'environnements utérins pris à différents jours de la gestation (Cinétique) sont prélevés (Date). Les ARN sont extraits, dosés par Nanodrop ([C]) et dilués (Dilution) pour répondre aux exigences de la plate-forme P3S (Groupe Hospitalier Pitié-Salpêtrière). La qualité des ARN est mesurée sur un BioAnalyseur (Agilent) ; le score de qualité est déduit de la détection des ARN ribosomaux (RIN).

DIVERSITÉ

RÉSULTATS

Mesurer la diversité d'un jeu de données de transcriptome peut se faire de plusieurs manières. Tout d'abord, il est possible de mesurer la diversité des échantillons individuellement (diversité α). Nous pouvons aussi considérer que deux souris femelles, d'une même fratrie et élevées ensembles, sont en théorie biologiquement très proche et en triant les cellules nous éliminons les biais possibles dues à la proportion de différents types cellulaires. À la manière d'un écologue qui ferait plusieurs prélèvements dans un champ, nous pouvons considérer les échantillons d'un même groupe comme des prélèvements d'un même système et calculer ainsi la diversité du système (diversité γ).

Pour cela, j'ai développé une succession d'algorithmes codés en langage R pour calculer la diversité d'un échantillon et d'un groupe d'échantillons en fonction de la valeur de q. La librairie utilisée pour calculer les indices est *entropart* de E. Marcon et B. Herault (Marcon and Hérault, 2015).

En règle générale, un vecteur d'abondance est utilisé pour calculer la diversité. En écologie, ce vecteur n'est pas exact par définition car il existe une chance que toutes les espèces n'aient pas été comptabilisées. Il existe alors des estimateurs et des correcteurs de biais pour contourner cette problématique. Néanmoins, dans notre cas, c'est-à-dire avec l'utilisation des puces à ADN, nous connaissons exactement la richesse maximale de nos échantillons car, une fois de plus, nous ne pouvons pas estimer ce que nous ne cherchons pas. Dans ce cas, il n'est pas nécessaire d'estimer l'abondance des espèces non observées et nous pouvons utiliser directement un vecteur de probabilité. La fonction *Diversity()* de cette librairie prend en argument ce type de vecteur, c'est-à-dire la fréquence relative de chaque élément du vecteur par rapport à la somme des éléments du vecteur. Nous pouvons donc calculer la fréquence relative d'un gène i d'un échantillon j contenant un nombre g de gènes par :

Équation 22:

$$f_{ij} = \frac{x_{ij}}{\sum_{k=1}^{g} x_{kj}}$$

Le vecteur ainsi obtenu équivaut à un vecteur de probabilité sommant à 1.

Prenons pour exemple un échantillon de l'expérience LPS et nommons le X. Calculons Z1 et Z2, respectivement les vecteurs d'abondance et de probabilité issus de X. Calculons ensuite la diversité pour $q = 1$. Dans le cas du calcul de la diversité à partir du vecteur d'abondance, la fonction *Diversity()* utilise par défaut la correction de Chao-Wang-Jost (Chao and Jost, 2015). Nous constatons alors que la différence entre les deux mesures, exprimée en nombre d'espèces, est très faible.

```
library(entropart)
Z1<-as.AbdVector(X)
Z2<-as.ProbaVector(X)
Diversity(Z1,q=1)
      ChaoWangJost
      2644.526
Diversity(Z2,q=1)
      None
      2643.046
```

La diversité est ainsi calculée pour $q = 1$ (indice de Shannon) et $q = 2$ (indice de Simpson), deux indices couramment utilisés en écologie, pour les trois jeux de données. Les valeurs de diversité de chaque échantillon du jeu de données LPS sont présentées en Figure 36. L'indice de Shannon nous donne une diversité supérieure à 2590 pour tous les échantillons. Cela revient à dire que les échantillons comprenant 9787 transcrits chacun peuvent se résumer à un peu plus de 2590 transcrits. Les échantillons provenant des souris contrôles sont à gauche de l'image (en rouge) et les échantillons des souris ayant reçu des injections de LPS à droite (en bleu). Nous constatons qu'il existe visiblement une différence entre les deux conditions en termes de diversité. Il semble en effet que la diversité soit inférieure pour les souris sous LPS que pour les souris sous PBS. Afin de nous assurer de cette différence, je procède au calcul des moyennes des valeurs de diversité pour chaque groupe en utilisant la méthode du *bootstrap* décrites dans la section « Tenir compte de la variabilité inter-individuelle ». Le *bootstrap* comprend mille sélections aléatoires, un nombre qui ne surpasse pas l'ensemble des combinaisons possibles pour la condition la plus faible en nombre d'échantillon (souris sous LPS) dont le nombre de possibilités d'échantillonnages avec remise est de $5^5 = 3125$. Je profite du *bootstrap* pour estimer la moyenne et la variance des distributions de

données, corrigé par le biais les valeurs initiale si nécessaire. Ces nouvelles estimations serviront à calculer la statistique de différence des moyennes par le test t.

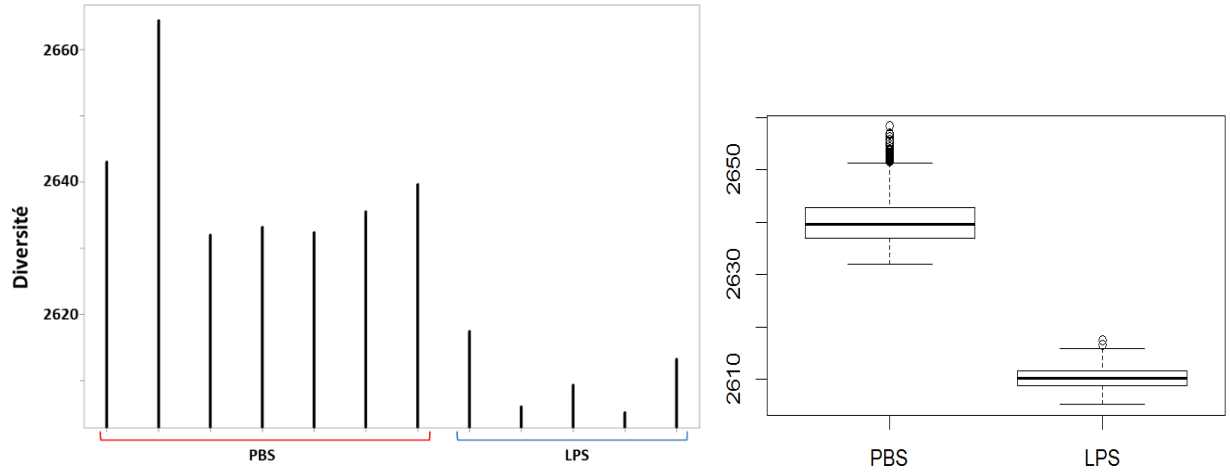


Figure 36. Diversité au sein du jeu de données LPS selon l'indice de Shannon.
Graphique de gauche : Diversité individuelle des différents échantillons du jeu de données calculée pour $q = 1$. Graphique de droite : distribution des diversités moyennes, calculées à partir de mille tirages aléatoires pour chaque groupe d'échantillons, par la représentation des boîtes à moustaches. Le trait noir épais représente la médiane des moyennes. La boîte inclus 50% des valeurs, les traits en pointillés de chaque côté de la boîte représentent chacun 1,5 fois l'espace interquartile. Les cercles noirs sont des valeurs extrêmes.

La distribution des moyennes n'est pas du tout chevauchante, indiquant une différence nette et surtout constante entre les deux groupes expérimentaux (test t, p-value < $1{,}2.10^{-4}$). Il existe donc une différence entre les deux groupes expérimentaux, minime en termes d'intensité car elle correspond à une baisse de 1,2% de la diversité.

J'ai alors regardé si cette diminution existait aussi pour une valeur de q supérieure. La Figure 37 montre les résultats obtenus par le calcul de l'indice de Simpson pour ces mêmes données. Ces résultats indiquent une tendance similaire à celle observée précédemment pour l'indice de Shannon avec une diminution de la diversité pour les souris traitées au LPS (test t, p-value < 0,026). En revanche, nous constatons ici que les boîtes à moustaches se chevauchent un peu, indiquant que la différence est moins nette que pour l'indice de Shannon (baisse de 0,7%). En effet, certains échantillons des souris sous LPS ont des diversités similaires à celles des souris traitées par une solution saline. Ceci n'est pas observé avec l'indice de Shannon. Cela indiquerait donc que la différence de diversité entre les deux conditions est particulièrement portée par des transcrits de faibles fréquences pour lesquels l'indice de Simpson est assez peu sensible.

Par ailleurs, le comportement de l'échantillon numéro 2 des souris sous PBS est intéressant car sa diversité est systématiquement beaucoup plus forte que les autres échantillons. Ce qui se traduit par une taille de la boîte à moustache plus importante pour ce groupe expérimental que pour l'autre. Dans ce cas précis, la différence de diversité avec les autres échantillons est plus importante avec l'indice de Simpson que l'indice de Shannon. Il explique donc en grande partie l'augmentation moyenne de la diversité du groupe d'échantillon.

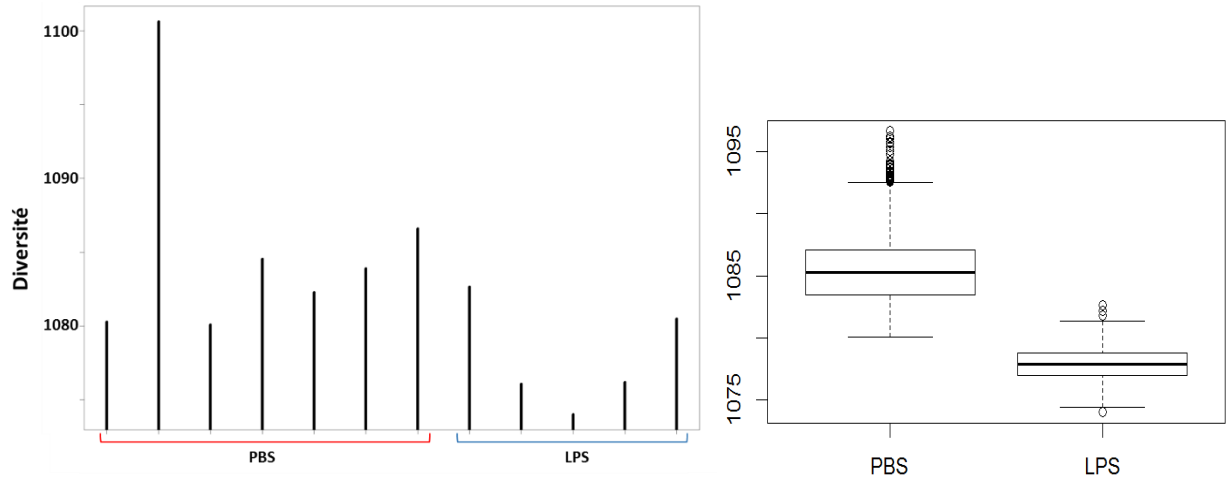


Figure 37. Diversité au sein du jeu de données LPS selon l'indice de Simpson.
Graphique de gauche : Diversité individuelle des différents échantillons du jeu de données calculée pour $q = 2$. Graphique de droite : Distribution des diversités moyennes, calculées à partir de mille tirages aléatoires pour chaque groupe d'échantillons, par la représentation des boîtes à moustaches. Le trait noir épais représente la médiane des moyennes. La boîte inclus 50% des valeurs, les traits en pointillés de chaque côté de la boîte représentent chacun 1,5 fois l'espace interquartile. Les cercles noirs sont des valeurs extrêmes.

Regardons maintenant les résultats obtenus sur le jeu de données Tolérance fœto-maternelle. Les figures Figure 38A et Figure 38B montrent les résultats obtenus avec l'indice de Shannon. Nous constatons cette fois que la diversité ne baisse pas de manière significative entre les souris non-gestantes et les premiers jours de la gestation. L'impact sur la diversité est observé à partir de E10 où la diversité augmente significativement par rapport aux jours précédents, les distributions ne se chevauchant plus, et ce pour l'ensemble des échantillons concernés.

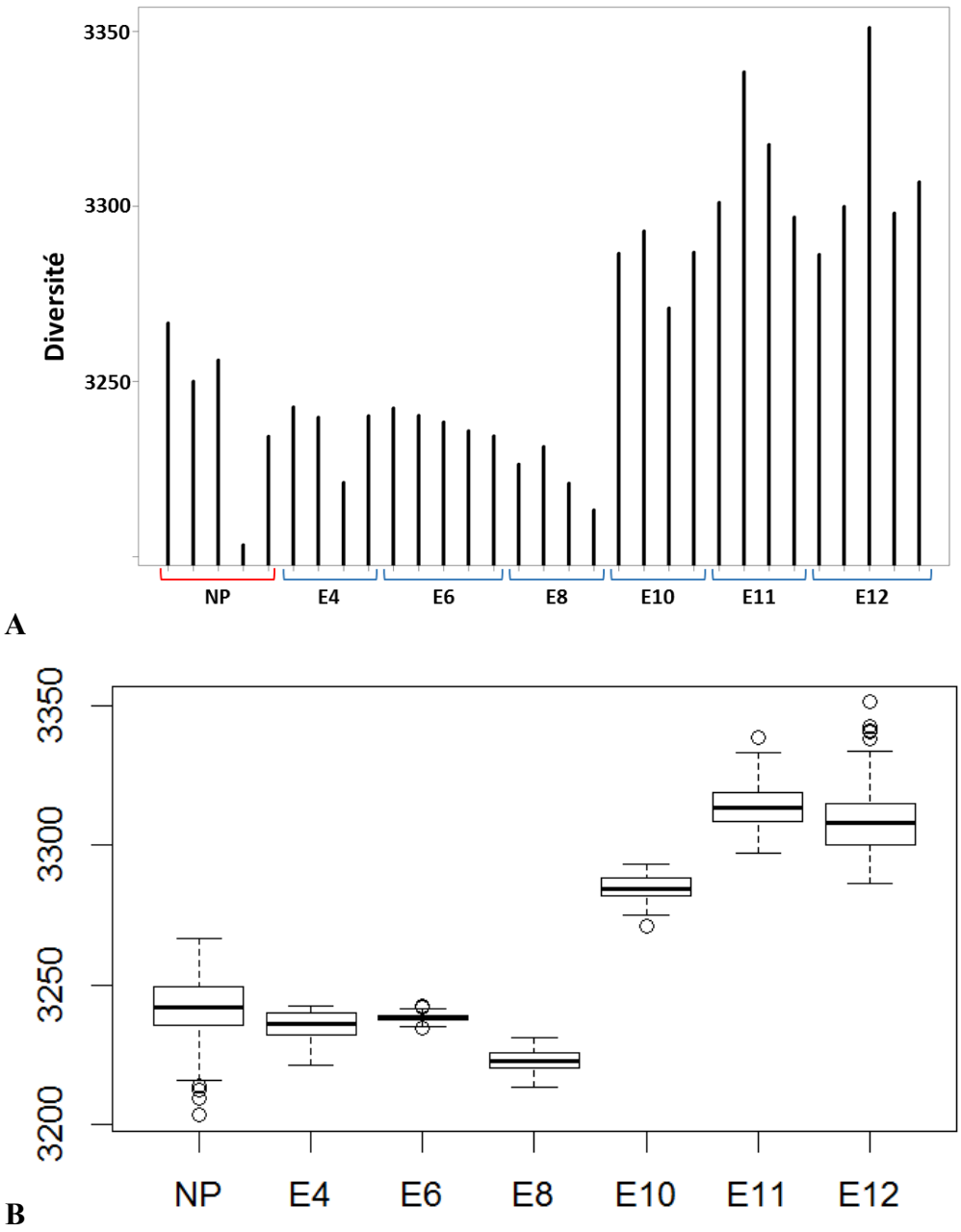


Figure 38. Diversité au sein du jeu de données Tolérance fœto-maternelle selon l'indice de Shannon.

A : Diversité individuelle des différents échantillons du jeu de données calculée pour $q = 1$. B : Distribution des diversités moyennes, calculées à partir de mille tirages aléatoires pour chaque groupe d'échantillons, par la représentation des boîtes à moustaches. Le trait noir épais représente la médiane des moyennes. La boîte inclus 50% des valeurs, les traits en pointillés de chaque côté de la boîte représentent chacun 1,5 fois l'espace interquartile. Les cercles noirs sont des valeurs extrêmes.

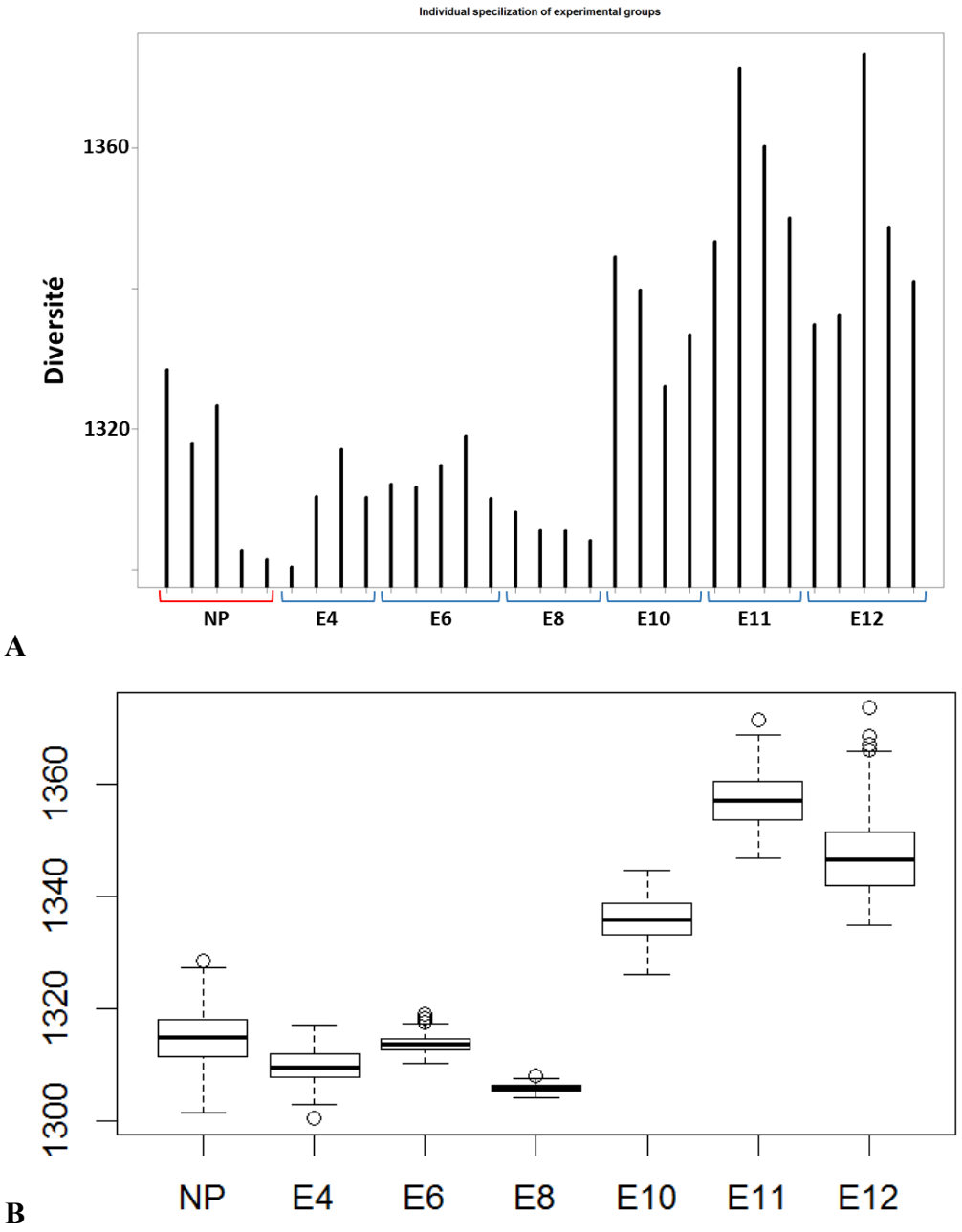


Figure 39. Diversité au sein du jeu de données Tolérance fœto-maternelle selon l'indice de Simpson.
A : Diversité individuelle des différents échantillons du jeu de données calculée pour $q = 2$. B : Distribution des diversités moyennes, calculées à partir de mille tirages aléatoires pour chaque groupe d'échantillons, par la représentation des boîtes à moustaches. Le trait noir épais représente la médiane des moyennes. La boîte inclus 50% des valeurs, les traits en pointillés de chaque côté de la boîte représentent chacun 1,5 fois l'espace interquartile. Les cercles noirs sont des valeurs extrêmes.

Les résultats obtenus avec l'indice de Simpson sont cette fois-ci tout à fait semblables à ceux obtenus avec l'indice de Shannon. La diversité n'est pas significativement modifiée dans les

premiers jours, mais elle augmente significativement à partir de E10 (Figure 39). La ressemblance des profils de distribution est elle aussi très forte, notamment pour les valeurs de diversité à partir de E10, indiquant une certaine stabilité de la différence de diversité d'un échantillon à l'autre pour différentes valeurs de q.

Étudions les résultats obtenus pour l'expérience Treg dont les différents graphiques sont représentés en figures. L'indice de Shannon indique une tendance à l'augmentation de la diversité après la stimulation à l'IL-2 mais non significative (p-value = 0,326) (Figure 40). Cette tendance est encore plus forte avec l'indice de Simpson mais toujours non significative (p-value = 0,185) (Figure 41).

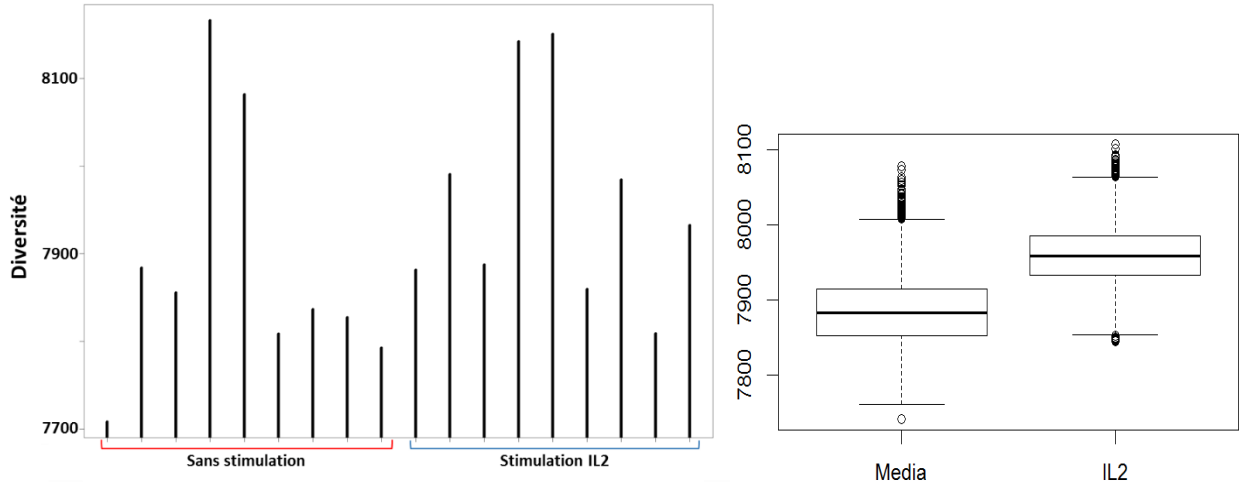


Figure 40. Diversité au sein du jeu de données Treg selon l'indice de Shannon.

Graphique de gauche : Diversité individuelle des différents échantillons du jeu de données calculée pour $q = 1$. Graphique de droite : Distribution des diversités moyennes, calculées à partir de mille tirages aléatoires pour chaque groupe d'échantillons, par la représentation des boîtes à moustaches. Le trait noir épais représente la médiane des moyennes. La boîte inclus 50% des valeurs, les traits en pointillés de chaque côté de la boîte représentent chacun 1,5 fois l'espace interquartile. Les cercles noirs sont des valeurs extrêmes. Les échantillons sont classés par appariement : le premier échantillon du groupe sans stimulation est apparié au premier échantillon du groupe avec stimulation.

Nous pouvons noter trois autres points intéressants sur ce jeu de données :

- Les diversités des échantillons présentent les mêmes profils pour les deux indices ; les échantillons avec un fort indice de Shannon, ont aussi un fort indice de Simpson et vice-versa (corrélation de 0,88 ; p-value = 0.0016 entre les résultats des indices de Shannon et

de Simpson pour le groupe sans stimulation. Corrélation de 0,91 ; p-value = 0.0005 entre les résultats des indices de Shannon et de Simpson pour le groupe avec stimulation).

- Les données sont appariées et cela se ressent sur les valeurs de diversité car il existe une relation forte entre les diversités des échantillons sans stimulation et avec stimulation (corrélation de 0,87 ; p-value = 0,0022 pour les résultats de l'indice de Shannon. Corrélation de 0,82 ; p-value = 0,0073 pour les résultats de l'indice de Simpson).
- La variabilité des indices semble aussi importante dans les deux conditions.

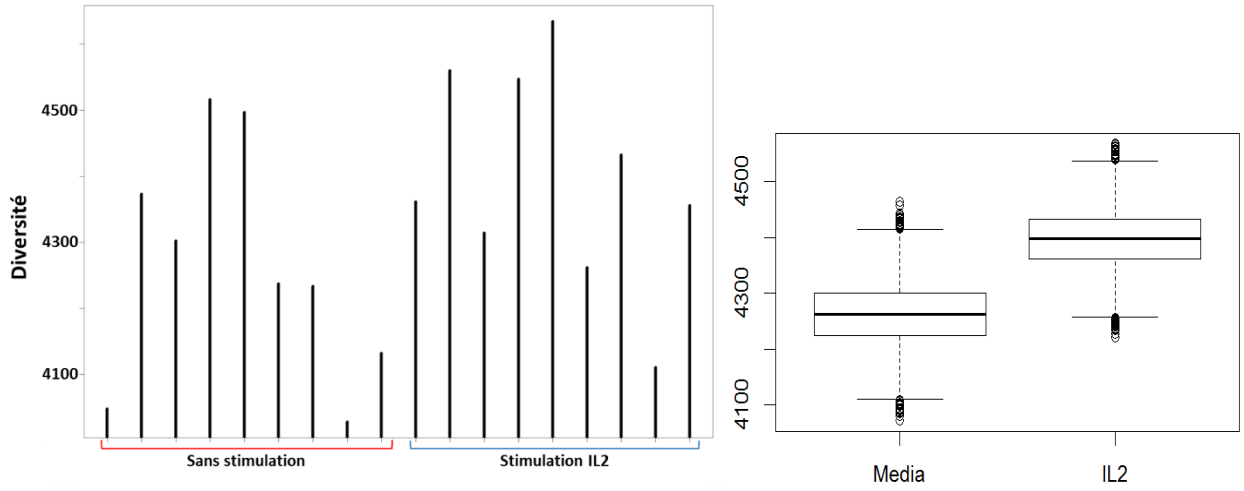


Figure 41. Diversité au sein du jeu de données Treg selon l'indice de Simpson.
Graphique de gauche : Diversité individuelle des différents échantillons du jeu de données calculée pour $q = 2$. Graphique de droite : Distribution des diversités moyennes, calculées à partir de mille tirages aléatoires pour chaque groupe d'échantillons, par la représentation des boîtes à moustaches. Le trait noir épais représente la médiane des moyennes. La boîte inclus 50% des valeurs, les traits en pointillés de chaque côté de la boîte représentent chacun 1,5 fois l'espace interquartile. Les cercles noirs sont des valeurs extrêmes. Les échantillons sont classés par appariement : le premier échantillon du groupe sans stimulation est apparié au premier échantillon du groupe avec stimulation.

Ce dernier point m'intéresse tout particulièrement car, s'il on reprend les observations faites dans l'Introduction à partir des analyses en composantes principales des différents jeux de données, j'avais indiqué l'existence d'une diminution de la variabilité inter-individuelle dans le groupe sujet au stimulus. Afin d'étudier ce phénomène, je me suis intéressé à la similarité des échantillons pour chaque groupe expérimental. Les outils à disposition des écologues fournissent un moyen de calculer un indice de similarité correspondant à :

Équation 23:

$$Sim = \frac{\left({}^{q}D(H_\alpha) \Big/ {}^{q}D\left(H_\gamma\right) \right) - 1/N}{1 - 1/N}$$

N est le nombre d'échantillon dans un groupe expérimental, ${}^{q}D(H_\alpha)$ est la moyenne des valeurs de diversité α d'ordre q des échantillons d'un groupe expérimental (Équation 24) et ${}^{q}D\left(H_\gamma\right)$ est la diversité γ d'ordre q d'un groupe expérimental. Cette dernière valeur est calculée après avoir sommé, pour chaque transcrit, les valeurs d'expression à travers N échantillons.

Équation 24:

$${}^{q}D(H_\alpha) = \frac{\sum_{i=1}^{N} {}^{q}D\left(H_{\alpha_i}\right)}{N}$$

L'indice de similarité est compris entre 0 et 1 car, en écologie, la diversité γ est supérieure à la moyenne des diversités α. Sommer les comptes de plusieurs échantillons engendre l'insertion de nouvelles espèces dans le vecteur d'abondances final. Or, l'ajout d'une espèce engendre automatiquement une augmentation de la diversité. Si les échantillons sont identiques, alors le vecteur d'abondance sera la même pour la diversité γ et les diversités α. Les deux diversités seront alors identiques et l'indice de similarité vaudra 1. À l'inverse, plus les échantillons sont différents, plus la moyenne de la diversité α sera différente de la diversité γ et, par conséquent, plus la similarité sera faible. La méthode prend de fait en compte l'existence de nombres d'échantillons différents dans les deux groupes expérimentaux.

Afin d'étudier cette similarité, j'ai procédé à ce calcul de similarité pour plusieurs valeurs de q.

La Figure 42 présente les résultats obtenus sur les données LPS. En abscisse nous retrouvons l'ordre de l'indice de diversité. L'ordonnée indique la valeur de l'indice de similarité. Pour $q = 0$, l'indice nous donne la richesse de chaque échantillon, c'est-à-dire le nombre d'espèces. Or, ce nombre est identique pour tous les échantillons, puce à ARN oblige, donc la similarité est maximisée à 1. Dès que q prend une valeur supérieure à 0, les similarités des deux groupes expérimentaux diminuent. Cette diminution est beaucoup plus prononcée pour le groupe contrôle (PBS) que le groupe LPS. L'écart s'accentue à mesure que q augmente et la tendance se confirme jusqu'à q égal à 5. L'écart maximal entre les deux similarités est d'ailleurs obtenu pour cette dernière valeur de q, indiquant ainsi que plus on ignore les transcrits de faible abondance, plus la

variabilité de la diversité au sein du groupe contrôle est importante. En revanche, elle est très stable pour le groupe soumis au LPS à partir de $q = 1$.

Une analyse similaire effectuée sur le jeu de données Tolérance fœto-maternelle, montre des profils plus complexes Figure 43. Il se dégage un début de constante néanmoins : le groupe témoin, composé des échantillons d'utérus de souris non-gestantes, à une similarité des diversités plus faible que les groupes d'échantillons provenant des souris gestantes, au moins jusqu'à $q = 3$. On notera une parfaite gradation de l'augmentation de la similarité pour les premières dates chez les souris gestantes : E4 est moins similaire que E6 qui est moins similaire que E8. La tendance s'inverse à partir de E10 qui présente une forte similarité comme E8, les deux courbes se croisant notamment pour $q = 1{,}1$. Après ce point, la similarité devient plus faible que pour E8 et est d'autant plus faible que q est grand.

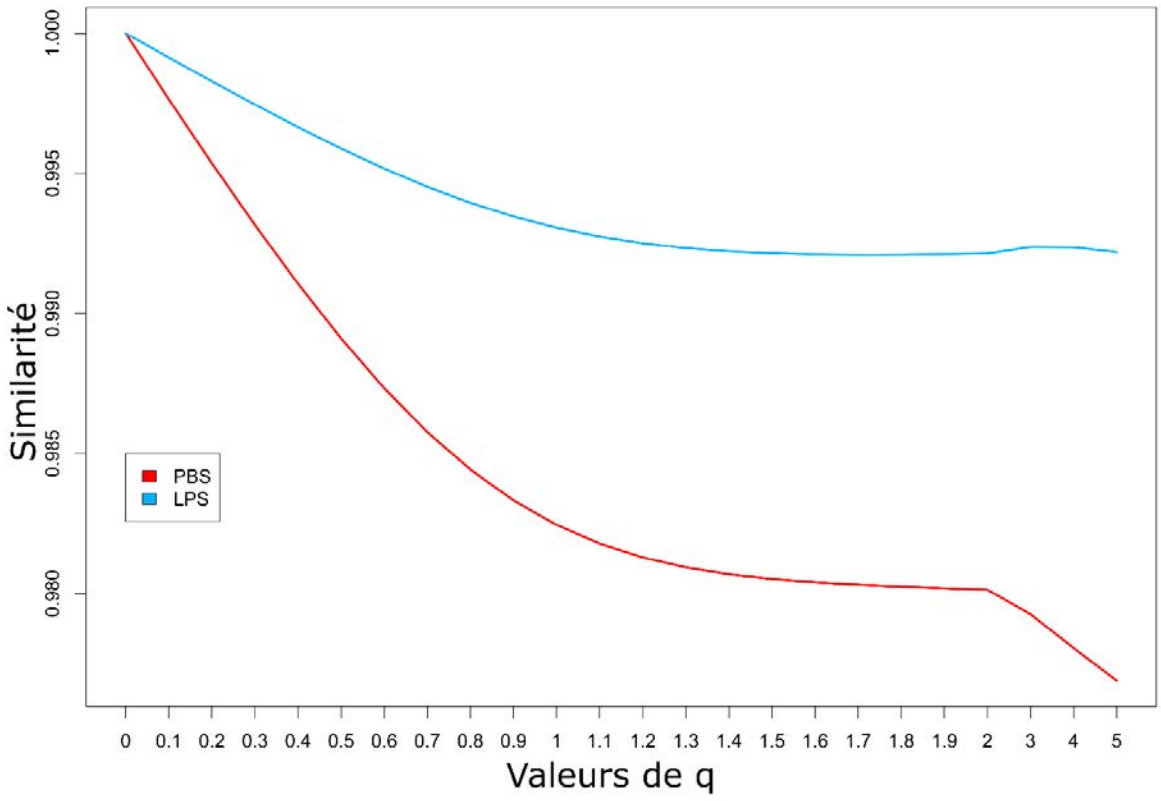


Figure 42. Similarité dans les groupes expérimentaux du jeu de données LPS.

La diversité des échantillons est calculée pour plusieurs valeurs de q ($q \in [0; 5]$). La similarité des diversités pour chaque groupe est alors calculée selon l'Équation 23. En rouge, l'évolution de la similarité des échantillons du groupe contrôle ; en bleu celle du groupe sous LPS.

Les groupes E11 présente des similarités systématiquement inférieures à E8 et E10 pour toutes les valeurs de q. La chute de la similarité est d'autant plus brutale pour des valeurs de q supérieures à 2. Même constat pour le groupe E12 qui voit sa similarité inférieure à tous les autres groupes de souris gestantes pour toutes les valeurs de q supérieure à 0,4. Seul le groupe contrôle reste très inférieur au groupe E12 tant que q est inférieur à 4. Pour q valant 4 ou 5, la similarité du groupe contrôle augmente fortement jusqu'à devenir plus importante que pour les groupes E11 et E12.

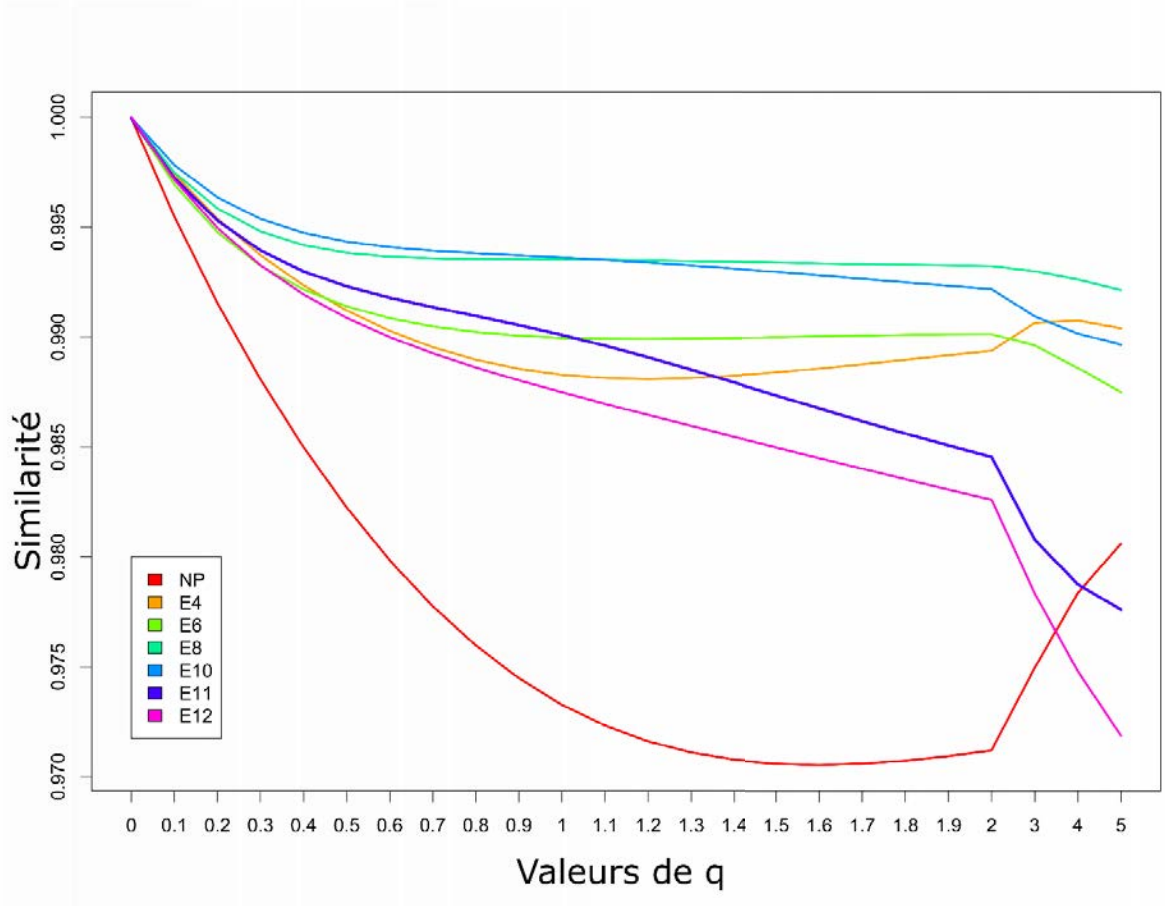


Figure 43. Similarité dans les groupes expérimentaux du jeu de données Tolérance fœto-maternelle.
La diversité des échantillons est calculée pour plusieurs valeurs de q ($q \in [0;5]$). La similarité des diversités pour chaque groupe est alors calculée selon l'Équation 23.

J'ai procédé de la même manière avec le jeu de données Treg et les résultats sont présentés en Figure 44. La tendance, initiée par les deux expériences précédentes, se confirme ici avec une valeur de similarité du groupe contrôle systématiquement inférieure au groupe ayant subi le stimulus. La différence est cette fois-ci très ténue par rapport aux expériences précédentes mais très stable car effective pour toutes les valeurs de q.

Dans les deux expériences précédentes, les valeurs de similarités des groupes de souris des conditions contrôles descendent sous la valeur de 0,98, tandis que les groupes soumis à un stimulus voient leur similarité restée au-dessus de 0,99 (avec les exceptions notables de E11 et E12). Le jeu de données Treg présente quant à lui des valeurs de similarité qui descendent en dessous de 0,98 pour les deux conditions (contrôle et sous stimulus).

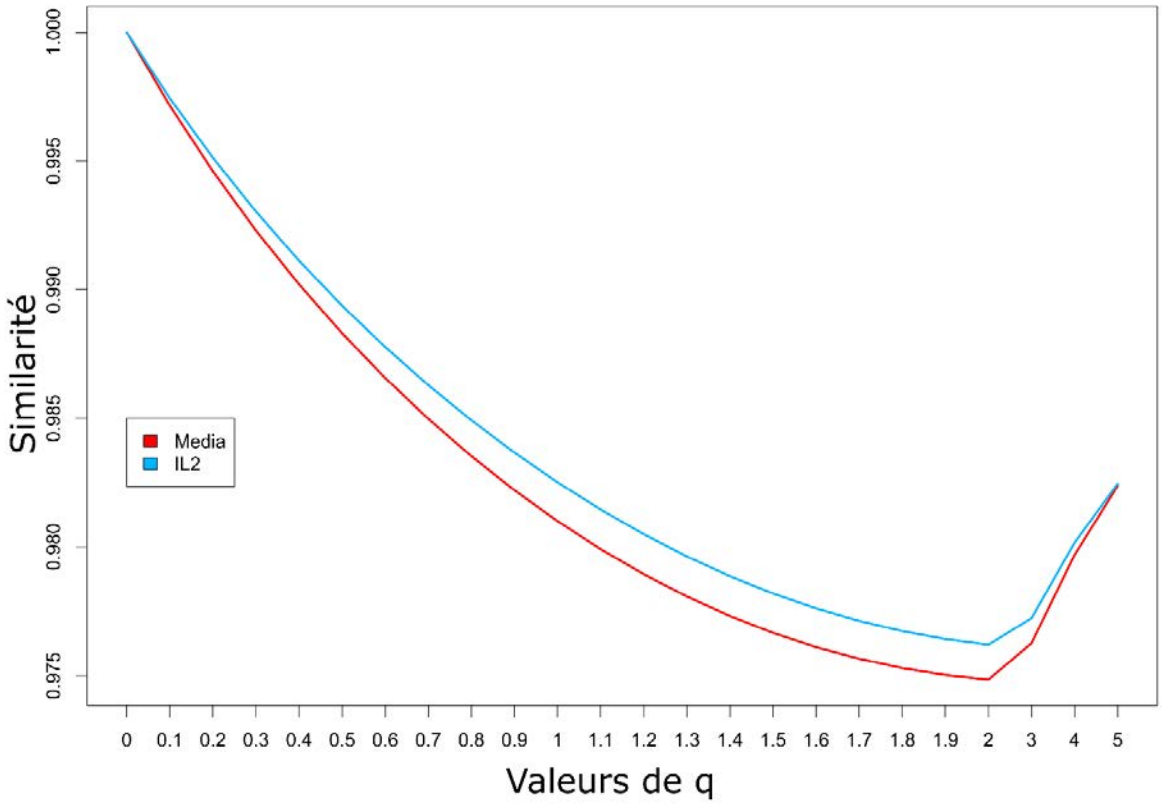


Figure 44. Similarité dans les groupes expérimentaux du jeu de données Treg.
La diversité des échantillons est calculée pour plusieurs valeurs de q ($q \in [0;5]$). La similarité des diversités pour chaque groupe est alors calculée selon l'Équation 23.

Appliqués à trois jeux de données différents, les indices de diversité montrent qu'ils apportent des clés nouvelles à la compréhension des données de transcriptome. La diversité étant liée au nombre de transcrits différents produits, ces indices donnent des indications sur la structure globale d'un échantillon. Ils permettent ainsi de mettre en évidence des comportements singuliers d'échantillons par rapport aux autres, comme nous pouvons le voir pour les échantillons 4 et 5 de la Figure 40.

Ces deux échantillons ont des diversités systématiquement plus importantes que les autres échantillons dans les deux conditions expérimentales.

Les indices de diversité permettent aussi d'évaluer un comportement global au sein d'un groupe d'échantillons par l'analyse de leur similarité. Les résultats indiquent alors une augmentation de la similarité pour les échantillons soumis à un stimulus par rapport aux échantillons témoins. Cette similarité n'est pas systématiquement vraie pour toutes les valeurs q testées. En effet, la Figure 43 nous indique qu'il est possible que les mesures de similarité s'entrecroisent en fonction de l'importance que nous donnons aux transcrits de faibles abondances.

DISCUSSION

La diversité d'un transcriptome est un indicateur assez simple à interpréter. O. Martinez en faisait une excellente démonstration, indiquant qu'un organe est d'autant plus divers qu'il exprime un grand nombre de gènes (Martínez and Reyes-Valdés, 2008). Il serait alors tentant de constater dans nos expériences que, comme le décrit O. Martinez, l'utérus est plus divers que la rate. Mais cette comparaison est très certainement hasardeuse dans notre cas. En effet, si la diversité est directement liée au nombre de gènes exprimés, je rappelle ici que lors du traitement des données, le nombre de transcrits conservés pour chaque échantillon est déterminé par l'ensemble des échantillons du jeu de données. Il n'est donc pas le même d'une expérience à l'autre. Ainsi, les jeux de données LPS et Tolérance fœto-maternelle comportent respectivement 9797 et 12375 transcrits, faisant par définition de l'utérus un organe plus divers. En réalité, ce nombre de transcrits est influencé par plusieurs facteurs. Le nombre de transcrits dépend tout d'abord de la richesse du transcriptome pour le compartiment observé ; plus ce compartiment est riche, plus il y aura des transcrits différents observés. Ainsi, un échantillon complexe, constitué de plusieurs dizaines de types cellulaires différents sera, en principe, plus riche qu'une population cellulaire triée. Il dépend également du nombre d'échantillon inclus dans l'expérience. La méthode de traitement des données que j'ai mise en place implique qu'ajouter un échantillon donne une chance supplémentaire de conserver un gène avec une p-value inférieure à 0,001. Enfin, il dépend aussi de la variabilité inter-individuelle des organismes observés. Les expériences LPS et Tolérance fœto-maternelle sont basées sur des souris sélectionnées pour être génétiquement proche les unes des

autres contrairement à l'expérience Treg constituée d'échantillons humains donc génétiquement plus éloignés les uns des autres. Dernier point important, le nombre de transcrits dépend aussi de la technologie utilisée. Dans notre cas, les deux premières expériences sont réalisées sur des puces Illumina, l'expérience Treg sur des puces Affymetrix. Les différentes technologies ont des sensibilités différentes à certains transcrits car la qualité de détection dépend des sondes dont les méthodes de synthèse diffèrent grandement d'un fabricant à un autre. Dans les trois jeux de données utilisés, nous constatons qu'il apparaît une modification de la diversité après stimulation du système. Ces modifications ne sont pas orientées dans le même sens selon l'expérience : la diversité diminue dans la rate avec injection de LPS, et augmente avec l'implantation fœtale ou la stimulation de LTreg par de l'IL-2.

Pour l'expérience Tolérance fœto-maternelle, l'augmentation traduit la modification de la balance des différentes populations cellulaires. À E4, les transformations de la paroi utérine incluent l'épaississement de l'endomètre et le recrutement de cellules, notamment du système immunitaire. L'implantation fœtale, présente dès E6, induit, là aussi, l'apparition de nouveaux types cellulaires enrichissant le transcriptome de l'environnement utérin en nouveaux transcrits.

Dans le cas de l'expérience Treg, la mécanique n'est pas la même. Les cellules sont triées et présentent donc un phénotype stable sur la base des marqueurs CD4, CD25 et FOXP3. Nous avons vu que dans Ferraro *et al.* les auteurs montraient que les gènes des LTreg s'exprimaient de manière plus diverse à travers les individus. Cette diversité doit en principe se retrouver au niveau individuel avec une population cellulaire qui est stable phénotypiquement mais variable en termes d'expression de gène. L'augmentation de la diversité traduit donc, selon moi, deux phénomènes :

- Au niveau de la cellule, l'augmentation de l'expression d'un nombre de transcrits différents plus important car mêlant le maintien du phénotype LTreg et les voies d'activation induites par l'IL-2.
- Au niveau de l'individu, les différentes cellules voient leur diversité augmenter de manière différente du fait de la variabilité inter-cellule.

Combinés, ces deux phénomènes expliqueraient l'augmentation de la diversité dans les échantillons stimulés par l'IL-2.

La diminution systématique de la diversité dans les échantillons de rates de souris ayant reçu du LPS est à mon sens liée à un déséquilibre de la balance des cellules présentes dans l'échantillon. L'introduction du LPS conduit à sa prise en charge par les cellules présentatrices d'antigène. Ces

dernières vont alors présenter cet antigène aux cellules lymphocytaires induisant une très importante expansion des cellules spécifiques de l'antigène. La diversité du transcriptome serait alors diminuée par son orientation vers la réponse au LPS. Néanmoins, on observe que cette diminution effective pour l'indice de Shannon est moins significative pour l'indice de Simpson. L'explication du phénomène mis en jeu ici est alors peut-être moins à regarder du côté des gènes fortement exprimés que de celui des gènes plus faiblement exprimés. Ceci fait écho aux travaux de J. Mar (Mar et al., 2011) ; elle étudie des cellules souches neuronales issues de différentes pathologies (schizophrénie et maladie de Parkinson). Les auteurs montrent que les variances d'expression des gènes permettent de mettre en évidence des voies de signalisation particulières pouvant expliquer le phénotype des cellules souches neuronales dans ces différents contextes pathologiques ; par exemple, il y a moins de diversité dans les cellules souches neuronales des patients atteints de schizophrénie. Or, nous l'avions vu dans l'Introduction, les cellules souches nécessitent une certaine diversité qui leur confère la plasticité nécessaire à la différenciation en plusieurs sous-populations différentes. Le point à retenir ici est que la modification de la diversité d'un échantillon à l'autre, n'est peut-être pas de la même ampleur en fonction de la valeur que l'on donne au paramètre q de la fonction de généralisation. Analyser ces différences pourrait mettre en évidence des groupes de gènes qui participent de manière plus importante à la variation de la diversité entre des échantillons. Il s'agit d'un axe de recherche pour la suite de cette étude.

Nous pouvons constater que la modification de la diversité d'un échantillon après l'induction d'un stimulus est étroitement liée à la diversité de cet échantillon avant le stimulus. Ce dernier point est démontré dans le jeu de données Treg où la distribution des valeurs de diversité après traitement à l'IL-2 suit la distribution des valeurs de diversité sans traitement. Seuls les échantillons 4 et 5 ne suivent pas cette logique. Ils sont tous deux très forts sans stimulation IL-2 et restent très forts après stimulation sans modification apparente de leur diversité.

Dans tous les cas de figure, nous voyons que l'induction d'un stimulus induit une modification stable et orientée de la diversité. Il est alors intéressant de se questionner sur les implications de cette modification pour la variabilité inter-individuelle. Pour cela j'ai appliqué la mesure de la similarité décrite par l'Équation 23. Elle consiste à comparer les valeurs de diversité des différents échantillons (diversité α) d'un groupe expérimental à la diversité globale du groupe (diversité γ). La mesure de la diversité γ est obtenue en sommant les abondances de chaque transcrit à travers les échantillons. Or, nous l'avons dit, la diversité est sensible au nombre d'échantillons. La diversité

de deux groupes expérimentaux de tailles différentes, comme c'est le cas pour deux de nos expériences, peut donc conduire à des interprétations erronées. Dans le cas de l'Équation 23, ce biais est éliminé par l'incorporation du nombre d'échantillons dans le calcul de la similarité. Les différents groupes expérimentaux sont donc comparables.

Il s'agit ici d'effectuer une analyse de la variation de la diversité des transcriptomes de la même manière qu'il est possible de regarder la variation de la variance des gènes, comme c'est le cas dans la publication de E. Glaab et R. Schneider (2012). Les auteurs y démontrent l'intérêt d'une analyse de variance de voies de signalisation connues sur des jeux de données de cancer de la prostate. La recherche systématique de différence de variance dans l'expression des gènes au sein de voies de signalisation leur a permis de mettre en évidence trois voies de signalisation qui n'avaient pas été détectées jusque-là car les tests statistiques classiques ne détectaient pas de différence dans les moyennes d'expression des gènes. Ces voies de signalisations sont liées au cycle de l'urée, au signal VEGF (*vascular endothelial growth factor*) et la cytotoxicité médiée par les NK. Pour ces trois voies de signalisation, les variances étaient significativement augmentées dans le cas du cancer de la prostate. Cette étude nous montre bien l'intérêt de regarder ce type de paramètre car il apporte une information nouvelle aux analyses standards du transcriptome. Néanmoins, les auteurs ne discutent pas de l'interprétation biologique de tels résultats.

Les différences entre notre méthodologie et celle développée dans la publication de E Glaab et R. Schneider sont tout d'abord que nous regardons la variation d'une mesure qui résume l'état d'un transcriptome. Pour leurs parts, E Glaab et R. Schneider analyse les variances des gènes individuellement. C'est la distribution de ces variances dans les voies de signalisations, citées ci-dessus, qui, comparée entre des donneurs sains et des patients, indiquent la modification générale de la variance des gènes. Le deuxième aspect qui diffère entre nos méthodes est que l'emploie des indices de diversité permet de faire varier le paramètre q et donc de suivre, au fil de la diminution de l'importance des gènes de faible abondance à mesure que q augmente, l'évolution de la variation de la diversité que nous évaluons ici par une similarité. Les auteurs, quant à eux définissent des groupements de gènes à analyser sur la base des connaissances scientifiques (voies de signalisation KEGG). Je reviendrai sur cette manière d'analyser les données dans la partie « Classer les transcrits ».

Les résultats obtenus pour l'étude de la similarité de la diversité sont remarquables par leur convergence vers une même dynamique. L'application d'un stimulus induit systématiquement une

augmentation de la similarité des échantillons. Néanmoins, il existe des différences notables entre les différentes expériences. Dans le jeu de données LPS la différence de similarité atteint son maximum pour les valeurs de diversité d'ordres élevés alors qu'elle devient nulle dans le jeu de données Treg et même inversée dans le jeu de données Tolérance fœto-maternelle. L'augmentation de l'ordre q correspond à la minimisation progressive de l'impact des transcrits de faibles abondance dans l'évaluation de la diversité. Nous observons donc des segmentations de la diminution de la diversité en fonction de l'abondance des transcrits. Ainsi dans le jeu de données Tolérance fœto-maternelle, les gènes les plus fortement exprimés, présentant donc les abondances les plus fortes, sont plus similaires dans les utérus des souris non gestantes que dans les utérus de souris gestantes 11 et 12 jours après le coït.

L'annotation par l'outil en ligne DAVID (https://david.ncifcrf.gov/) des 250 gènes les plus abondants de nos trois jeux de données indiquent qu'ils participent à la production de protéines (Table 3). La différence de diversité de ces gènes indique des changements de leur abondance et traduisent ainsi une part de l'activité cellulaire de l'échantillon.

Table 3. Enrichissement fonctionnel des gènes les plus abondants

	# Gènes	# Reconnus	Enrichissement pour le terme "Ribosome" correction Benjamini-Hochberg
LPS	250	198	$3{,}5.10^{-53}$
Tolérance	250	219	$2{,}0.10^{-43}$
Treg	250	240	$3{,}7.10^{-57}$

Compte tenu des jeux données utilisés, une hypothèse serait que ces variations de la diversité aux fortes valeurs de q soient le fruit de l'hétérogénéité du système observé. Dans le cas de la rate, nous avons affaire à un mélange de plusieurs populations cellulaires (LT, LB, DC, etc.). Le jeu de données Treg est quant à lui exclusivement composé de cellules LTreg. Enfin, l'utérus de souris non-gestantes est principalement composé de cellules musculaires (myomètre), les transformations successives de l'environnement n'intervenant qu'après la fécondation. Un système hétérogène va donc engendrer plus de variabilité dans l'expression des gènes de forte abondance qu'un système homogène. On distingue, par ailleurs, très bien l'évolution de cette modification dans le jeu de données Tolérance fœto-maternelle : la diversité est moins variable pour les valeurs de q élevées

que les valeurs faibles dans le groupe NP. Elle augmente progressivement au cours de la gestation et est particulièrement élevée à E11 et E12 où cette variabilité est plus forte que pour le groupe NP. La similarité des valeurs de diversité des échantillons augmente avec l'application d'un stimulus ; néanmoins, comment cela se traduit-il au niveau de la variabilité inter-individuelle ? Nous avons étudié cette question par l'utilisation de la mesure de spécialisation.

SPÉCIALISATION

RÉSULTATS

Pour rappel, nous appelons spécialisation d'un échantillon la somme des valeurs de spécificité des transcrits qui le compose, multiplié par leurs fréquences relatives respectives (Équation 18). Plus un échantillon contient des transcrits avec une forte spécificité, plus sa spécialisation sera forte. Elle sera d'ailleurs d'autant plus forte que la spécialisation est associée à une forte abondance du transcrit. En revanche, la spécialisation est basée sur le calcul de la spécificité des gènes qui est corrélée au nombre d'échantillons observés. Dans nos différents exemples, nous n'avons pas nécessairement le même nombre d'échantillons dans chaque groupe.

Afin de remédier à cela, je me suis inspiré du travail que j'ai effectué dans (Dérian et al., 2016). Je calcule donc la spécialisation pour chaque combinaison d'échantillons qu'il est possible de réaliser. Pour cela, nous considérons le nombre de combinaisons sans ordre établi de k échantillons qu'il est possible de réaliser à partir d'un ensemble de n échantillons :

Équation 25:

$$A_n^k = \frac{n!}{(n-k)!}$$

Ne souhaitant pas avoir de redondances dans les combinaisons dues au tirage aléatoire des échantillons, le nombre de combinaisons intéressantes vaut alors :

Équation 26:

$$C_n^k = \frac{A_n^k}{k!}$$

Prenons le cas du jeu de données LPS :

1) Nous avons 7 échantillons de PBS et 5 de LPS.
2) Le nombre de combinaisons de 5 échantillons choisis parmi 7, sans remise et avec ordre est de 21 (Équation 26).
3) La spécialisation de chaque échantillon est calculée pour chaque combinaison.
4) La spécialisation finale d'un échantillon est la médiane de ces valeurs de spécialisation.

J'ai donc créé une fonction *combinaison* qui énumère, sous forme de liste, les différentes combinaisons d'échantillons possibles à partir d'un jeu de données.

```
comb<-combinaison(data,5,7) # liste de combinaisons de 5 echantillons parmi 7.
mat<-matrix(0,length(comb),7) # matrice de résultat
For (i in 1 :length(comb)){ # début de la boucle 1
X<-data[,comb[[i]]] # selection des données selon la combinaison
X<-apply(X,2,as.ProbaVector) # transformation des données en probabilités
Si<-apply(X,1, Specificity.gene) # calcul de la spécificité des gènes
Spi<-X*Si
Special<-apply(Spi,2,sum)# calcul de la specialisation
for (j in 1:length(comb[[i]])){ # début de la boucle 2
mat[i,comb[[i]][j]]<-Special[j] # résultats attribués aux échantillons
} # fin boucle 2
} # fin boucle 1
# La matrice mat contient des zéro du fait de la non sélection de 2 échantillons
à chaque combinaison. Les zéros sont donc supprimés avant le calcul des médianes.
Mat<-removeZero(mat)
Res<- apply(mat,2,median, na.rm=TRUE) # médiane des spécialisations
```

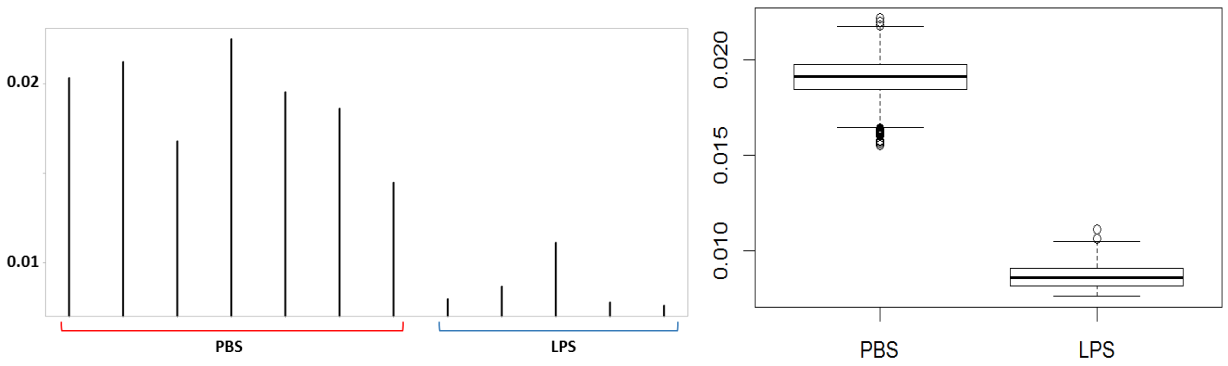


Figure 45. Spécialisation au sein du jeu de données LPS.

Graphique de gauche : Spécialisation individuelle des différents échantillons du jeu de données. Graphique de droite : Distribution des spécialisations moyennes, calculées à partir de mille tirages aléatoires pour chaque groupe d'échantillons, par la représentation des boîtes à moustaches. Le trait noir épais représente la médiane des moyennes. La boîte inclus 50% des valeurs, les traits en pointillés de chaque côté de la boîte représentent chacun 1,5 fois l'espace interquartile. Les cercles noirs sont des valeurs extrêmes.

Cette méthode est appliquée aux jeux de données LPS et Tolérance fœto-maternelle.

Le jeu de données LPS présente alors un profil très différencié pour ces deux groupes expérimentaux. La spécialisation au sein du groupe contrôle est environ deux fois plus importante que dans le groupe sous stimulation LPS (Figure 45). Nous pouvons alors formuler l'hypothèse que la spécialisation diminue dans le groupe des échantillons qui possèdent une similarité de diversité forte.

L'effet est tout aussi visible dans le jeu de données Tolérance fœto-maternelle. Nous constatons la même gradation de la diminution de la spécialisation des échantillons au cours du temps jusqu'à E8 puis son augmentation progressive, elle aussi, à partir de E10 (Figure 46). Notons que la taille des boîtes à moustache suit la même logique, à savoir une diminution jusqu'à E8 puis une augmentation à partir de E10.

Le jeu de données Treg confirme l'hypothèse de départ, présentant également des échantillons avec une valeur de spécialisation plus faible quand ils sont soumis à un stimulus que lorsqu'ils ne le sont pas (Figure 47). Ici aussi, la variabilité des valeurs de spécialisation est plus forte dans le groupe contrôle que dans le groupes sous IL-2 même si la différence est moins prononcée que pour les autres expériences avec des boîtes à moustaches chevauchantes (p-value = 0,011).

La spécialisation peut aussi nous servir de base pour l'analyse individuelle des patients. C'est particulièrement intéressant pour le jeu de données Treg qui montre, pour certains donneurs, des spécialisations aussi élevées après le stimulus qu'avant. Pour comprendre ce qui caractérisait ces patients, j'ai analysé le comportement de plusieurs milliers de signatures moléculaires sur les listes de gènes de chaque patient, ordonnés par leur spécificité, par la méthode GSEA. Contrairement à une liste ordonnée sur les abondances des transcrits qui verrait les gènes les plus fortement exprimés en haut de la liste et les gènes les moins exprimées en bas de la liste, les gènes en haut de la liste seront les gènes ayant une forte abondance associée à une forte spécificité ; en bas de la liste se trouveront les gènes avec de faibles fréquences relatives associées à de faible spécificité (Équation 17). Nous obtenons alors un score d'enrichissement qui traduit la spécificité d'une signature pour un individu. Nous espérons ainsi faire ressortir ce qui caractérise un échantillon d'un autre.

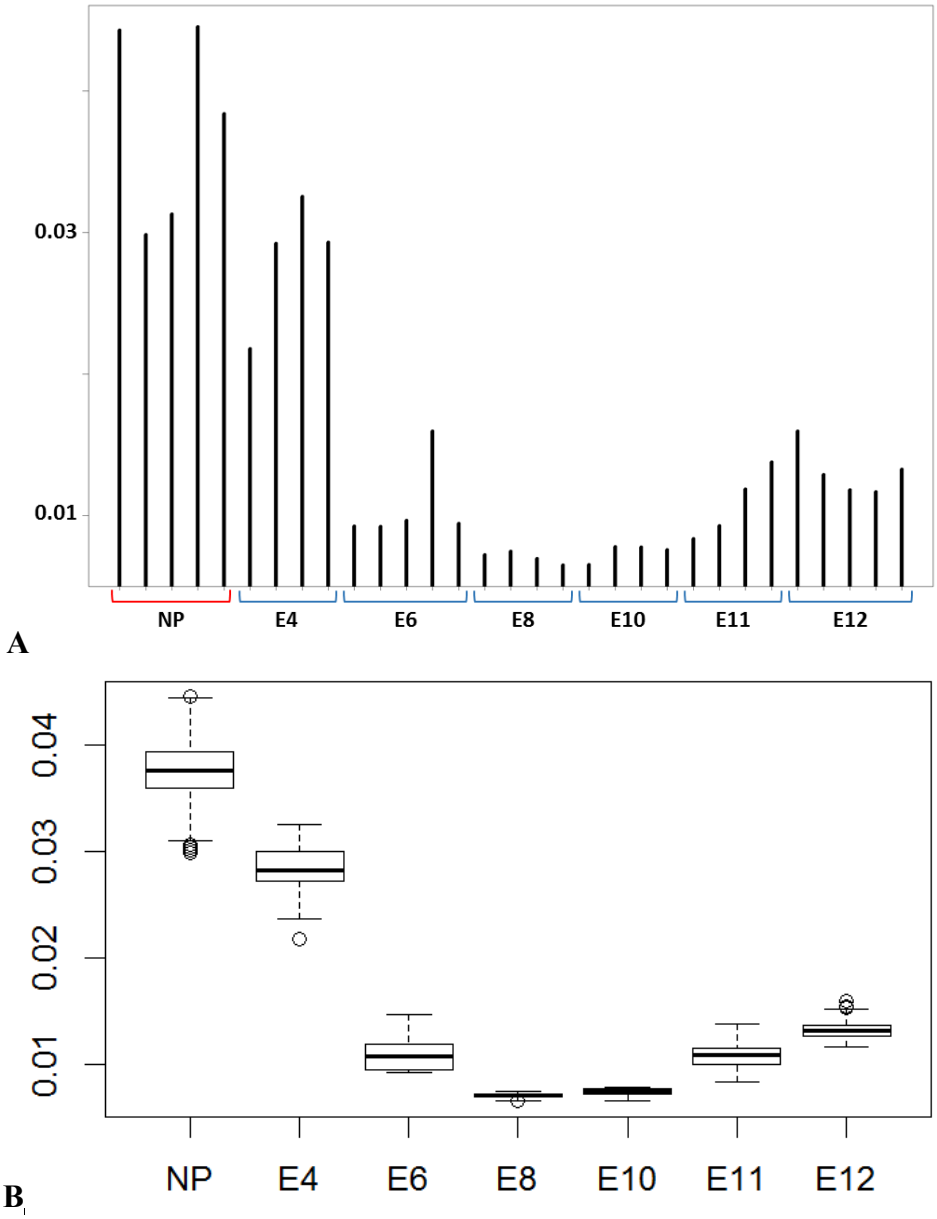


Figure 46. Spécialisation au sein du jeu de données Tolérance fœto-maternelle.

A : Spécialisation individuelle des différents échantillons du jeu de données. B : Distribution des spécialisations moyennes, calculées à partir de mille tirages aléatoires pour chaque groupe d'échantillons, par la représentation des boîtes à moustaches. Le trait noir épais représente la médiane des moyennes. La boîte inclus 50% des valeurs, les traits en pointillés de chaque côté de la boîte représentent chacun 1,5 fois l'espace interquartile. Les cercles noirs sont des valeurs extrêmes.

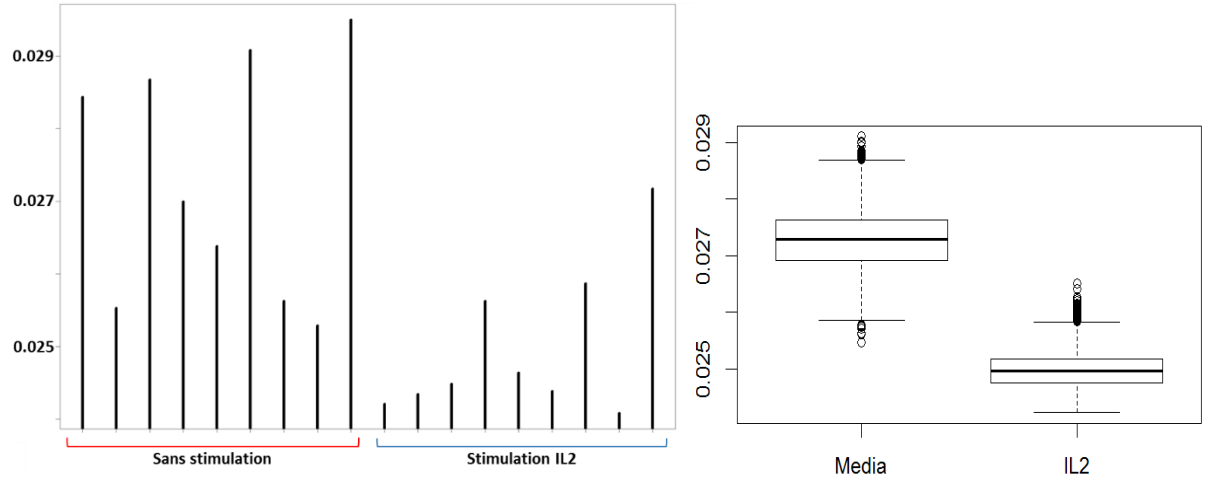


Figure 47. Spécialisation au sein du jeu de données Treg.

Graphique de gauche : Spécialisation individuelle des différents échantillons du jeu de données. Graphique de droite : Distribution des spécialisations moyennes, calculées à partir de mille tirages aléatoires pour chaque groupe d'échantillons, par la représentation des boîtes à moustaches. Le trait noir épais représente la médiane des moyennes. La boîte inclus 50% des valeurs, les traits en pointillés de chaque côté de la boîte représentent chacun 1,5 fois l'espace interquartile. Les cercles noirs sont des valeurs extrêmes.

Comparons le patient numéro 7 aux patients 1 et 6. Le patient numéro 7 est le seul qui ne voit pas sa spécialisation diminuer. Les patients 1 et 6 connaissent quant à eux des chutes drastiques de leur spécialisation. Le jeu de données de signatures testé est un mélange de signatures issues de la littérature et compilées au sein du MSigDB (http://software.broadinstitute.org/gsea/msigdb/). J'ai sélectionné le sous-ensemble C2 contenant des signatures de voies de signalisation connues issues de :

- Biocarta (217 signatures) : www.genecarta.com
- KEGG (186 signatures) : www.genome.jp/kegg/
- Reactome (674 signatures) : http://www.reactome.org/

Toutes sont des bases de données de voies de signalisation. A ces signatures, s'ajoutent des signatures dites de voies de signalisation canonique et de perturbations génétiques et chimiques, issues des sites suivants :

- Matrisome : http://matrisomeproject.mit.edu
- Pathway Interaction Database : http://pid.nci.nih.gov
- SigmaAldrich : http://www.sigmaaldrich.com/life-science.html
- Signaling Gateway : http://www.sigmaaldrich.com/life-science.html
- Signal Transduction KE : http://stke.sciencemag.org

- SuperArray : http://www.superarray.com

Enfin, nous y ajoutons une collection de signature fournies par notre méthode d'analyse (Pham et al., 2014) pour un total de plus de 5000 signatures. Les signatures sont testées sur les listes ordonnées des six échantillons des patients 1, 6 et 7. Les données sont centrées et log-transformées. Les signatures ayant un enrichissement significatif pour une grande spécificité (FDR q-value <0.001) sont sélectionnées et analysées grâce à un diagramme de Venn (Figure 48).

Pour chaque patient, le nombre de signatures sélectionnées est environ deux fois plus élevé dans les échantillons sous IL-2 que sans stimulation. Les signatures qui nous intéressent ici sont celles qui sont spécifiques à chaque patient : 44 signatures pour le patient 7, 13 pour le patient 1 et 64 pour le patient 6 dans la condition sans stimulation. Parmi ces signatures spécifiques, on retrouve des signatures en lien avec la machinerie immunitaire (voir document en Annexe #3). Le donneur 6, par exemple possède un enrichissement pour une signature liée à la réponse IL-2 (« Reactome_IL_2_Signaling »), indiquant que les gènes qui la compose sont à la fois fortement exprimés et de manière spécifique à ce patient par rapport aux deux autres donneurs.

Après stimulation, les donneurs 1 et 6 sont assez peu enrichis en signature liée à la machinerie immunitaire. Le donneur 7 en revanche possède un nombre important de signatures spécifiques (124, soit 20% de ses signatures sélectionnées). Parmi elles on retrouve une signature liée à l'IL-4 mais aussi l'IL-10, l'IL-12 et l'interféron.

Une expérience similaire effectuée sur les donneurs 4, 7 et 9 est représentée en Figure 49. Ces patients partagent le fait d'avoir les plus fortes spécialisations dans la condition traitée par l'IL-2. Une fois encore, le nombre de signatures est deux fois plus important dans la condition IL-2 que dans la condition contrôle.

Dans la condition traitée à l'IL-2, le donneur 4 montre des signatures enrichies pour des gènes liés à la réponse IL-2, à l'IL-7 et à l'IL-17, à la réponse antivirale. Le donneur 7 perd la spécificité des signatures liées à l'IL-4, partagée avec les deux autres donneurs, tout comme celle liée à l'IL-12. La signature liée à l'IL-10 est quant à elle partagée avec le donneur 4 seulement. Son profil est en revanche toujours lié à une signature liée à FOXP3 (« Gavin-Foxp3_Targets_Cluster_P6 »). Le profil du donneur 9 contient, pour sa part, des signatures liées aux signaux induit par les cytokines et le chimiokines, à la voie NFkB, à l'IL-1 et l'IL-2.

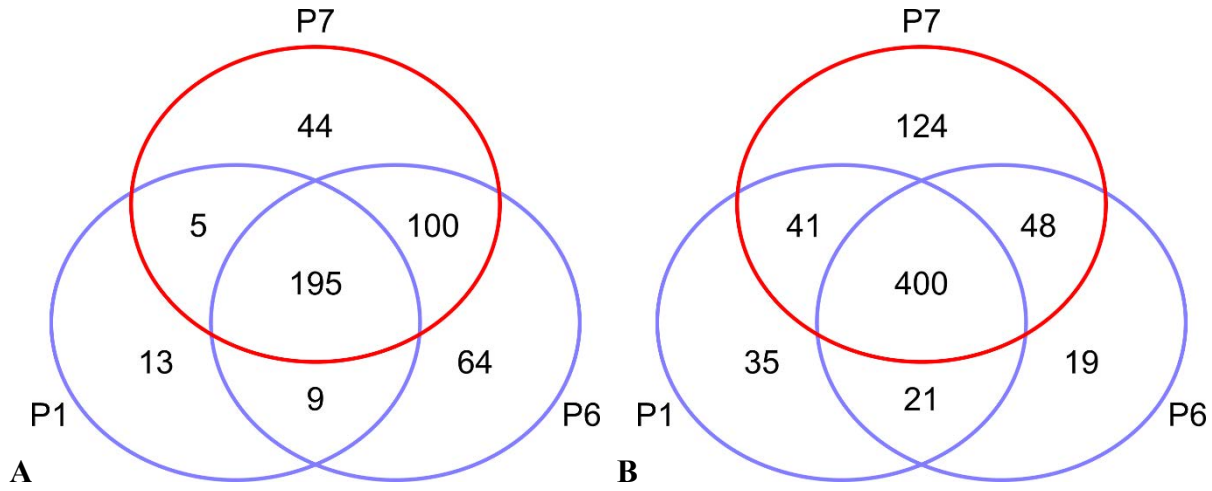


Figure 48. Signatures spécifiques dans le jeu de données Treg : Patients 1, 6 et 7.
Diagrammes de Venn des signatures significativement enrichies en gènes spécifiques (FDR q-value<0.001) pour les échantillons des donneurs 1, 6 et 7 du jeu de données Treg, sans stimulation (A) et avec stimulation (B).

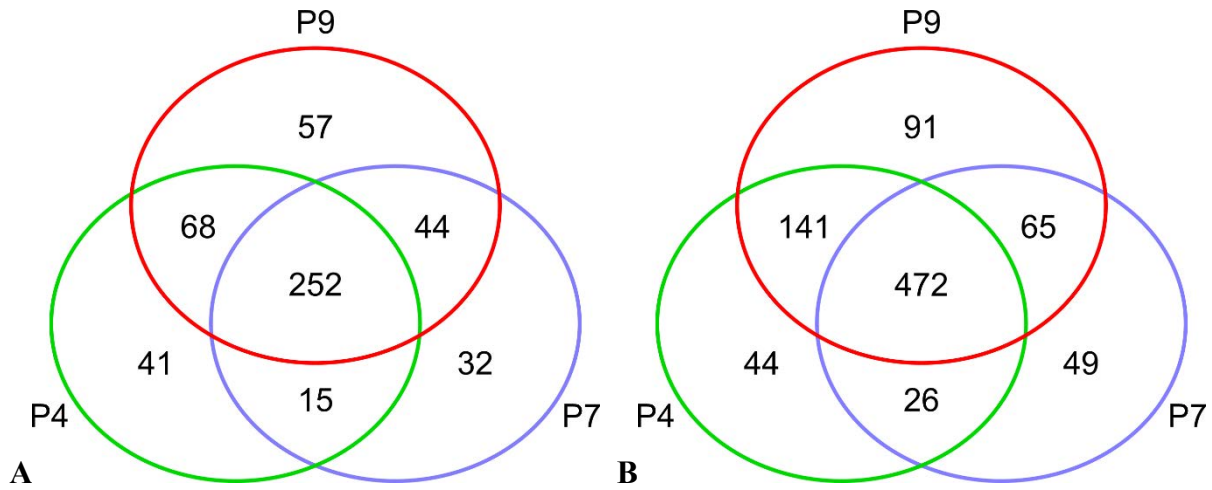


Figure 49. Signatures spécifiques dans le jeu de données Treg : patients 4,7 et 9.
Diagrammes de Venn des signatures significativement enrichies en gènes spécifiques (FDR q-value<0.001) pour les échantillons des donneurs 4, 7 et 9 du jeu de données Treg, sans stimulation (A) et avec stimulation (B).

La mesure de spécificité d'un transcrit et celle de spécialisation d'un échantillon qui en découle, mettent en avant une caractéristique intéressante sur un jeu de données de transcriptome. La spécialisation permet d'informer sur le caractère singulier d'un transcriptome par rapport à d'autres. Nous le constatons dans le jeu de données Treg où les échantillons 1, 3, 6 et 9 présentent des mesures de spécialisation plus élevées que les autres échantillons, indiquant qu'ils expriment un certain nombre de transcrits de manière singulière, ou de manière partagée avec un faible

nombre d'autres échantillons. Il s'agit là d'indications qu'ils ne nous étaient pas accessibles auparavant. Les résultats présentés dans cette section montrent donc bien le caractère innovant de la mesure de spécialisation.

La spécificité, elle aussi, est une innovation dans notre manière d'étudier un jeu de données de transcriptome. En ordonnant les gènes par leur spécificité, nous faisons ressortir les paramètres moléculaires qui différencient les échantillons à l'échelle individuelle. Elle permet ainsi d'apprécier des comportements de gènes de manière individuelle, mais aussi de signature moléculaires, qui fournissent des informations importantes sur l'état de l'échantillon et qui ne seraient pas mis en avant par d'autres méthodes, notamment celles basées sur la mesure de variance.

DISCUSSION

En prenant la mesure de la spécialisation comme mesure de variabilité inter-individuelle, nous sortons des schémas classiques de ce type d'analyse (Jiang et al., 2014; Mar et al., 2011). La mesure ajoute une notion de spécificité qui est absente des mesures effectuées par l'analyse de variance par exemple. Dans les deux publications précédemment citées sont toutes deux basées sur des calculs de diversité utilisant des statistiques descriptives classiques (corrélation, coefficient de variation, distance deux à deux…). Ces mesures sont adaptées pour ce type d'analyse ; la mesure de spécificité que nous utilisons apporte une information complémentaire en nous permettant d'évaluer l'existence, dans les échantillons d'un groupe expérimental, de transcrits qui voient leur expression être spécifique d'un ou de quelques échantillons. Les résultats ce chapitre indiquent que le nombre de transcrits ayant ce type de profil est influencé par le statut expérimental. Après stimulation, les échantillons présentent une spécialisation systématiquement plus faible qu'avant la stimulation.

D'une manière générale on constate aussi que la spécialisation des échantillons est très faible à l'intérieur des groupes expérimentaux. Il suffit, pour s'en convaincre de la comparer à la valeur maximale de chaque groupe (Table 4) : nous sommes loin des valeurs obtenues par O. Martinez dans son papier mais cela est tout à fait normal (Martínez and Reyes-Valdés, 2008). O. Martinez comparaît des organes entre eux, la spécialisation étant alors très importante car les fonctions biologiques associées à ces organes sont différentes et induisent l'expression de gènes différents.

Dans notre cas, nous comparons des échantillons provenant d'un même système ; les différences sont alors bien plus ténues qu'entre organes. D'expérience, la corrélation d'expression entre deux échantillons d'un même compartiment est supérieure à 0,9. C'est d'ailleurs un critère possible pour la détection d'échantillons problématiques lors des études du transcriptome.

Table 4. Valeurs de Spécialisation des groupes expérimentaux.

	Condition	Spécialisation maximale	Spécialisation moyenne
LPS	PBS	2,32	0,018
	LPS	2,32	0,009
Tolérance fœto-maternelle	NP	2	0,037
	E4	2	0,028
	E6	2	0,011
	E8	2	0,006
	E10	2	0,007
	E11	2	0,011
	E12	2	0,013
Treg	noStim	3,16	0,027
	IL-2	3,16	0,025

Nous pouvons garder ici la même logique que celle développée par O. Martinez dans son papier (Martínez and Reyes-Valdés, 2008), à savoir utiliser cette mesure pour détecter des comportements singuliers. Par exemple, le jeu de données Tolérance fœto-maternelle montre que la spécialisation est forte pour tous les échantillons NP et E4, indiquant que les échantillons possèdent une part de leur transcriptome qui s'exprime de manière singulière. Retrouver des valeurs fortes pour l'ensemble des individus est donc synonyme d'une variabilité inter-individuelle forte elle aussi. À partir de E6, les valeurs de spécialisation diminuent pour tous les échantillons indiquant une diminution de la variabilité inter-individuelle. Celle-ci ne recommence à monter qu'à partir de E11, indiquant ainsi que les individus se font plus variables après E11.

Cette succession d'événements suggère que la perte de spécificité est induite par un processus biologique, celui de l'implantation fœtale et ce, dès les premiers jours après celle-ci. Puis le processus étant en place, un certain relâchement de la spécificité se met en place du fait des variations inter-individuelles, faisant qu'un individu serait en avance par rapport à un autre dans le processus de gestation par exemple.

En revanche, ceci est beaucoup moins clair pour le jeu de données Treg. En effet, si en moyenne la spécialisation diminue dans les échantillons IL-2, des différences notables apparaissent entre les individus. Ainsi, les patients 4, 7 et 9 possèdent des valeurs de spécialisation plus élevées que les autres patients. Cela indique qu'ils expriment des gènes de manière spécifique alors même que l'on s'attendrait à ce qu'ils soient peu divers du fait du stimulus, sur une population cellulaire triée qui plus est. Ces résultats indiquent donc qu'il existe, pour cette expérience, une variabilité inter-individuelle après le stimulus, comme si les différents individus ne réagissaient pas de la même manière. En réalité, les individus réagissent globalement de la même manière, en témoigne le nombre de gènes statistiquement différemment exprimés (1310, BH p-value < 0.001). En revanche ils ne répondent pas nécessairement avec la même intensité.

Nous avions vu, avec la publication de E. Glaab et R. Schneider (2012), que les auteurs envisageaient l'analyse de la variation des variances d'expression de groupe de gènes liés biologiquement. Dans la même logique, j'ai procédé à une analyse similaire, sur le jeu de données Treg, en utilisant des signatures moléculaires décrites dans la littérature pour tester leur enrichissement en gènes de forte spécificité. L'utilisation de la mesure de spécificité est particulièrement intéressante car elle permet de faire l'analyse pour chaque patient, là où l'utilisation de la variance ne peut se faire qu'en prenant le groupe dans son ensemble. Cela s'explique par le fait qu'il existe une cohérence biologique au fait de multiplier les fréquences relatives des gènes dans un échantillon par leurs mesures de la spécificité au sein d'un groupe d'échantillons. En effet, un gène est d'autant plus spécifique, qu'il est exprimé fortement dans un échantillon par rapport aux autres. Combiner les deux valeurs, met alors bien en évidence les transcrits qui caractérisent cette spécificité. Les résultats obtenus sur les différents patients nous indiquent que le profil du donneur 1 est enrichi, par rapport aux profils des patients 6 et 7, pour une signature liée à l'IL-2 (« Marzec_IL-2_Signaling_DN »). Il s'agit d'une signature qui est retrouvée sous-régulée dans des cultures de cellule dérivées de lymphomes cutanés à cellules T stimulés par de l'IL-2 par rapport à des LTCD4 sains (Marzec et al., 2008), celle-là même que l'on retrouve chez le donneur 4 après la stimulation. Ces gènes ont donc un lien direct avec l'IL-2 et pourtant se retrouvent enrichis chez un donneur sur les trois donneurs analysés. Cela indique qu'il existe des variations dans le niveau de réponse à l'IL-2 des différents donneurs. Toujours dans la même expérience, un autre exemple nous est donné par le donneur 7 : son profil est enrichi spécifiquement pour une signature liée à FOXP3 (« Gavin_Foxp3_Targets_Cluster_P6 »). Cette signature contient

des gènes qui dépendent de FOXP3 et qui jouent un rôle dans le cycle cellulaire (Gavin et al., 2007). Ils sont en particulier sous-régulés par la protéine PDE3B, exprimée par les LT naïfs mais pas les LTreg : *« […] reduced PDE3B levels implied cAMP-mediated adaptation in [LTreg] to chronic TCR and IL-2 signalling. Notably, naive T-cell fail to downregulate PDE3B upon TCR engagement, and we found that FOXP3 binds a highly conserved region in the first intron of Pde3b. Thus,, reduced PDE3B expression represents the first unique marker of [LTreg] and may be considered more definitive than FOXP3 itiself as it reports FOXP3 function. ».*

Quand on compare les donneurs ayant les plus fortes valeurs de spécialisation après le traitement à l'IL-2 (P4, P7 et P9), on retrouve cette signature de Gavin *et al.* chez ce même donneur 7. Une autre signature du même papier retrouvée chez le donneur 9, concerne des gènes fortement modulés par FOXP3 (positivement ou négativement) dans les LTreg par rapport au LTCD4. On trouve également chez le donneur 9, une signature, « Zheng_FoxP3_Targets_Up », qui rassemble pour sa part des gènes qui possèdent des promoteurs sur lesquels se fixe FOXP3. Ce dernier induit la surexpression des gènes de la signature dans les LTreg, à la fois dans le thymus, et donc pendant leur développement, et dans le sang périphérique. Toutes ces signatures impliquent FOXP3 dont a vu qu'il est au centre du programme suppressif des LTreg. Nous constatons pourtant qu'il existe des différences entre les donneurs quant à l'intensité de l'expression de ces gènes.

On le voit, ce type d'analyse permet de sortir des profils de spécificité pour chaque individu et s'inscrit donc dans la lignée des méthodes d'analyse d'échantillons individuels comme ssGSEA (Barbie et al., 2009), avec pour particularité de mettre en avant l'expression spécifique des gènes plutôt que leur sur- ou sous-expression. On peut néanmoins noter aussi quelques limites à l'emploi de cette mesure :

- Les résultats indiquent un enrichissement d'un grand nombre de signatures liées au cycle cellulaire, à la traduction et à la transcription. Ces signatures sont en effet composées de gènes qui sont fortement exprimés quelle que soit la condition observée. Ils ressortent donc en haut des listes ordonnées par leur forte abondance dans le transcriptome de l'échantillon.
- La spécificité ne fonctionne que dans un sens ; elle est maximale lorsqu'un gène est exprimé dans un échantillon seulement. En revanche, elle sera extrêmement faible si le gène est exprimé dans tous les échantillons sauf un. C'est logique d'un point de vu biologique : si le gène est exprimé dans plusieurs échantillons, il ne peut pas être spécifique. Néanmoins il

serait intéressant de pouvoir considérer la spécificité de l'absence d'expression au même titre que la spécificité de l'expression.

Enfin, il paraît indispensable de définir des méthodologies d'analyses systématiques des résultats d'enrichissement de signatures, telles que celles proposées dans (Nehar-Belaid et al., 2016), pour extraire au mieux l'information permettant de caractériser un individu par rapport à un autre.

AUTRES APPLICATIONS

DONNÉES APPARIÉES

Le choix de travailler sur le jeu de données Treg nous donne l'opportunité d'appliquer la méthodologie mise en place au cours de cette thèse sur des données appariées. L'intérêt des données appariées est de pouvoir observer de manière plus fine l'impact de la variabilité inter-individuelle. J'ai donc procédé à la transformation des données d'expression des gènes des différents échantillons en données de variations des gènes de chaque donneur. La variation d'un gène est alors calculée sur la base du log2 du ratio d'abondance dans la condition traitée à l'IL-2 par rapport à la condition contrôle. Le sens de la variation n'ayant pas d'importance pour l'analyse, j'ai transformé les log-ratios en valeurs absolues de ces log-ratios. Nous obtenons alors des valeurs de variations de gènes exprimant à quel point un transcriptome change d'un traitement à l'autre. Pour un donneur, plus le nombre de gènes avec une forte variation sera grand, plus le donneur verra son transcriptome différent entre les deux conditions.

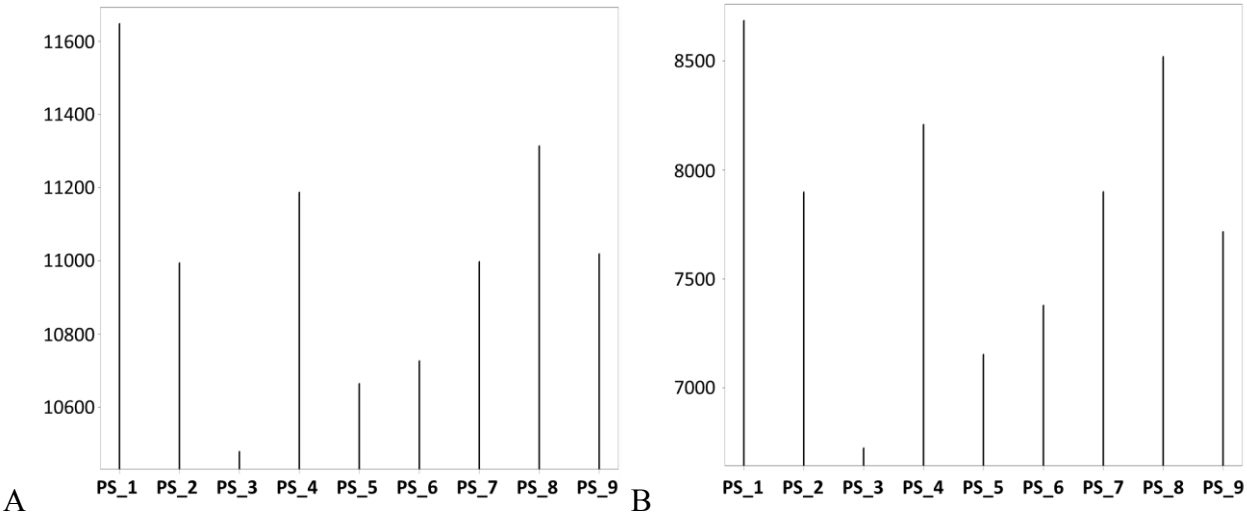


Figure 50. Diversité des log-ratios du jeu de données Treg.

A : Diversité individuelle des différences d'expression du jeu de données Treg calculée pour $q = 1$. B : Diversité individuelle des différences d'expression du jeu de données Treg calculée pour $q = 2$.

La Figure 50 montre que pour les indices de Shannon et de Simpson, les valeurs de diversité individuelle des distributions de variation d'expression sont très différentes d'un donneur à l'autre. En revanche, la topographie des valeurs de diversité est identique pour les deux valeurs de q, indiquant une stabilité dans la variation entre $q = 1$ et $q = 2$. Il n'est pas question ici de faire une analyse de la similarité, puisque nous n'avons plus deux groupes à comparer. En revanche, nous pouvons apprécier le comportement de la diversité des variations d'expression chaque donneur pour différentes valeurs de q (Figure 51). De manière intéressante, nous constatons que l'ordre des diversités sont très stables pour q compris entre 0 et 2. Le donneur 1, par exemple est systématiquement plus divers que les autres donneurs. À l'inverse le donneur 3 est systématiquement moins divers que les autres donneurs. À partir de q valant 3, les valeurs de diversité changent de comportement. Le donneur 1 devient alors l'un des donneurs avec la plus faible diversité. De la même manière que pour les analyses précédentes, plus nous montons dans les valeurs de q, plus nous regardons la diversité des gènes possédant une forte variation d'expression entre les deux conditions, en minimisant l'impact des gènes ayant des variations d'expression faibles.

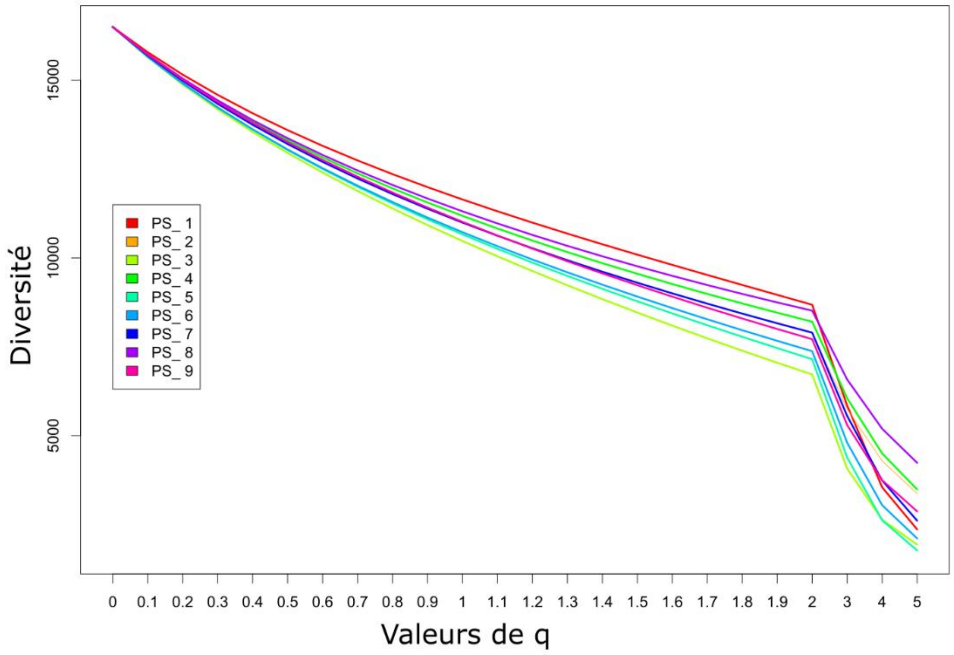


Figure 51. Diversité des log-ratio selon différentes valeurs de q.

Valeur de diversité des variations d'expression pour chaque donneur en fonction de la valeur de q.

On peut alors constater que les donneurs qui sont les moins divers pour les valeurs de q élevées (donneurs 1, 3, 5, 6, et 9) sont aussi ceux qui voient leur spécialisation chuter le plus fortement entre les deux conditions (Figure 47). Les donneurs 2 et 8 ont aussi des différences de spécialisation entre les deux conditions mais ils sont, dans la condition non traitée, les deux donneurs avec les plus petites valeurs de spécialisation. Les donneurs 4 et 7 échappent à cette logique, puisque le donneur 4 montre une diminution forte de la spécialisation après le traitement à l'IL-2, tandis que le donneur 7 ne montre pas de diminution.

C'est finalement l'analyse de la spécialisation des variations d'expressions qui nous apprends peut-être le plus sur le comportement des donneurs. Cette mesure nous donne l'importance des fortes variations d'expression de transcrits qui sont spécifiques d'un donneur par rapport aux autres (Figure 52). Nous pouvons constater alors que le donneur 3 possède un nombre particulièrement important de tels transcrits par rapport aux autres donneurs.

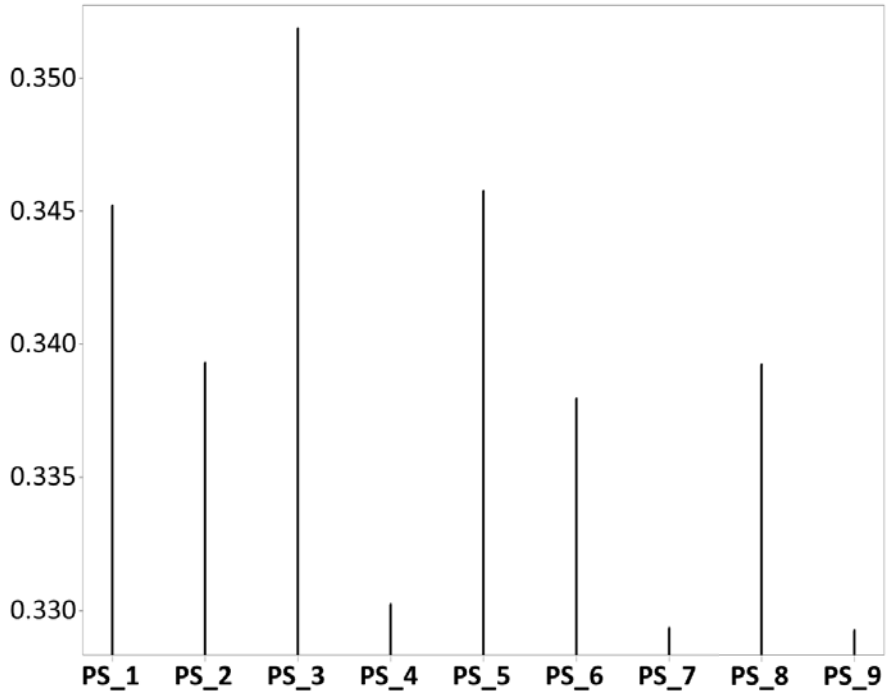


Figure 52. Spécialisation des variations de gènes au sein du jeu de données Treg.
Valeurs de spécialisation des variations d'expression de gènes de chaque donneur, calculées sur la base des log-ratios d'expression entre les échantillons traités à l'IL-2 et les échantillons non traités.

On notera que les donneurs qui possèdent les valeurs de spécialisation les plus faibles (P4, P7 et P9) sont aussi ceux qui possèdent les valeurs de spécialisation les plus fortes dans la condition

traitée à l'IL-2 (Figure 47). Il y a presque une inversion des distributions des valeurs de spécialisation entre ces deux figures.

L'étude des gènes dont la variation d'expression est particulièrement spécifique d'un échantillon à l'autre par la méthode GSEA montre qu'une majorité des signatures enrichies sont issues de l'analyse ACI du jeu de données. L'analyse par ACI cherche les regroupements de gènes qui expliquent aux mieux la structure du jeu de données. Il apparaît donc plausible qu'un certain nombre de ces signatures soient directement liées aux variations d'expression du jeu de données.

L'ordre des signatures ACI n'est pas le même en fonction des donneurs et on retrouve des différences dans les signatures enrichies pour les différents donneurs. Ainsi, les donneurs P1, P5 et P7 ont des profils de spécialisation enrichis pour des signatures du cycle cellulaire. Les donneurs P3, P7 et P9 partagent une signature liée à FOXP3 (« Ono_Foxp3_Target_Up ») où sont rassemblés des gènes dont l'expression augmente dans des cellules LTCD4 traitées pour augmenter l'expression de protéine FOXP3 par transduction. Cette signature n'est, pas du tout enrichie pour le donneur 1 (FDR q-value=0.57). Le donneur 4, quant à lui montre un enrichissement en signatures liées à la survie (« Hann_Resistance_To_Bcl2_Inhibitor_Dn ») et à l'inflammation (« Biocarta_Inflam_Pathway »). Le profil de donneur est par ailleurs, le seul à être enrichi pour une des deux signatures que j'ai obtenues à partir du jeu de données et qui contient les gènes trouvés statistiquement sur- ou sous-exprimés dans la condition traitée par rapport au contrôle : il s'agit de la signature des gènes sous-exprimés.

CLASSER LES TRANSCRITS

En écologie, il est possible de regrouper les espèces d'un prélèvement en classes et de calculer la diversité de ces classes à travers différents échantillons. La définition de ces classes se fait généralement sur la base de propriétés biologiques particulières. Deux espèces très proches phénotypiquement peuvent ainsi être traitées comme une seule espèce. Des espèces qui possèdent des fonctions similaires dans l'écosystème (être prédateur de gros gibier par exemple), peuvent également être traitées comme une seule espèce. L'idée nous est alors venue de considérer cette possibilité pour les analyses de transcriptome.

L'idée la plus séduisante est alors de considérer comme espèces, non plus les transcrits mais des groupes de transcrits liés biologiquement, comme le font Glaab et Schneider (2012) dans publication que je cite dans la partie « Discussion » du chapitre « Diversité ». Lors de nos recherches précédentes, nous nous sommes particulièrement attachés à analyser les transcriptomes grâce à la génération de signatures moléculaire (Dérian et al., 2016; Pham et al., 2014). Une signature moléculaire, telle que nous l'entendons, est une collection de transcrit qui possèdent des connections entre eux au regard du jeu de données, qu'ils agissent de manière coordonnée ou synergétique. Cependant un problème majeur empêche de raisonner ainsi. En effet, dans les systèmes biologiques, l'expression d'un gène n'est pas le seul fruit d'une action individuelle ; elle est souvent le résultat de l'activation de plusieurs voies différentes. Pour s'en convaincre, il suffit de regarder les signatures extraites par l'ACI sur un jeu de données quelconque. Ces signatures vont systématiquement partager des transcrits. Or, si nous envisageons de classer les transcrits en entités plus importantes, un gène ne peut pas se trouver dans deux entités en même temps, au risque de perturber complètement les résultats. Prenons l'exemple d'un gène dont l'expression vaut la valeur arbitraire de 1000, ce gène étant impliqué dans deux voies de signalisation. La première voie contribue à l'expression de ce gène à hauteur de 90%, l'autre voie à hauteur de 10%. Impliquer ce gène dans la première voie de signalisation n'engendre pas de problème majeur si ce n'est une petite surestimation de l'abondance de la voie de signalisation. En revanche, l'impliquer dans la deuxième voie de signalisation engendre une perturbation non négligeable du résultat. D'une manière générale, les erreurs s'accumulant avec le nombre de transcrits présents dans la voie de signalisation, il arrive très vite que l'estimation de l'abondance de la voie est fausse.

Nous avons alors envisagé d'utiliser l'ACI pour estimer la part de l'expression du gène expliquée par les différentes composantes. Cela nous permettrait alors d'utiliser cette part pour considérer le niveau d'expression d'un transcrit dans la signature. Mais, là aussi, plusieurs problèmes résident. Tout d'abord, il faut considérer que l'algorithme fastICA ne fournit pas systématiquement le même résultat pour deux analyses successives (voir le chapitre Extraire l'information). Dans notre méthodologie, nous suivons alors les recommandations de P. Chiapetta et procédons à plusieurs itérations. Cette procédure induit généralement la découverte d'un nombre de composantes supérieur à celui initialement cherché par l'algorithme. D'expérience, ce nombre de composantes est d'autant plus grand que le jeu de données est complexe, au sens d'une faible détermination de classes biologiques fortes dans le jeu de données (donneurs sains *versus* patients, témoins *versus*

traitement, …). Ainsi, la détermination de la part d'expression d'un gène dans les différentes signatures est automatiquement fausse car la somme des contributions serait alors systématiquement supérieure à 100%. Le deuxième problème se situe au niveau des signatures elles-mêmes : comment évaluer la part d'expression d'un gène d'une signature issue d'ACI dans un autre jeu de données que celui dont elle est issue ? Le dernier point est celui de la signification des signatures. Par définition, une signature, qu'elle soit issue de l'ACI ou d'une analyse comparative de l'expression par un test statistique, est fortement liée au jeu de données dont elle est issue. Même si l'annotation par des outils d'enrichissement, comme DAVID® ou Ingenuity®, aident grandement, il est peu souhaitable d'assurer qu'une signature fortement enrichie pour telle voie de signalisation est une signature universelle de cette voie. À titre d'exemple, il y a quelques années, nous avons travaillé sur la détection d'une signature LTreg (Pham et al., 2014). Avant de commencer nos travaux, nous avons recherché les signatures décrites dans la littérature. Cette demi-douzaine de signatures, ne partageaient que très peu de gènes en commun.

Considérant ces différents points, j'ai alors envisagé d'utiliser des signatures de voies de signalisation provenant de KEGG. Les gènes impliqués dans ces voies de signalisation sont associés suite à l'agglomération des connaissances scientifiques sur les domaines impliqués. Elles ne sont donc pas issues d'un seul jeu de données. Considérant que je ne souhaitais pas de redondance de gènes dans les différentes voies, j'ai systématiquement supprimé les gènes en communs. Là aussi, la limite de la technique nous empêche de considérer cette méthode. En effet, quand on étudie le système immunitaire, enlever les gènes redondants de plusieurs voies de signalisations impliquées dans ce système, revient à enlever plus de la moitié des gènes. Cela est un dernier argument contre l'utilisation des signatures pour la classification des gènes en entités plus grandes : l'utilisation des signatures n'aboutit jamais à l'utilisation de l'ensemble des gènes d'un jeu de données, ne serait-ce parce que nous ne connaissons pas le rôle de chaque transcrit. Nous perdons ainsi une part de l'information mesurée, alors que le but de l'utilisation des indices de diversité est justement d'analyser les données d'une manière globale et non une partie d'entre elle.

Dans ces conditions, la seule classification pertinente pour moi est celle de la taille des gènes. L'hypothèse développée sous cet angle est la suivante : pour réagir rapidement, un système doit augmenter de manière rapide le nombre de transcrits issus de gènes d'intérêt. Or il existe une grande disparité dans la taille des gènes, impliquant une différence du temps nécessaire à leur transcription.

Les gènes impliqués dans la réponse rapide du système pourraient alors être généralement plus courts.

Pour tester cette hypothèse, j'ai classé en huit niveaux de taille, les gènes de souris issus de la base de données BioMart (http://www.biomart.org/). J'ai ainsi obtenu les tailles de 26368 gènes connus de la souris. L'assignation des gènes à l'une des huit classes s'est effectuée à partir de la distribution des tailles sur une classification hiérarchique (distance euclidienne, agglomération Ward). La Figure 53 donne le découpage des longueurs de gènes pour la souris. Les gènes dont la taille est strictement supérieure à 250 000 sont classés dans une seule catégorie, les gènes très longs.

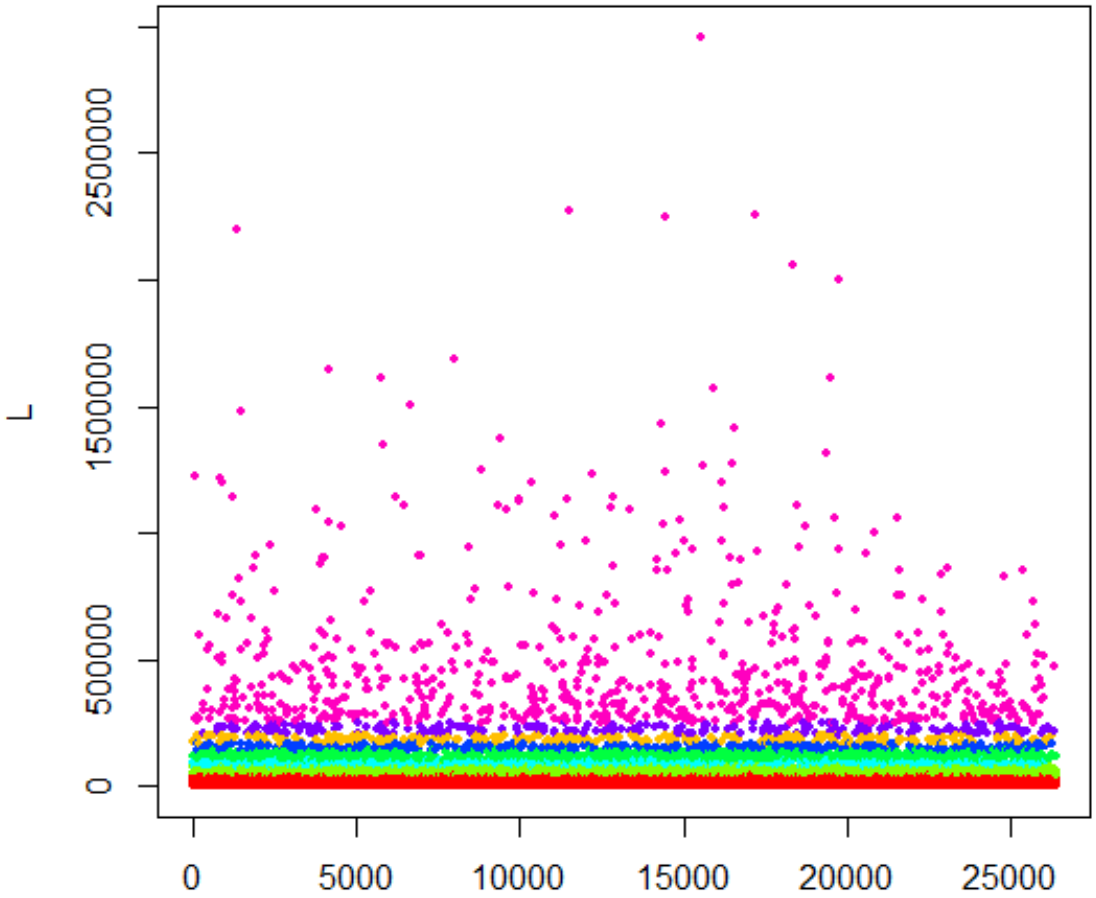


Figure 53. Analyse de la distribution des tailles de gènes.

Taille (en ordonnée) des 26368 gènes de la souris (en abscisses) issus de la base de données MGI (*Mouse Genome Informatics*). Les gènes d'une taille supérieure à 250 000 nucléotides sont placés dans la même classe (rose). Les autres classes respectent le découpage en sept clusters de la classification hiérarchique des données (distance Euclidienne/agglomération Ward).

En sommant les abondances des gènes assignés à chaque classe, on obtient une nouvelle distribution des abondances pour huit entités seulement. Nous appliquons alors la méthodologie développée dans cette thèse à ce nouveau jeu de données. Les résultats présentés ci-dessous sont issus des jeux de données LPS et Tolérance fœto-maternelle.

La Figure 54 montre les résultats de la mesure de la diversité de Shannon. La diversité diminue systématiquement dans les deux jeux de données dès l'application d'un stimulus. Moins de diversité signifie qu'une ou plusieurs des classes prennent une part plus importante de l'abondance après le stimulus. Dans ces cas précis, c'est la classe des transcrits les plus petits qui augmentent, les autres classes restent stables ou diminuent.

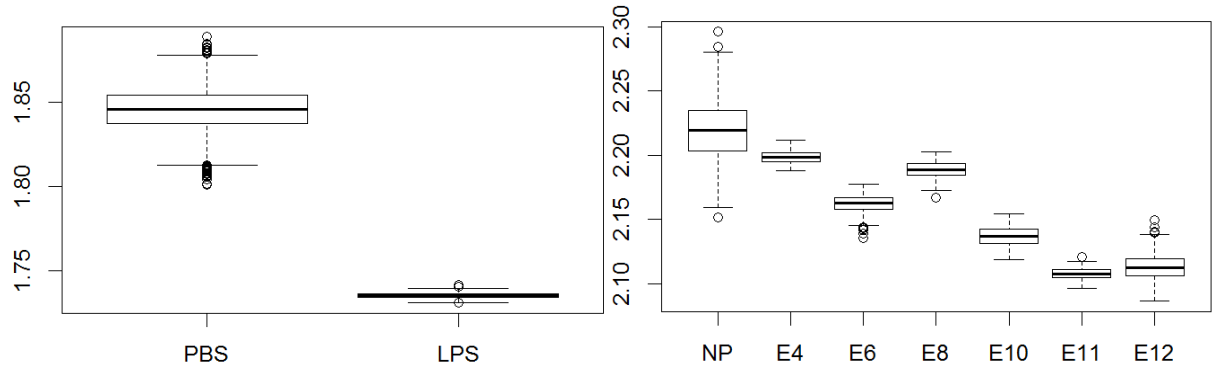


Figure 54. Longueur des gènes : diversité selon l'indice de Shannon.

Distribution de la diversité (pour $q = 1$) individuelle de l'abondance des gènes classés selon leur taille dans les jeux de données LPS (gauche) et Tolérance fœto-maternelle (droite), par la représentation des boîtes à moustaches. Le trait noir épais représente la médiane des moyennes. La boîte inclus 50% des valeurs, les traits en pointillés de chaque côté de la boîte représentent chacun 1,5 fois l'espace interquartile. Les cercles noirs sont des valeurs extrêmes.

Nous retrouvons par ailleurs en Figure 55 un comportement de la similarité des individus qui nous est maintenant familier ; la similarité augmente de manière drastique après l'induction d'un stimulus. Enfin, nous voyons en Figure 56 que cette augmentation de la similarité est accompagnée d'une diminution de la spécialisation des échantillons à l'intérieur des groupes expérimentaux.

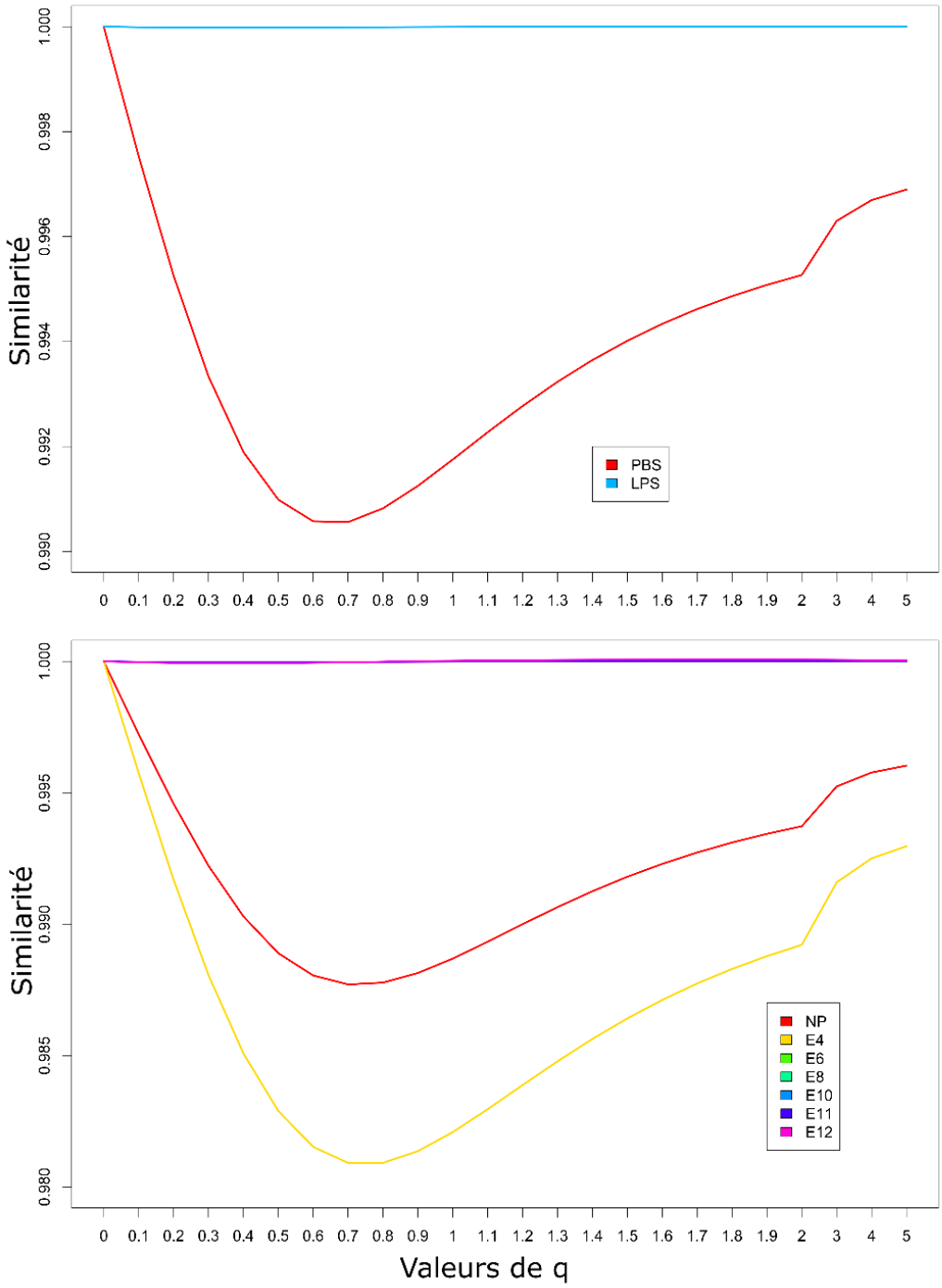


Figure 55. Longueur des gènes : Similarité.

La diversité des échantillons est calculée pour plusieurs valeurs de q ($q \in [0;5]$). La similarité des diversités pour chaque groupe est alors calculée selon l'Équation 23.

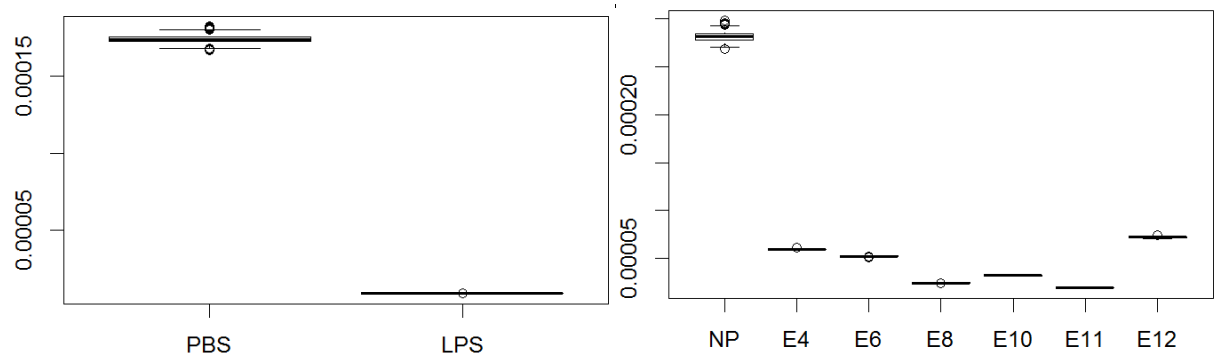


Figure 56. Longueur des gènes : Spécialisation.

Distribution des spécialisations moyennes, calculées à partir de mille tirages aléatoires pour chaque groupe expérimental des jeux de données LPS (gauche) et Tolérance fœto-maternelle (droite), par la représentation des boîtes à moustaches. Le trait noir épais représente la médiane des moyennes. La boîte inclus 50% des valeurs, les traits en pointillés de chaque côté de la boîte représentent chacun 1,5 fois l'espace interquartile. Les cercles noirs sont des valeurs extrêmes.

Cette dernière section nous montre comment il est possible d'adapter les mesures de diversité et de spécialisation pour répondre à des questions biologiques particulières. L'utilisation de données appariées est particulièrement intéressante car si nous avions vu que les mesures de diversité et de spécialisation avaient déjà la capacité de faire ressortir des phénomènes singuliers sur les transcriptome d'un échantillon, ils démontrent ici leur intérêt pour mettre en avant des modifications dynamiques singulières (avant/après stimulation). Ceci offre la possibilité de réaliser des analyses fines du comportement individuel face à la stimulation.

Effectuer ce genre d'analyse sur des regroupements de gènes est très certainement intéressant mais les limitations techniques sont un obstacle à une étude directe des voies de signalisation par exemple. Notre laboratoire continue de rechercher des moyens d'exploiter cette méthode pour développer de nouvelles hypothèses biologiques.

DISCUSSION GÉNÉRALE

La biologie des systèmes est devenue un une composante importante de la recherche en biologie et révolutionne la manière d'appréhender des objets biologiques très différents (Bowen et al., 2016; Plenge, 2016). Les développements, de plus en plus sophistiqués, en biotechnologies et bio-informatiques permettent d'appréhender la variabilité inter-individuelle sous un jour nouveau. Les données à haut-débit illustrent cela en analysant un nombre d'événements tel qu'il est possible d'envisager de mesurer la quasi-totalité de l'information d'un système biologique même si l'analyse de ces données restent un challenge (Veneziano et al., 2016).

L'analyse du transcriptome est un bon exemple de ces évolutions. Avec l'avènement de la technique de PCR (Mullis and Faloona, 1987), les premières analyses se concentraient péniblement sur une dizaine de gènes pour une même expérience, du fait de la lourdeur du système et du coup des réactifs. Néanmoins, grâce à cela, les biologistes ont mesuré pour la première fois les variations de l'expression de certains gènes dans des conditions expérimentales différentes et de développer des protocoles de détection de virus (Kwok et al., 1987). Les puces à ADN vont révolutionner le domaine en permettant l'analyse de plusieurs dizaines de milliers de transcrits en même temps au prix d'une précision de la mesure de moindre qualité (Petrik, 2001; Wildsmith and Elcock, 2001). Mais là encore, la technologie présente une limite importante ; elle ne mesure que ce qu'une fraction connue du système. Ce défaut, tout relatif au regard de la quantité d'information déjà disponible, sera définitivement éliminé par la génération des systèmes de séquençage à haut-débit. Cette fois, il n'y a plus de limitation du nombre de transcrits : pour une profondeur suffisante, l'utilisateur peut espérer avoir une mesure de la quantité de l'ensemble des transcrits présents dans un échantillon.

Les analyses du transcriptome telles qu'elles sont pratiquées en routine dans les laboratoires se concentrent sur l'extraction et la compréhension de sous-ensembles de transcrits, formant une signature moléculaire, pour peu que ce groupe de transcrits permette de discriminer des conditions expérimentales différentes (Touzot et al., 2015). Nous avons d'ailleurs nous-même participé à l'élaboration de stratégies d'analyses qui vont dans le sens de cette démarche (Pham et al., 2014). Il nous est alors apparu intéressant de ne plus considérer le transcriptome à travers le prisme des

analyses standards effectuées dans le domaine (recherche de gènes significativement différemment exprimés, de signatures moléculaires statistiquement modulées, …), mais de monter encore d'un niveau d'échelle, comparer le transcriptome dans sa globalité. Le transcriptome est assurément un système biologique, tant ses composants, les transcrits, sont liés les uns aux autres de manières directes ou indirectes (Conway and Schoolnik, 2003), et ce sont justement ces interactions complexes qui nous incitent à considérer les mesures d'expression comme un seul et même objet et non plus comme une succession de mesures individuelles. L'utilisation des indices de diversité et de la mesure de spécialisation prend alors tout son sens car ces indices donnent une mesure du transcriptome dans son intégralité.

L'aspect principal de cette thèse concerne l'étude de la variabilité inter-individuelle. L'intérêt de la variabilité inter-individuelle est cruciale car elle s'inscrit dans le développement d'un nouveau champ médical : la médecine personnalisée. À titre d'exemple, dans notre laboratoire, une étude de la variabilité inter-individuelle de lignées de souris au cours du vieillissement montre que les perturbations du système immunitaire engendrées par le vieillissement sont très différentes en quantité et qualité d'un individu à l'autre. Les raisons de telles différences s'expliquent notamment par l'historicité du système immunitaire, unique pour chaque individu, mais aussi le fonds génétique des souris étudiées (Pham, 2013).

Notre laboratoire travaille depuis plusieurs années en étroite collaboration avec des services cliniques pour étudier l'impact de nouvelles molécules thérapeutiques sur des patients atteints de maladies auto-immunes (Bonnet et al., 2016; Rosenzwajg et al., 2015; Saadoun et al., 2011; Terrier et al., 2012). Les résultats ont montré qu'il existe une variabilité de réponse à ces traitements. Historicité du système et particularismes génétiques jouent probablement là aussi un rôle important. Néanmoins, l'injection d'un produit, quel qu'il soit, dans un organisme, ne peut pas s'effecteur sans un changement d'état du système si petit soit-il. Certaines expériences, utilisées dans cette thèse, montrent que l'induction d'un stimulus semble réduire la différence entre les individus, par rapport à une situation physiologique. En réalité, même certaines de nos situations physiologiques sont des situations de stimulation (injection de PBS, culture cellulaire *in vitro*). Dans l'article que nous avons publié cette année (Dérian et al., 2016), nous montrons que l'une des signatures qui permet de discriminer les vecteurs qui induisent une bonne réponse immunitaire de ceux qui n'en induisent pas est une signature riche en annotation immunologique. Rien de plus normal, sauf que les vecteurs qui induisent une mauvaise réponse immunitaire quelques jours après

l'injection du vecteur, présentent eux aussi ne sur-expression de cette signature. La différence entre les deux classes de vecteurs se fait alors sur l'intensité de cette sur-expression. L'impact de l'injection est donc identique pour tous les vecteurs, mais pas l'intensité de la réponse. Si l'impact est identique, cela signifie que les voies de signalisations, notamment, mises en jeu durant la réponse sont aussi les mêmes pour tout ou partie. Il en résulte que quelle que soit l'intensité de la réponse, il existe une dérive du système vers un point commun pour tous les individus et, par conséquent, une diminution de la variabilité inter-individuelle après un stimulus. On peut y voir une analogie sociologique avec le phénomène grèves nationales : si l'on considère la population française dans son ensemble, la diversité socio-économique entre les individus est très importante et l'histoire des individus est unique. Il en résulte que la compréhension du monde qui nous entoure est elle aussi, pour ainsi dire, unique à chaque individu tant les possibilités d'interprétation des événements mondiaux sont nombreuses. Pourtant, il suffit d'un stimulus, en l'occurrence une proposition de loi, pour agglomérer, très rapidement, une partie de la population vers un seul et même objectif, le refus de cette loi, alors même que les individus sont probablement en désaccord sur d'autre sujet. La variabilité inter-individuelle initiale se réduit, pour un temps, à un discours non nécessairement commun mais orienté vers un même but.

L'analyse de la variabilité du transcriptome montre son intérêt dans un tout autre contexte biologique, les études sur le cancer (Jiang et al., 2014; Nguyen et al., 2016). Partant du principe que les tumeurs cancéreuses sont constituées de cellules à la diversité génétique particulièrement importante (De Sousa E Melo et al., 2013), les auteurs ont cherché à comprendre en quoi cette diversité pouvait marqué un critère pour différencier des groupes de tumeurs. Il ressort de ces études que les tumeurs qui voient leurs diversités génétiques mais aussi transcriptomiques être les plus importantes, sont celles qui résistent le mieux aux traitements anti-cancer. La diversité transcriptomique importante d'une tumeur est donc synonyme de mauvais pronostic. Ces modèles sont différents de ceux que j'analyse dans cette thèse. En effet, il s'agit de systèmes cellulaires connus pour leur instabilité génétique et leur division non contrôlée, faisant des tumeurs des modèles d'évolution accélérée. L'augmentation de la diversité des cellules revient alors à ce que j'écrivais en introduction, un moyen d'ouvrir un champ des possibles plus vaste pour répondre à un stress.

Dans ces articles, les jeux de données ont été analysées par des combinaisons de méthodes basées sur le coefficient de variation ou des distances deux à deux. Je propose dans cette thèse l'utilisation

des indices de diversité pour l'analyse du transcriptome. Ces indices, très utilisés en écologie notamment, permettent de prendre en considération l'ensemble des transcrits et de donner une représentation l'information portée par ces transcrits. O. Martinez (Martínez and Reyes-Valdés, 2008) a montré toute l'efficacité de telles mesures pour la compréhension de la variabilité inter-organe chez l'homme : certains organes sont plus divers ou plus spécialisés que d'autres. Notre démarche a pour but d'évaluer cette méthodologie dans un contexte différent de celui du papier d'O. Martinez : l'impact d'un stimulus sur un groupe d'échantillons expérimentalement homogènes.

Les travaux décrits ci-dessus nous montrent qu'il existe une différence de diversité du transcriptome entre les individus. Cette diversité est systématiquement perturbée après l'action d'un stimulus. Cette perturbation est orientée dans un sens unique pour une même condition mais non nécessairement identique à travers différentes expériences. Ainsi, un transcriptome peut voir sa diversité augmenter après une stimulation, indiquant que l'abondance des transcrits est plus homogène dans cette dernière condition. Cela se traduit par plus de gènes exprimés que dans la condition sans stimulation. L'augmentation du nombre de gènes exprimés peut résulter d'un phénomène de complexification du système comme on peut le constater dans l'expérience Tolérance fœto-maternelle. Les transformations locales et le développement du fœtus augmentent la diversité des cellules dans l'échantillon et par conséquent, la diversité du système. Un phénomène un peu différent intervient dans le jeu de données Treg. La diversité augmente aussi par l'augmentation du nombre de gènes exprimés dans le système. Le nombre de gènes détectés avec une p-value de détection inférieure à 0.001 est supérieure pour la condition traitée pour six donneurs sur neuf. Le phénomène traduit ici l'activation de voies de signalisations induites par le stimulus en sus des voies déjà impliquées dans le maintien du phénotype LTreg.

Ce changement d'état de la diversité du transcriptome s'accompagne d'une diminution de la variabilité des diversités au sein du groupe expérimental. Le calcul de la similarité utilisée ici n'est pas le seul à envisager pour évaluer la cohésion de différents échantillons. Par exemple, nous aurions pu envisager la mesure de la dispersion des diversités en calculant une distance entre les paires d'échantillons. C'est le diamètre d'ordre r pour $r = 0$, x_i et x_j étant deux points d'un système de points $X = \{x_1, \dots x_i, \dots, x_j, \dots, x_n\}$:

Équation 27:

$$D_r = \left(\frac{2}{n(n-1)} \sum_{i<j} d(x_i, x_j)^r \right)^r$$

Par ailleurs il existe aussi une mesure dite de divergence de Kulback-Liebler (Kullback and Leibler, 1951) qui détermine la divergence entre la diversité γ d'un système et les diversités α qui le composent :

Équation 28:

$$D_{KL} = \left| H_\gamma - H_\alpha \right|$$

H_γ et H_α sont respectivement les entropies, de Shannon par exemple, du système dans son entier et d'un échantillon en particulier. L'usage de l'Équation 23 dans cette thèse tient au fait qu'elle est, comme l'équation l'Équation 28, conçue pour analyser des mesures de diversité. L'Équation 23 est par ailleurs établie pour tenir compte d'un nombre d'échantillons différents dans les groupes expérimentaux observés. La contrainte de cette mesure est qu'elle est conditionnée à ce que la diversité γ soit plus forte que les diversités α. Or, je n'ai pas de certitude que cela soit systématiquement le cas pour le transcriptome. Un moyen d'assurer cette condition serait de calculer la diversité γ non plus à partir des échantillons d'un groupe expérimental mais à partir de l'ensemble des échantillons de l'expérience. Si cela ne suffisait pas, il faudrait envisager l'utilisation de l'Équation 27 ou de l'Équation 28 pour le calcul de l'homogénéité des diversités des échantillons.

Un résultat particulièrement prometteur de ce travail est le comportement de la spécialisation des échantillons après la stimulation. Celle-ci est diminuée de manière très stable chez les différents individus. Contrairement à une analyse de la variance qui indiquerait seulement un changement d'état du système, comme le suggèrent les analyses en composantes principales effectuées sur les différents jeux de données utilisés dans cette thèse, elle indique à la fois le changement orienté du système vers un point commun et elle met en évidence des particularismes individuels. L'utilisation de la mesure de spécificité des gènes, au sein d'un groupe expérimental, comme base pour une étude d'enrichissement de signatures nous montre qu'il est possible d'extraire des informations d'autant plus intéressantes que ces signatures ne partagent pas nécessairement de gènes avec ceux découverts par les méthodes statistiques traditionnelles. La possibilité d'utiliser cet outil sur des

données transformées comme nous l'avons fait (log-ratio des données appariées) fait de cette mesure une nouvelle arme à l'arsenal du bio-statisticien.

La mesure de la spécificité est basée sur une adaptation de l'entropie de Shannon. Or, on le sait, cette entropie donne du poids aux événements de faibles intensités. Nous pouvons imager utiliser une version dérivée de la diversité d'ordre 2, tel que :

Équation 29:

$$S_i = \frac{1}{t}\left(\sum_{j=1}^{t}\left(\frac{p_{ij}}{p_i}\right)^2\right)$$

Cette variation pourrait donner des résultats intéressants sur des données pour lesquelles le nombre d'échantillons est particulièrement conséquent en limitant l'importance des échantillons de faible abondance.

Cette élévation de l'ordre q est peut-être un élément très intéressant dans ces outils (diversité et spécificité), car il permet de prendre en compte l'ensemble du système mais autorise aussi l'étude de ce système à des degrés différents de tolérance aux faibles abondances. Ce « découpage » procure une meilleure compréhension des variations d'un système : sont-ce plutôt les transcrits de forte abondance ou de faible qui sont impactés par le stimulus ou qui expliquent la diversité du système ?

Les applications liées à ces mesures sont d'une manière générale aussi diverses qu'il existe des études du transcriptome différentes. On peut par exemple penser à comparer la diversité du transcriptome dans d'autres modèles biologiques, comme c'est déjà le cas avec le microbiote (Birzele et al., 2016; Opstelten et al., 2016). Couplés aux méthodes d'analyse du transcriptome à partir d'échantillon de sang (Chaussabel, 2015), ces mesures devraient permettre de répondre de manière nouvelle aux variations des différents systèmes mis en jeu lors d'une pathologie. Je pense à la diversité des transcriptomes, mais aussi à celle du protéome et des abondances de différents types cellulaires. La spécificité, quant à elle, répond à la volonté de la recherche actuelle pour une meilleure mesure et la compréhension des modifications à l'échelle de l'individu.

BIBLIOGRAPHIE

Abumaree, M.H., Stone, P.R., and Chamley, L.W. (2006). The effects of apoptotic, deported human placental trophoblast on macrophages: possible consequences for pregnancy. J. Reprod. Immunol. *72*, 33–45.

Aghion, J., and Poirier, F. (2000). La biologie de l'implantation. Médecine Sci. *16*, 324–328.

Aluvihare, V.R., Kallikourdis, M., and Betz, A.G. (2004). Regulatory T cells mediate maternal tolerance to the fetus. Nat. Immunol. *5*, 266–271.

Barber, E.M., and Pollard, J.W. (2003). The uterine NK cell population requires IL-15 but these cells are not required for pregnancy nor the resolution of a Listeria monocytogenes infection. J. Immunol. Baltim. Md 1950 *171*, 37–46.

Barbie, D.A., Tamayo, P., Boehm, J.S., Kim, S.Y., Moody, S.E., Dunn, I.F., Schinzel, A.C., Sandy, P., Meylan, E., Scholl, C., et al. (2009). Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature *462*, 108–112.

Bennet, A.F. (1987). Interindividual variability : an underutilized resource.

Birzele, L.T., Depner, M., Ege, M.J., Engel, M., Kublik, S., Bernau, C., Loss, G.J., Genuneit, J., Horak, E., Schloter, M., et al. (2016). Environmental and mucosal microbiota and their role in childhood asthma. Allergy.

Blake, W.J., KAErn, M., Cantor, C.R., and Collins, J.J. (2003). Noise in eukaryotic gene expression. Nature *422*, 633–637.

Blazewicz, J., Borowski, M., Chaara, W., Kedziora, P., Klatzmann, D., Lukasiak, P., Six, A., and Wojciechowski, P. (2012). GeVaDSs - decision support system for novel Genetic Vaccine development process. BMC Bioinformatics *13*, 91.

Bochtler, T., Fröhling, S., and Krämer, A. (2015). Role of chromosomal aberrations in clonal diversity and progression of acute myeloid leukemia. Leukemia *29*, 1243–1252.

Boedigheimer, M.J., Wolfinger, R.D., Bass, M.B., Bushel, P.R., Chou, J.W., Cooper, M., Corton, J.C., Fostel, J., Hester, S., Lee, J.S., et al. (2008). Sources of variation in baseline gene expression levels from toxicogenomics study control animals across multiple laboratories. BMC Genomics *9*, 285.

Bonnet, B., Vigneron, J., Levacher, B., Vazquez, T., Pitoiset, F., Brimaud, F., Churlaud, G., Klatzmann, D., and Bellier, B. (2016). Low-Dose IL-2 Induces Regulatory T Cell-Mediated Control of Experimental Food Allergy. J. Immunol. Baltim. Md 1950 *197*, 188–198.

Bowen, J.R., Ferris, M.T., and Suthar, M.S. (2016). Systems biology: A tool for charting the antiviral landscape. Virus Res. *218*, 2–9.

Cappuccio, A., Zollinger, R., Schenk, M., Walczak, A., Servant, N., Barillot, E., Hupé, P., Modlin, R.L., and Soumelis, V. (2015). Combinatorial code governing cellular responses to complex stimuli. Nat. Commun. *6*, 6847.

Chao, A., and Jost, L. (2015). Estimating diversity and entropy profiles via discovery rates of new species. Methods Ecol. Evol. *6*, 873–882.

Chaussabel, D. (2015). Assessment of immune status using blood transcriptomics and potential implications for global health. Semin. Immunol. *27*, 58–66.

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.

Cheng, G., Yu, A., and Malek, T.R. (2011). T-cell tolerance and the multi-functional role of IL-2R signaling in T-regulatory cells. Immunol. Rev. *241*, 63–76.

Chiappetta, P., Roubaud, M.C., and Torrésani, B. (2004). Blind source separation and the analysis of microarray data. J. Comput. Biol. J. Comput. Mol. Cell Biol. *11*, 1090–1109.

Chubb, J.R., Trcek, T., Shenoy, S.M., and Singer, R.H. (2006). Transcriptional pulsing of a developmental gene. Curr. Biol. CB *16*, 1018–1025.

Conway, T., and Schoolnik, G.K. (2003). Microarray expression profiling: capturing a genome-wide portrait of the transcriptome. Mol. Microbiol. *47*, 879–889.

Darrasse-Jèze, G., Bergot, A.-S., Durgeau, A., Billiard, F., Salomon, B.L., Cohen, J.L., Bellier, B., Podsypanina, K., and Klatzmann, D. (2009). Tumor emergence is sensed by self-specific CD44hi memory Tregs that create a dominant tolerogenic environment for tumors in mice. J. Clin. Invest. *119*, 2648–2662.

Dérian, N., Bellier, B., Pham, H.P., Tsitoura, E., Kazazi, D., Huret, C., Mavromara, P., Klatzmann, D., and Six, A. (2016). Early Transcriptome Signatures from Immunized Mouse Dendritic Cells Predict Late Vaccine-Induced T-Cell Responses. PLoS Comput. Biol. *12*, e1004801.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. Ann. Stat. *7*, 1–26.

Ege, M.J., Frei, R., Bieli, C., Schram-Bijkerk, D., Waser, M., Benz, M.R., Weiss, G., Nyberg, F., van Hage, M., Pershagen, G., et al. (2007). Not all farming environments protect against the development of asthma and wheeze in children. J. Allergy Clin. Immunol. *119*, 1140–1147.

Ellis, J.A., Gow, S.P., Waldner, C.L., Shields, S., Wappel, S., Bowers, A., Lacoste, S., Xu, Z., and Ball, E. (2016). Comparative efficacy of intranasal and oral vaccines against Bordetella bronchiseptica in dogs. Vet. J. Lond. Engl. 1997 *212*, 71–77.

Elowitz, M.B., Levine, A.J., Siggia, E.D., and Swain, P.S. (2002). Stochastic gene expression in a single cell. Science *297*, 1183–1186.

Fahrner, J.A., Liu, R., Perry, M.S., Klein, J., and Chan, D.C. (2016). A novel de novo dominant negative mutation in DNM1L impairs mitochondrial fission and presents as childhood epileptic encephalopathy. Am. J. Med. Genet. A. *170*, 2002–2011.

Feng, M., Yang, Z., Pan, L., Lai, X., Xian, M., Huang, X., Chen, Y., Schröder, P.C., Roponen, M., Schaub, B., et al. (2016). Associations of Early Life Exposures and Environmental Factors With Asthma Among Children in Rural and Urban Areas of Guangdong, China. Chest *149*, 1030–1041.

Ferraro, A., D'Alise, A.M., Raj, T., Asinovski, N., Phillips, R., Ergun, A., Replogle, J.M., Bernier, A., Laffel, L., Stranger, B.E., et al. (2014). Interindividual variation in human T regulatory cells. Proc. Natl. Acad. Sci. U. S. A. *111*, E1111–1120.

Ferretti, C., Bruni, L., Dangles-Marie, V., Pecking, A.P., and Bellet, D. (2007). Molecular circuits shared by placental and cancer cells, and their implications in the proliferative, invasive and migratory capacities of trophoblasts. Hum. Reprod. Update *13*, 121–141.

Feugeas, J.-P., Tourret, J., Launay, A., Bouvet, O., Hoede, C., Denamur, E., and Tenaillon, O. (2016). Links between Transcription, Environmental Adaptation and Gene Variability in Escherichia coli: Correlations between Gene Expression and Gene Variability Reflect Growth Efficiencies. Mol. Biol. Evol.

Foissac, F., Urien, S., Hirt, D., Frange, P., Chaix, M.-L., Treluyer, J.-M., and Blanche, S. (2011). Pharmacokinetics and virological efficacy after switch to once-daily lopinavir-ritonavir in treatment-experienced HIV-1-infected children. Antimicrob. Agents Chemother. *55*, 4320–4325.

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.

Gavin, M.A., Rasmussen, J.P., Fontenot, J.D., Vasta, V., Manganiello, V.C., Beavo, J.A., and Rudensky, A.Y. (2007). Foxp3-dependent programme of regulatory T-cell differentiation. Nature *445*, 771–775.

Genbacev, O., Zhou, Y., Ludlow, J.W., and Fisher, S.J. (1997). Regulation of human placental development by oxygen tension. Science *277*, 1669–1672.

Ginsburg, G.S., and Willard, H.F. (2009). Genomic and personalized medicine: foundations and applications. Transl. Res. J. Lab. Clin. Med. *154*, 277–287.

Glaab, E., and Schneider, R. (2012). PathVar: analysis of gene and protein expression variance in cellular pathways using microarray data. Bioinformatics *28*, 446–447.

Golding, I., Paulsson, J., Zawilski, S.M., and Cox, E.C. (2005). Real-time kinetics of gene activity in individual bacteria. Cell *123*, 1025–1036.

Harper, C.V., Finkenstädt, B., Woodcock, D.J., Friedrichsen, S., Semprini, S., Ashall, L., Spiller, D.G., Mullins, J.J., Rand, D.A., Davis, J.R.E., et al. (2011). Dynamic analysis of stochastic transcription cycles. PLoS Biol. *9*, e1000607.

Hicks, S.C., and Irizarry, R.A. (2014). When to use Quantile Normalization?

Holtan, S.G., Creedon, D.J., Haluska, P., and Markovic, S.N. (2009). Cancer and pregnancy: parallels in growth, invasion, and immune modulation and implications for cancer therapeutic agents. Mayo Clin. Proc. *84*, 985–1000.

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

Hyvärinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans. Neural Netw. Publ. IEEE Neural Netw. Counc. *10*, 626–634.

Jiang, T., Shi, W., Natowicz, R., Ononye, S.N., Wali, V.B., Kluger, Y., Pusztai, L., and Hatzis, C. (2014). Statistical measures of transcriptional diversity capture genomic heterogeneity of cancer. BMC Genomics *15*, 876.

Jin, L.-P., Chen, Q.-Y., Zhang, T., Guo, P.-F., and Li, D.-J. (2009). The CD4+CD25 bright regulatory T cells and CTLA-4 expression in peripheral and decidual lymphocytes are down-regulated in human miscarriage. Clin. Immunol. Orlando Fla *133*, 402–410.

Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostat. Oxf. Engl. *8*, 118–127.

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

Keim, P., Gruendike, J.M., Klevytska, A.M., Schupp, J.M., Challacombe, J., and Okinaka, R. (2009). The genome and variation of Bacillus anthracis. Mol. Aspects Med. *30*, 397–405.

Klatzmann, D., and Abbas, A.K. (2015). The promise of low-dose interleukin-2 therapy for autoimmune and inflammatory diseases. Nat. Rev. Immunol. *15*, 283–294.

Koopman, L.A., Kopcow, H.D., Rybalov, B., Boyson, J.E., Orange, J.S., Schatz, F., Masch, R., Lockwood, C.J., Schachter, A.D., Park, P.J., et al. (2003). Human decidual natural killer cells are a unique NK cell subset with immunomodulatory potential. J. Exp. Med. *198*, 1201–1212.

Koreth, J., Matsuoka, K., Kim, H.T., McDonough, S.M., Bindra, B., Alyea, E.P., Armand, P., Cutler, C., Ho, V.T., Treister, N.S., et al. (2011). Interleukin-2 and Regulatory T Cells in Graft-versus-Host Disease. N. Engl. J. Med. *365*, 2055–2066.

Kullback, S., and Leibler, R.A. (1951). On Information and Sufficiency. Ann. Math. Stat. *22*, 79–86.

Kwok, S., Mack, D.H., Mullis, K.B., Poiesz, B., Ehrlich, G., Blair, D., Friedman-Kien, A., and Sninsky, J.J. (1987). Identification of human immunodeficiency virus sequences by using in vitro enzymatic amplification and oligomer cleavage detection. J. Virol. *61*, 1690–1694.

Landry, C.R., Lemos, B., Rifkin, S.A., Dickinson, W.J., and Hartl, D.L. (2007). Genetic properties influencing the evolvability of gene expression. Science *317*, 118–121.

Lannoy, N., and Hermans, C. (2016). Principles of genetic variations and molecular diseases: applications in hemophilia A. Crit. Rev. Oncol. Hematol. *104*, 1–8.

Lee, S.-I., and Batzoglou, S. (2003). Application of independent component analysis to microarrays. Genome Biol. *4*, R76.

Lehner, B., and Kaneko, K. (2011). Fluctuation and response in biology. Cell. Mol. Life Sci. CMLS *68*, 1005–1010.

Levy, B.D., Clish, C.B., Schmidt, B., Gronert, K., and Serhan, C.N. (2001). Lipid mediator class switching during acute inflammation: signals in resolution. Nat. Immunol. *2*, 612–619.

Liaw, A., and Wiener, M. (2002). Classification and Regression by randomForest. R News *2*, 18–22.

Lin, B.R., Frausto, R.F., Vo, R.C., Chiu, S.Y., Chen, J.L., and Aldave, A.J. (2016). Identification of the First De Novo UBIAD1 Gene Mutation Associated with Schnyder Corneal Dystrophy. J. Ophthalmol. *2016*, 1968493.

Liu, S., Sun, K., Jiang, T., and Feng, J. (2015). Natural epigenetic variation in bats and its role in evolution. J. Exp. Biol. *218*, 100–106.

Manikkam, M., Haque, M.M., Guerrero-Bosagna, C., Nilsson, E.E., and Skinner, M.K. (2014). Pesticide methoxychlor promotes the epigenetic transgenerational inheritance of adult-onset disease through the female germline. PloS One *9*, e102091.

Mar, J.C., Matigian, N.A., Mackay-Sim, A., Mellick, G.D., Sue, C.M., Silburn, P.A., McGrath, J.J., Quackenbush, J., and Wells, C.A. (2011). Variance of gene expression identifies altered network constraints in neurological disease. PLoS Genet. *7*, e1002207.

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

Marcon, E., and Hérault, B. (2015). Entropart : An R Package to Measure and Partition Diversity. J. Stat. Softw. *67*.

Martínez, O., and Reyes-Valdés, M.H. (2008). Defining diversity, specialization, and gene specificity in transcriptomes through information theory. Proc. Natl. Acad. Sci. U. S. A. *105*, 9709–9714.

Marzec, M., Halasa, K., Kasprzycka, M., Wysocka, M., Liu, X., Tobias, J.W., Baldwin, D., Zhang, Q., Odum, N., Rook, A.H., et al. (2008). Differential effects of interleukin-2 and interleukin-15 versus interleukin-21 on CD4+ cutaneous T-cell lymphoma cells. Cancer Res. *68*, 1083–1091.

Masel, J. (2011). Genetic drift. Curr. Biol. CB *21*, R837–838.

McCarthy, D.J., and Smyth, G.K. (2009). Testing significance relative to a fold-change threshold is a TREAT. Bioinformatics *25*, 765–771.

McGonagle, D., and McDermott, M.F. (2006). A proposed classification of the immunological diseases. PLoS Med. *3*, e297.

Mishalian, I., Granot, Z., and Fridlender, Z.G. (2016). The diversity of circulating neutrophils in cancer. Immunobiology.

Miyanari, Y., and Torres-Padilla, M.-E. (2012). Control of ground-state pluripotency by allelic regulation of Nanog. Nature *483*, 470–473.

Moffett-King, A. (2002). Natural killer cells and pregnancy. Nat. Rev. Immunol. *2*, 656–663.

Moine, M. (2002). Indicateurs de diversité et exploitation statistique d'une question ouverte. 6ème Journ. Int. Anal. Stat. Données Textuelles.

Monod, J. (1971). Le Hasard et la nécessité: essai sur la philosophie naturelle de la biologie moderne (Paris: Ed. du Seuil).

Morris, M.C., Gilliam, E.A., and Li, L. (2014). Innate immune programing by endotoxin and its pathological consequences. Front. Immunol. *5*, 680.

Müller, C., Schillert, A., Röthemeier, C., Trégouët, D.-A., Proust, C., Binder, H., Pfeiffer, N., Beutel, M., Lackner, K.J., Schnabel, R.B., et al. (2016). Removing Batch Effects from Longitudinal Gene Expression - Quantile Normalization Plus ComBat as Best Approach for Microarray Transcriptome Data. PloS One *11*, e0156594.

Mullis, K.B., and Faloona, F.A. (1987). Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. Methods Enzymol. *155*, 335–350.

Munn, D.H., and Mellor, A.L. (2016). IDO in the Tumor Microenvironment: Inflammation, Counter-Regulation, and Tolerance. Trends Immunol. *37*, 193–207.

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.

Newman, J.R.S., Ghaemmaghami, S., Ihmels, J., Breslow, D.K., Noble, M., DeRisi, J.L., and Weissman, J.S. (2006). Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. Nature *441*, 840–846.

Nguyen, A., Yoshida, M., Goodarzi, H., and Tavazoie, S.F. (2016). Highly variable cancer subpopulations that exhibit enhanced transcriptome variability and metastatic fitness. Nat. Commun. *7*, 11246.

Novick, A., and Weiner, M. (1957). Enzyme induction as an all-or-none phenomenon. Proc. Natl. Acad. Sci. U. S. A. *43*, 553–566.

Opstelten, J.L., Plassais, J., van Mil, S.W.C., Achouri, E., Pichaud, M., Siersema, P.D., Oldenburg, B., and Cervino, A.C.L. (2016). Gut Microbial Diversity Is Reduced in Smokers with Crohn's Disease. Inflamm. Bowel Dis. *22*, 2070–2077.

Otsuka, A., Nonomura, Y., and Kabashima, K. (2016). Roles of basophils and mast cells in cutaneous inflammation. Semin. Immunopathol. *38*, 563–570.

Ozbudak, E.M., Thattai, M., Kurtser, I., Grossman, A.D., and van Oudenaarden, A. (2002). Regulation of noise in the expression of a single gene. Nat. Genet. *31*, 69–73.

Passlick, B., Flieger, D., and Ziegler-Heitbrock, H.W. (1989). Identification and characterization of a novel monocyte subpopulation in human peripheral blood. Blood *74*, 2527–2534.

Paszek, P., Ryan, S., Ashall, L., Sillitoe, K., Harper, C.V., Spiller, D.G., Rand, D.A., and White, M.R.H. (2010). Population robustness arising from cellular heterogeneity. Proc. Natl. Acad. Sci. U. S. A. *107*, 11644–11649.

Patel, D.J., and Wang, Z. (2013). Readout of epigenetic modifications. Annu. Rev. Biochem. *82*, 81–118.

Paulsson, J., and Ehrenberg, M. (2000). Random signal fluctuations can reduce random fluctuations in regulated components of chemical regulatory networks. Phys. Rev. Lett. *84*, 5447–5450.

Petrik, J. (2001). Microarray technology: the future of blood testing? Vox Sang. *80*, 1–11.

Pham, H.P. (2013). L'importance de la variabilité inter-individuelle dans l'étude de la dynamique et de la diversité du répertoire des lymphocytes T au cours du vieillessement.

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.

Phipson, B., Lee, S., Majewski, I.J., Alexander, W.S., and Smyth, G.K. (2016). Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. Ann. Appl. Stat. *10*, 946–963.

Pittendrigh, C.S. (1958). Adaptation, Natural Selection, and Behavior. In Behavior and Evolution, A. Roe, and G.G. Simpson, eds. (New Haven: Yale University Press),.

Piunti, A., and Shilatifard, A. (2016). Epigenetic balance of gene expression by Polycomb and COMPASS families. Science *352*, aad9780.

Plenge, R.M. (2016). Disciplined approach to drug discovery and early development. Sci. Transl. Med. *8*, 349ps15.

Raser, J.M., and O'Shea, E.K. (2004). Control of stochasticity in eukaryotic gene expression. Science *304*, 1811–1814.

Rosenzwajg, M., Churlaud, G., Mallone, R., Six, A., Dérian, N., Chaara, W., Lorenzon, R., Long, S.A., Buckner, J.H., Afonso, G., et al. (2015). Low-dose interleukin-2 fosters a dose-dependent regulatory T cell tuned milieu in T1D patients. J. Autoimmun. *58*, 48–58.

Saadoun, D., Rosenzwajg, M., Joly, F., Six, A., Carrat, F., Thibault, V., Sene, D., Cacoub, P., and Klatzmann, D. (2011). Regulatory T-cell responses to low-dose interleukin-2 in HCV-induced vasculitis. N. Engl. J. Med. *365*, 2067–2077.

Sadlack, B., Merz, H., Schorle, H., Schimpl, A., Feller, A.C., and Horak, I. (1993). Ulcerative colitis-like disease in mice with a disrupted interleukin-2 gene. Cell *75*, 253–261.

Sadlack, B., Kühn, R., Schorle, H., Rajewsky, K., Müller, W., and Horak, I. (1994). Development and proliferation of lymphocytes in mice deficient for both interleukins-2 and -4. Eur. J. Immunol. *24*, 281–284.

Sadlack, B., Löhler, J., Schorle, H., Klebb, G., Haber, H., Sickel, E., Noelle, R.J., and Horak, I. (1995). Generalized autoimmune disease in interleukin-2-deficient mice is triggered by an uncontrolled activation and proliferation of CD4+ T cells. Eur. J. Immunol. *25*, 3053–3059.

Sakaguchi, S., Toda, M., Asano, M., Itoh, M., Morse, S.S., and Sakaguchi, N. (1996). T cell-mediated maintenance of natural self-tolerance: its breakdown as a possible cause of various autoimmune diseases. J. Autoimmun. *9*, 211–220.

Sakaguchi, S., Sakaguchi, N., Shimizu, J., Yamazaki, S., Sakihama, T., Itoh, M., Kuniyasu, Y., Nomura, T., Toda, M., and Takahashi, T. (2001). Immunologic tolerance maintained by CD25+ CD4+ regulatory T cells: their common role in controlling autoimmunity, tumor immunity, and transplantation tolerance. Immunol. Rev. *182*, 18–32.

Sallusto, F. (2016). Heterogeneity of Human CD4(+) T Cells Against Microbes. Annu. Rev. Immunol. *34*, 317–334.

Sasaki, Y., Sakai, M., Miyazaki, S., Higuma, S., Shiozaki, A., and Saito, S. (2004). Decidual and peripheral blood CD4+CD25+ regulatory T cells in early pregnancy subjects and spontaneous abortion cases. Mol. Hum. Reprod. *10*, 347–353.

Schmid, R., Baum, P., Ittrich, C., Fundel-Clemens, K., Huber, W., Brors, B., Eils, R., Weith, A., Mennerich, D., and Quast, K. (2010). Comparison of normalization methods for Illumina BeadChip HumanHT-12 v3. BMC Genomics *11*, 349.

Schurmann, C., Heim, K., Schillert, A., Blankenberg, S., Carstensen, M., Dörr, M., Endlich, K., Felix, S.B., Gieger, C., Grallert, H., et al. (2012). Analyzing illumina gene expression microarray

data from different tissues: methodological aspects of data analysis in the metaxpress consortium. PloS One *7*, e50938.

Six, A., Bellier, B., Thomas-Vaslin, V., and Klatzmann, D. (2012). Systems biology in vaccine design. Microb. Biotechnol. *5*, 295–304.

Sobesky, R., Feray, C., Rimlinger, F., Derian, N., Dos Santos, A., Roque-Afonso, A.-M., Samuel, D., Bréchot, C., and Thiers, V. (2007). Distinct hepatitis C virus core and F protein quasispecies in tumoral and nontumoral hepatocytes isolated via microdissection. Hepatol. Baltim. Md *46*, 1704–1712.

Soumelis, V., Pattarini, L., Michea, P., and Cappuccio, A. (2015). Systems approaches to unravel innate immune cell diversity, environmental plasticity and functional specialization. Curr. Opin. Immunol. *32*, 42–47.

De Sousa E Melo, F., Vermeulen, L., Fessler, E., and Medema, J.P. (2013). Cancer heterogeneity--a multifaceted view. EMBO Rep. *14*, 686–695.

Stone, K.D., Prussin, C., and Metcalfe, D.D. (2010). IgE, mast cells, basophils, and eosinophils. J. Allergy Clin. Immunol. *125*, S73–80.

Strickland, S., and Richards, W.G. (1992). Invasion of the trophoblasts. Cell *71*, 355–357.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. *102*, 15545 –15550.

Sugimoto, M.A., Sousa, L.P., Pinho, V., Perretti, M., and Teixeira, M.M. (2016). Resolution of Inflammation: What Controls Its Onset? Front. Immunol. *7*, 160.

Terrier, B., Derian, N., Schoindre, Y., Chaara, W., Geri, G., Zahr, N., Mariampillai, K., Rosenzwajg, M., Carpentier, W., Musset, L., et al. (2012). Restoration of regulatory and effector T cell balance and B cell homeostasis in systemic lupus erythematosus patients through vitamin D supplementation. Arthritis Res. Ther. *14*, R221.

Thakkar, N., Salerno, S., Hornik, C.P., and Gonzalez, D. (2016). Clinical Pharmacology Studies in Critically Ill Children. Pharm. Res.

Thomas, S., Rouilly, V., Patin, E., Alanio, C., Dubois, A., Delval, C., Marquier, L.-G., Fauchoux, N., Sayegrih, S., Vray, M., et al. (2015). The Milieu Intérieur study - an integrative approach for study of human immunological variance. Clin. Immunol. Orlando Fla *157*, 277–293.

Timm, S., Frydenberg, M., Janson, C., Campbell, B., Forsberg, B., Gislason, T., Holm, M., Jogi, R., Omenaas, E., Sigsgaard, T., et al. (2016). The Urban-Rural Gradient In Asthma: A Population-Based Study in Northern Europe. Int. J. Environ. Res. Public. Health *13*.

Torres-Padilla, M.-E., and Chambers, I. (2014). Transcription factor heterogeneity in pluripotent stem cells: a stochastic advantage. Dev. Camb. Engl. *141*, 2173–2181.

Touzot, M., Dahirel, A., Cappuccio, A., Segura, E., Hupé, P., and Soumelis, V. (2015). Using Transcriptional Signatures to Assess Immune Cell Function: From Basic Mechanisms to Immune-Related Disease. J. Mol. Biol. *427*, 3356–3367.

Veneziano, D., Di Bella, S., Nigita, G., Laganà, A., Ferro, A., and Croce, C.M. (2016). Noncoding RNA: Current Deep Sequencing Data Analysis Approaches and Challenges. Hum. Mutat.

Verbeeck, R.K., Günther, G., Kibuule, D., Hunter, C., and Rennie, T.W. (2016). Optimizing treatment outcome of first-line anti-tuberculosis drugs: the role of therapeutic drug monitoring. Eur. J. Clin. Pharmacol. *72*, 905–916.

Warren, L., Bryder, D., Weissman, I.L., and Quake, S.R. (2006). Transcription factor profiling in individual hematopoietic progenitors by digital RT-PCR. Proc. Natl. Acad. Sci. U. S. A. *103*, 17807–17812.

Wildsmith, S.E., and Elcock, F.J. (2001). Microarrays under the microscope. Mol. Pathol. MP *54*, 8–16.

Williams, T.D. (2008). Individual variation in endocrine systems: moving beyond the "tyranny of the Golden Mean."Philos. Trans. R. Soc. Lond. B. Biol. Sci. *363*, 1687–1698.

Wu, J., and Tzanakakis, E.S. (2013). Distinct allelic patterns of nanog expression impart embryonic stem cell population heterogeneity. PLoS Comput. Biol. *9*, e1003140.

Zak, D.E., Andersen-Nissen, E., Peterson, E.R., Sato, A., Hamilton, M.K., Borgerding, J., Krishnamurty, A.T., Chang, J.T., Adams, D.J., and Hensley, T.R. (2012). Merck Ad5/HIV induces broad innate immune activation that predicts CD8+ T-cell responses but is attenuated by preexisting Ad5 immunity. Proc. Natl. Acad. Sci. *109*, E3503–E3512.

ANNEXE #1

PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Early Transcriptome Signatures from Immunized Mouse Dendritic Cells Predict Late Vaccine-Induced T-Cell Responses

Nicolas Dérian[1,2,3], Bertrand Bellier[1,2,3], Hang Phuong Pham[1,3¤a], Eliza Tsitoura[4¤b], Dorothea Kazazi[4], Christophe Huret[1,3¤c], Penelope Mavromara[4], David Klatzmann[1,2,3]*, Adrien Six[1,2,3]*

1 Sorbonne Universités, UPMC Univ Paris 06, UMRS 959, Immunology, Immunopathology, Immunotherapy, Paris, France, **2** AP-HP, Clinical Investigation Center in Biotherapy, Hôpital Pitié-Salpêtrière, Paris, France, **3** INSERM, UMRS 959, "Immunology, Immunopathology, Immunotherapy", Paris, France, **4** Molecular Virology Laboratory, Hellenic Pasteur Institute, Athens, Greece

¤a Current address: ILTOO Pharma, iPEPS—ICM Hôpital Pitié Salpêtrière, 47/83 Boulevard de l'Hôpital, Paris France
¤b Current address: Laboratory of Molecular and Cellular Pneumonology, Medical School, University of Crete, Heraklion, Greece
¤c Current address: CNRS UMR7216 Epigenetics and Cell Fate, Université Paris Diderot, Sorbonne Paris Cité, Paris, France
* david.klatzmann@upmc.fr (DK); adrien.six@upmc.fr (AS)

OPEN ACCESS

Citation: Dérian N, Bellier B, Pham HP, Tsitoura E, Kazazi D, Huret C, et al. (2016) Early Transcriptome Signatures from Immunized Mouse Dendritic Cells Predict Late Vaccine-Induced T-Cell Responses. PLoS Comput Biol 12(3): e1004801. doi:10.1371/journal.pcbi.1004801

Editor: Grégoire Altan-Bonnet, Memorial Sloan-Kettering Cancer Center, UNITED STATES

Received: May 22, 2015

Accepted: February 8, 2016

Published: March 21, 2016

Data Availability Statement: Microarray data are available on GEO: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=ijejgwimldunhax&acc=GSE66991. Data are publicly available since July 1st, 2015.

Funding: This work was sponsored by European Commission (http://ec.europa.eu) under contract No. LSHB-CT-2004-005246 "CompuVac: rational design and standardized evaluation of novel genetic vaccines" and Safer and Faster Evidence-based Translation (T SAFE 115003), and LabEx Transimmunom (ANR-11-IDEX-0004-02, http://www.

Abstract

Systems biology offers promising approaches for identifying response-specific signatures to vaccination and assessing their predictive value. Here, we designed a modelling strategy aiming to predict the quality of late T-cell responses after vaccination from early transcriptome analysis of dendritic cells. Using standardized staining with tetramer, we first quantified antigen-specific T-cell expansion 5 to 10 days after vaccination with one of a set of 41 different vaccine vectors all expressing the same antigen. Hierarchical clustering of the responses defined sets of high and low T cell response inducers. We then compared these responses with the transcriptome of splenic dendritic cells obtained 6 hours after vaccination with the same vectors and produced a *random forest* model capable of predicting the quality of the later antigen-specific T-cell expansion. The model also successfully predicted vector classification as low or strong T-cell response inducers of a novel set of vaccine vectors, based on the early transcriptome results obtained from spleen dendritic cells, whole spleen and even peripheral blood mononuclear cells. Finally, our model developed with mouse datasets also accurately predicted vaccine efficacy from literature-mined human datasets.

Author Summary

Vaccines are designed to elicit effective immune responses against antigens. The various vector platforms used in vaccine development are diverse and complex, rendering the selection of promising vaccines vector challenging. We have designed a modeling strategy

PLOS | COMPUTATIONAL BIOLOGY

agence-nationale-recherche.fr). CH, ET and DKa were supported by the CompuVac consortium (www.compuvac.eu). PHP was supported by a doctoral fellowship from the Ministère de la Recherche et de la Technologie (http://www.enseignementsup-recherche.gouv.fr). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

that predicts the propensity of vaccine vectors to elicit strong late T-cell responses using transcriptome material obtained 6 hours after vaccination. Our model, designed with mouse datasets, also predicted vector efficacy from mined human data. Thus, molecular signatures obtained 6 hours after vaccination can predict vaccine efficacy at 2 weeks post vaccination, which should help in vaccine development.

Introduction

The development of vaccines against complex chronic diseases such as HIV or cancer has been largely unsuccessful so far. Novel vaccine technologies are rationally designed to generate appropriate protective immune responses [1], notably efficient T-cell responses. Such vaccine vectors include plasmid DNA, viral and bacterial vectors, and virus-like particles (VLPs). The intrinsic characteristics of these vectors, including their capacity to stimulate innate immunity and to activate and target the antigen to antigen-presenting cells, determine in large part their immunogenicity and thus their potency as vaccine or gene therapy vectors [2–4]. However the rational design of vectors is limited by various aspects, such as the partial understanding of the factors governing the induction of optimal immunity (i.e. the activation of the innate immune system by various vector components, the effect upon adaptive immunity. . .) or the possible dependence of vector efficacy on the specificity of the target diseases.

Systems biology has been introduced in vaccine development to assist in circumventing these limitations and shorten the vaccine development process. Systems biology may not only help to better understand, analyze and reconstruct the complex immune interactions between the pathogen/vaccine and host immune system, but may also improve the *in silico* testing models for vaccine candidates. Systems biology approaches have proven capable to predict immune responses induced after vaccination [5,6]. For example, expression patterns of genes associated with the efficient processing of peptides for major histocompatibility complex presentation have been identified as useful surrogate markers of vaccine efficacy, obviating the need to perform challenge studies [7]. Signatures derived from antibody repertoire profiling on peptide microarrays during the natural course of influenza infection were shown to be predictive of the efficacy of influenza vaccines [8]. Multivariate analysis performed on human peripheral blood mononuclear cell (PBMC) microarray data, obtained 3 days after vaccination, identified innate immune response–related signatures that predicted the late adaptive immune response to the YF-17D yellow fever vaccine [9].

In this manuscript, we describe a methodology that enabled us to successfully predict the adaptive immune responses induced by large sets of vaccine vectors of different classes, ranging from infectious particles to VLPs and DNA. All these vectors expressed the same antigen, the immune response to which was measured using a validated standardized method. We developed our model based on the analysis of transcriptomic data, obtained 6 hours after vaccination, that could predict the antigen-specific immune responses induced at the peak of the response, 5–10 days later. It is noteworthy that this model, developed in mice, successfully predicted vaccine-induced responses from literature-mined human datasets.

Results

Vaccine vector classification according to antigen-specific T-cell expansion

Forty-one vectors classified in 13 categories of vaccines and all expressing the same antigen were evaluated and compared for their ability to induce an adaptive T-cell immune response after vaccination (S1 Table). The forty-one vectors included (i) recombinant viral vectors

derived from adenovirus (rAd), vaccinia (VACC), modified vaccinia Ankara (MVA) and lentivirus (LV), (ii) recombinant bacteria vectors derived from Bacille de Calmette et Guérin (BCG), (iii) recombinant VLPs made of the AP205 [10] or Qbeta (Qb) [11] proteins from bacteriophage, the VP2 proteins from murine polyoma virus (MPY) [12] or murine pneumotropic virus (MPT), the Gag capsid proteins from murine leukaemia virus (MLV) [13], the core from hepatitis B virus (HBc), and (iv) plasmid encoding a recombinant protein (DNA) or recombinant MLV-VLPs (plasmoVLPs) [13,14]. Each vaccine platform was engineered to display or express the immunodominant LCMV gp33-41 epitope model antigen [15] in order to compare the different vaccine-induced CD8+ T-cell specific responses. In the framework of CompuVac (www.compuvac.eu), we standardized the method for measuring the gp33-41-specific T-cell response using tetramer staining (Fig 1A). Mice were immunized with each vector and we evaluated the gp33-41-specific T-cell response in PBMCs at days 5, 7 and 10, following the frequency of circulating gp33-41/H-2Db tetramer+ CD8+ T cells. In each experiment we included control mice that were injected with PBS or rAd (rAd_1 batch) to provide negative and positive controls. Data for each experimental group were normalized as the experimental to rAd vector response ratio allowing cross-laboratory data comparisons.

We observed a wide range of immune responses that were triggered by the different vectors. The maximal CD8+ T-cell expansion was induced with bacteriophage-derived VLPs, while very low but significant responses were observed with MPT and HBc VLPs (Fig 1B). Interestingly, different vector designs within the same vector platform led to different responses. As an example, Qb-derived VLPs induced variable CD8+ T-cell expansion depending on their production processes that were designed to modify their TLR-ligand composition (i.e. Qb_5 devoid of viral RNA and CpG in contrast to Qb_1; Fig 1A). We took into consideration all the vectors and performed hierarchical clustering on normalized values that defined 3 clusters (C). The first cluster comprised vectors with low ratio values, characterizing weak inducers of antigen-specific T cells, hereafter referred as "Weak" vectors. The other 2 clusters included vectors inducing high or intermediate responses, defining the "Strong" vector class. This class comprised the different recombinant viral vectors (rAd, MVA, VACC, LV) expressing rather than displaying the antigen, and which have been extensively developed as CD8+ T-cell vaccines [16–18]. It also contained bacteriophage-adjuvanted VLPs, in agreement with previous reports [10,19].

Modelling strategy

As dendritic cell activation is key to the initiation of immune responses, we investigated whether transcriptome data from sorted spleen dendritic cells (DCs) sampled 6 hours after immunization could be predictive of the antigen-specific T-cell response measured several days later, at the peak of the response. To address this question, we devised a stepwise modelling scheme. DC-sorted transcriptome datasets were initially produced for 19 vectors on the Codelink platform, corresponding to 7 different vaccine platforms, for which the antigen-specific T-cell response was also measured (S1 Table).

The rationale for looking at signatures instead of individual genes was motivated by (i) the need to detect slight gene expression modifications (captured as the overall expression changes of correlated genes), (ii) the technical constraints of working on different microarray platforms (CodeLink, Illumina and Affymetrix), and (iii) the objective of producing a predictive model working across microarray platforms. Thus, our modelling scheme was based on our recently described strategy for signature discovery, using independent component analysis (ICA) followed by gene set enrichment analysis (GSEA) [20]. This allows circumventing the limitations due to the use of different platforms when analyzing individual gene expressions, by comparing statistical signature's enrichment across datasets.

PLOS | COMPUTATIONAL BIOLOGY

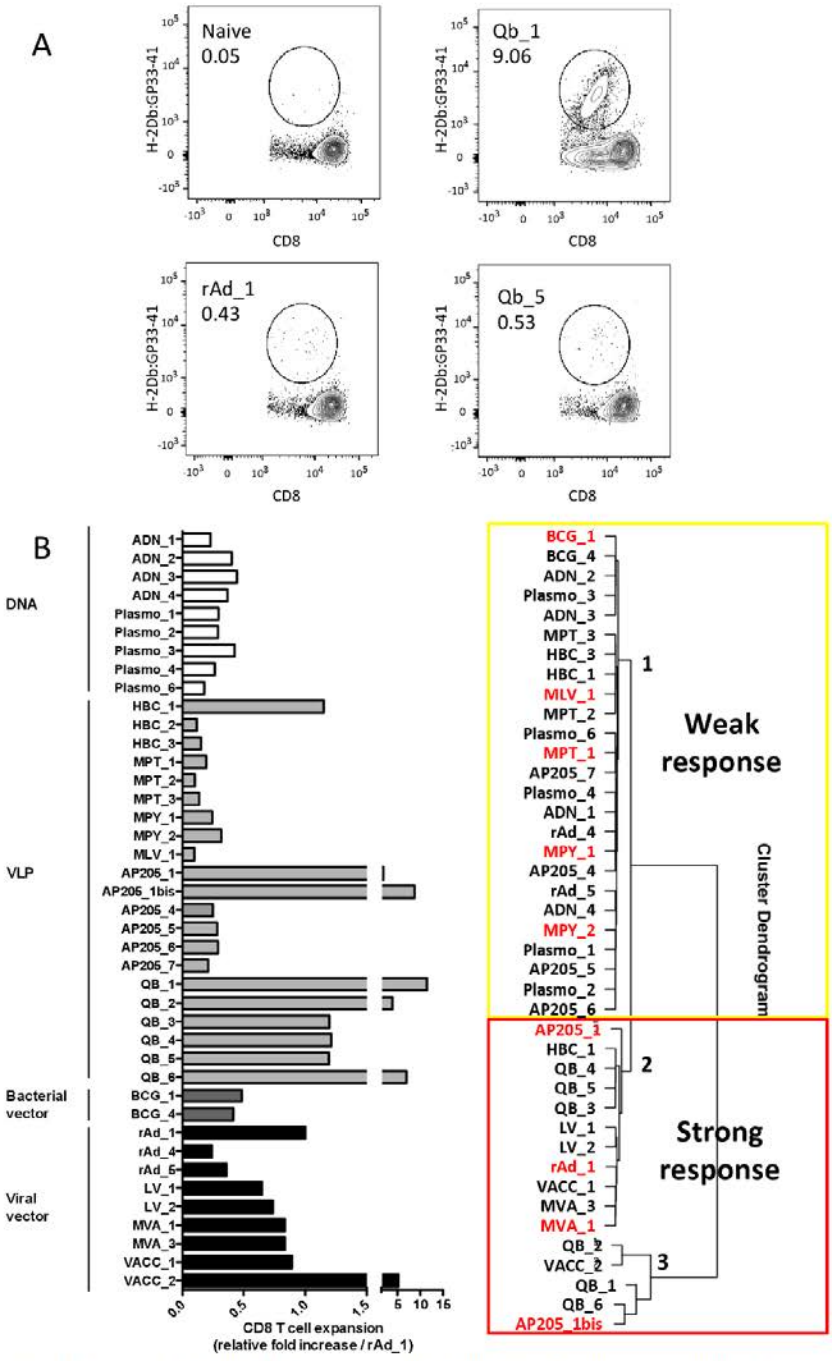


Fig 1. LCMV gp33-41 model antigen-expressing/displaying vector T-cell response analysis. A. Evaluation of gp33-41 specific T-cell frequency in mice immunized with Qb_1 or Qb_5 VLPs, with rAd_1 and control naive mice, by H-2Db:gp33-41 tetramer staining. B. For each vector tested, gp33-41 antigen-specific

responses were evaluated at days 5, 7 and 10 in groups of 3–5 vaccinated mice. The normalized "CD8 T-cell expansion" value was calculated as the average of the peak response for each mouse against the value obtained for the internal standard experimental group (rAd_1). C. Hierarchical clustering (Euclidean/Ward. D2) performed on normalized T-cell response values defined as "Weak" (cluster 1) and "Strong" (clusters 2 and 3) vectors. Vectors in red were used to build the initial prediction model (see Model stability and confidence in the Results section).

doi:10.1371/journal.pcbi.1004801.g001

ICA is an unsupervised algorithm extracting independent components Y from original datasets X by searching for the demixing matrix W:

$$Y = X \times W$$

W matrix is calculated by maximizing the non gaussianity of the components measured as the negentropy J:

$$J(y) = H(y_{Gauss}) - H(y),$$

where $H(y)$ and $H(y_{Gauss})$ are the Shannon entropy for a vector y and a random Gaussian vector with same variance as y [21].

The use of ICA to analyze microarray data is justified by the hypothesis that X is a mix of signals from underlying cellular pathways. Therefore, columns of Y contain a summary of gene contributions in the extracted components. The RNA expression value of a gene is thus the superposition of several signals of this gene in each component which add up. From each component y, two reduced gene sets can be extracted by selecting genes with critical contribution on both sides of the distribution [22].

We first performed ICA on the 19 available datasets, yielding 210 molecular signatures characterizing the variability within each dataset, and likely linked to vector properties. We then analyzed the differential gene expression between the controls and the tested vectors using bootstrapping [23,24], in order to increase the model's sensitivity. Bootstrapping consists in sampling series of additional datasets by randomly drawing samples with replacement of equal size from an original dataset, as described in Fig 2. We sampled 100 consecutive bootstrapped datasets from each of the 19 original datasets and generated 100 corresponding ranking lists of genes based on modified t-test statistics. The previously identified signatures were then tested for their behavior vis-à-vis the gene lists using GSEA, generating normalized enrichment scores (NES). Molecular signatures from GSEA software (>5000) were added at this step in order to increase the efficiency of the normalization procedure. NES of molecular signatures from ICA were then extracted for the next steps. This yielded a matrix, containing 1900 columns (100 bootstrapped datasets for each of the 19 original datasets) and 210 lines (the number of calculated NES). This matrix was then used to create random forest (RF) classification models (Fig 2). NES values and T-cell response classification were used as predictors and dependent variables, respectively, in the randomForest package, which as output provides classification results and associated probabilities for each T-cell response class.

Model stability and confidence

An initial predictive model was built with 9 vector datasets (in red in Fig 1B) for which the antigen-specific T-cell responses were available (900 bootstrapped datasets and 100 signatures). Predictions of 10 additional datasets, including independent experiments done with the same or different batches of these vectors, were very consistent (see Tables 1 and S2). The model sensitivity for the "Weak" and "Strong" vector classes (respectively equal to the specificity for the "Strong" and "Weak" classes) are 0.89 and 0.98, respectively. The positive predictive value

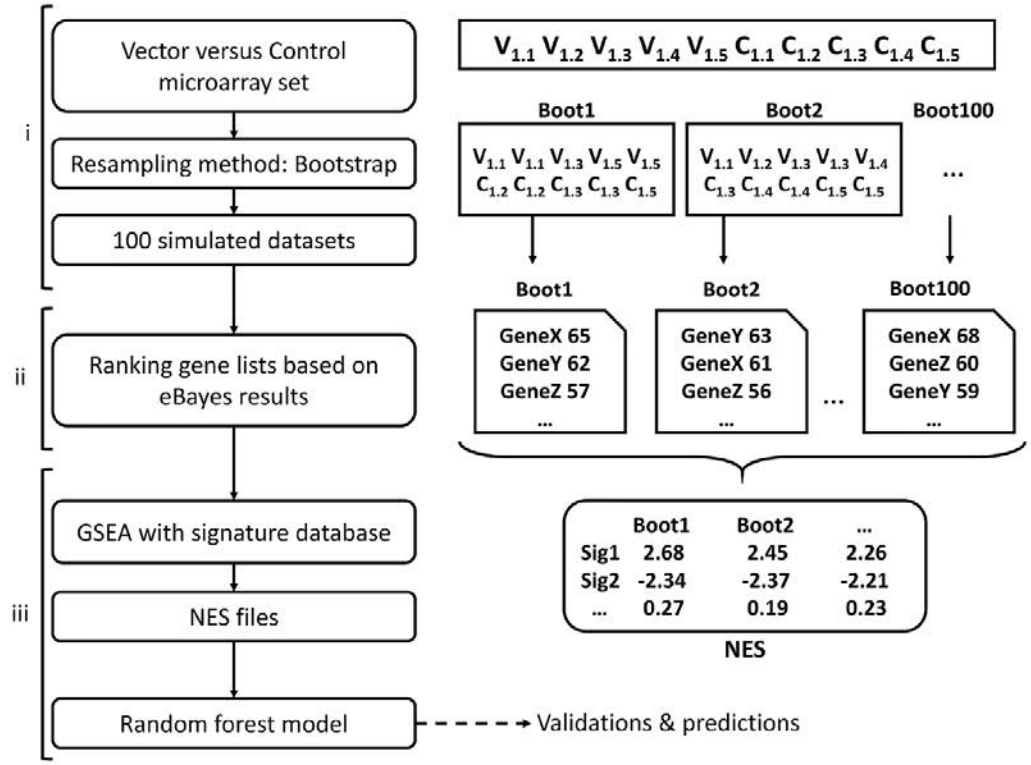


Fig 2. Modelling strategy. (i) For each pre-processed dataset, composed of microarray measures for mice injected with vector 1 (V1.1, V1.2 . . .) and control (C1.1, C1.2 . . .), one hundred datasets were created by bootstrapping samples among V and C. (ii) Ranked gene lists, according to the eBayes statistical comparison of vector and control conditions, were generated. (iii) Potential signatures were tested for enrichment on each of the 100 ranked gene lists by GSEA. The resulting NES matrix was then used to build the random forest model.

doi:10.1371/journal.pcbi.1004801.g002

(PPV) is stable for the two classes ("Weak": 0.96, "Strong": 0.93). This 9-vector model is already efficient to classify the vector platform with 0.94 accuracy. These results led us to construct the final predictive model (called RFM model) including all the 19 datasets, based on the analysis of the 210 signatures across the 1900 bootstrapped datasets. This complete training set contained enough information to discriminate clearly between the 2 vector classes, as demonstrated by the misclassification rate parameter reaching zero after 100 simulated trees.

The RandomForest algorithm provides a ranked list of the signatures based on their importance to the efficacy of the classification in the model. This score is based on the decrease of the Gini impurity criterion for each child node of a split. The result of this calculus is the mean of this decrease for each signature present in the trees of the forest. 27 most important signatures, having a mean decrease score higher than ten, were selected. Clustering methods were then

Table 1. Model's sensitivity and accuracy.

Model	Sensitivity Strong (Specificity Weak)	Sensitivity Weak (Specificity Strong)	PPV Strong (NPV Weak)	PPV Weak (NPV Strong)	Accuracy
9-vector	0.98	0.89	0.93	0.96	0.94
RFM	**1**	**0.97**	**0.96**	**1**	**0.98**

doi:10.1371/journal.pcbi.1004801.t001

PLOS | COMPUTATIONAL BIOLOGY

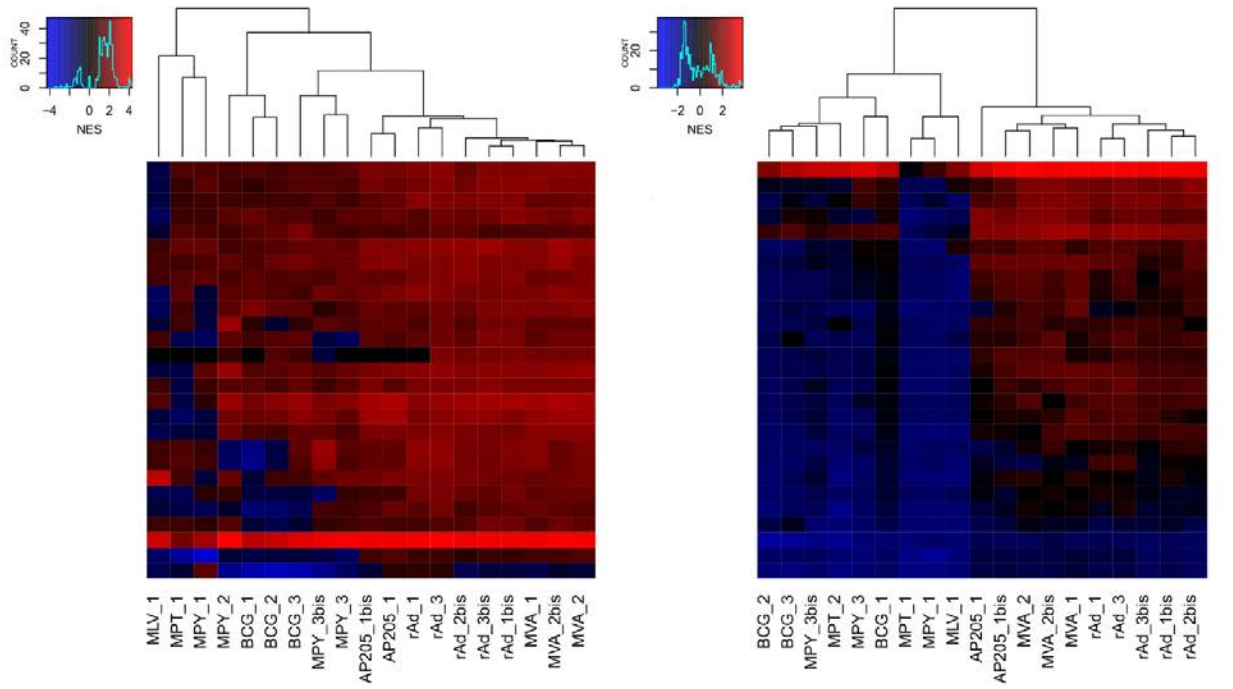


Fig 3. Hierarchical clustering (distance method: Euclidean; agglomeration method: Complete) of NES values of the 27 selected signatures, provided by the RFM model, on original vector datasets (A) and of mean NES values calculated on bootstrapped datasets (B).

doi:10.1371/journal.pcbi.1004801.g003

applied (i) on NES values of these 27 signatures calculated on original datasets (Fig 3A) and (ii) on the mean NES values calculated on the bootstrapped datasets (Fig 3B). The interest of bootstrap is clearly revealed with clusters more explicitly defined after bootstrap.

We then asked whether RFM was biased toward particular vector datasets. We first used the leave-one-out methodology, where 19 models were iteratively built using only 18 out of 19 datasets, and then assessing how accurately such models predict the 100 bootstraps from the left-out dataset. All vectors were classified as expected for at least 96 of the 100 bootstrapped datasets, except MPY_3 for which 16 bootstrapped datasets were misclassified (S3 Table). This result shows overall very high prediction stability and no significant bias of the RFM model.

We verified that RFM was not biased for a given vector platform. One hundred new models were constructed, each based on one randomly selected representative of the 7 vector platforms (rAd, AP205, MVA, MPY, MPT, MLV and BCG). For each vector, the probabilities to be classified as expected were calculated and the prediction distribution across the 100 models is shown in Fig 4. Vaccines from the "Strong" vector class (in red) showed good consistency in their prediction distribution, with no value under 0.6 (100% confidence). Vaccines from the "Weak" vector class showed more variability: in particular, 2 MPY vaccines (MPY_3 & MPY_3bis; same vector batch (#3) used in 2 independent experiments) were not classified as expected in 16 models out of 100 (84% confidence); these 16 misclassifying models all used MPY_2 as the MPY representative. Note that this specific preparation (#2) of MPY vaccine was produced using baculovirus machinery in insect-derived cells, while the other MPYs were produced in yeast.

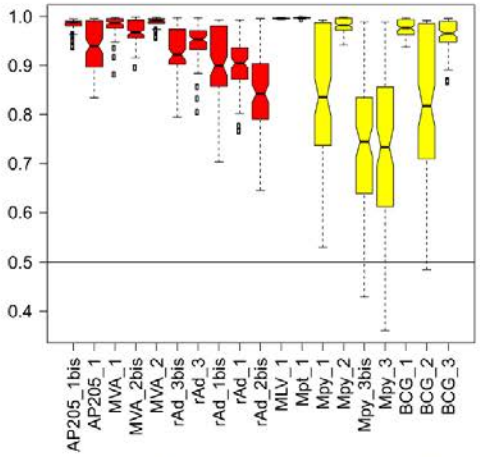


Fig 4. Vector prediction confidence. One hundred different models were created using one representative of each vector platform. The probabilities for a vector being classified as expected were calculated for its 100 bootstrapped datasets and averaged as a vector mean probability. Vector mean probabilities are displayed as boxplots. A value of 1 means that the bootstrapped dataset was successfully predicted 100 times over the 100 models. "Strong" and "Weak" vectors are colored red and yellow, respectively.

doi:10.1371/journal.pcbi.1004801.g004

Model validation with novel vectors

RFM was then used to predict the vector class of 4 new vectors belonging to 3 vector platforms: 2 batches of lentivirus (LV) vectors -a category of vaccine not represented during the model establishment, one new batch of AP205 (AP205_3) and one of MLV (MLV_2). We had independently determined that LV vectors induced strong antigen-specific T-cell responses after immunization and were classified in the "Strong" vector class (Fig 1). As shown in Tables 1 and 2A, these 4 bootstrapped datasets were classified as expected with high precision (>95%) while sensitivity and PPV of the model increased compared to the 9-vector model, especially the sensitivity for the "Weak" vector class now reaching 0.97 (from 0.89) with RFM. These results highlight that RFM (i) is not vaccine platform-dependent, (ii) correctly predicts a vector platform unknown to the model, and (iii) efficiently predicts both "Weak" and "Strong" vectors.

Model prediction of whole spleen and PBMC data

RFM was built on transcriptome data obtained from sorted spleen DCs. In our next experiment, we assessed whether RFM would be sensitive enough to classify transcriptome datasets derived from whole spleen samples obtained 6 hours after immunization, where DCs represent 1–2% of total splenocytes. As summarized in Table 2B, all bootstrapped datasets from whole spleens were well classified, with at least 91% of the expected classification, thus demonstrating our model's sensitivity in classifying vectors in whole spleen transcriptome datasets.

We then tested microarray datasets for whole spleen samples obtained 6, 48 and 72 hours after vaccination with one vector, the rAd vector that we used as a standard. Strikingly, only datasets sampled 6 hours after injection were classified as expected (as "Strong") (Table 2D).

Similarly, we tested the performance of our model in classifying vectors using PBMC-derived microarray datasets. The rationale for this experiment is that PBMCs, less than 1% of which are DCs, offer a more accessible sample source than spleen, especially in humans. As

Table 2. Vector class prediction efficiency.

			RFM model predictions*	
	Vectors	Material	Strong	Weak
	LV_1	DC	**100**	0
	LV_2	DC	**100**	0
A	AP205_3	DC	3	**97**
	MLV_2	DC	5	**95**
	AP205_1	Spleen	**99**	1
	MVA_1	Spleen	**100**	0
B	rAd_1	Spleen	**98**	2
	MLV_1	Spleen	9	**91**
	MPT_1	Spleen	0	**100**
	AP205_1	PBMC	**73**	27
	MVA_1	PBMC	**90**	10
	Qb_1	PBMC	**99**	1
	Qb_2	PBMC	**95**	5
C	Qb_3	PBMC	**100**	0
	Qb_4	PBMC	**100**	0
	Qb_5	PBMC	**98**	2
	MLV_1	PBMC	0	**100**
	MPT_1	PBMC	0	**100**
	rAd_1_6	Spleen	**98**	2
D	rAd_1_48	Spleen	0	**100**
	rAd_1_72	Spleen	2	**98**

*Number of the 100 bootstrapped datasets predicted as "Strong" or "Weak".

doi:10.1371/journal.pcbi.1004801.t002

shown in Table 2C, all but one vectors were classified as expected with high precision (≥ 90%). AP205_1 was classified as expected, though with less confidence (73%).

Model prediction of human PBMC data

Finally, we tested whether our model could classify datasets obtained from the literature. We found datasets from the Merck Ad5/HIV trial reported by Zak et al. [25] PBMC transcriptome data were generated from samples obtained at 6, 24 and 72 hours after vaccination. We bootstrapped the samples of Zak et al., taking patient-paired samples before and after vaccination. 100% and 91% of the bootstrapped paired samples were predicted as "Strong" at 24 and 72 hours, respectively (Table 3), in line with the authors' original observations. The same analysis

Table 3. Predictions of human PBMC transcriptome data derived 6, 24 and 72 hours after vaccination by MRKAd5/HIV published in [25].

		RFM model predictions*	
Time point	Material	Strong	Weak
6 h	PBMCs	31	**69**
24 h	PBMCs	**100**	**0**
72 h	PBMCs	**91**	**9**

*Number of the 100 bootstrapped datasets predicted as "Strong" or "Weak".

doi:10.1371/journal.pcbi.1004801.t003

performed with the 6-hour time point gave a "Strong" prediction for 31% of the bootstrapped paired samples. The latter finding is consistent with the conclusion of Zak et al. that transcriptomic modifications at 6 hours were not significant. These results demonstrate the capacity of RFM generated from mouse DC transcriptome datasets to classify human PBMC datasets.

Biological insight

Biological annotation of the 27 most important signatures of RFM reveals one signature (Sig1) with statistical functional enrichments related to immune processes (FDR p-values 10^{-4}–10^{-8}). This signature is highly focused on STAT-1 with 51 genes having strong biological connections (Fig 5A). Interestingly, Sig1 is upregulated in all the vectors, but with higher intensity in the "Strong" as compared to the "Weak" vectors.

No specific molecular pathway was clearly identified by QIAGEN's Ingenuity Pathway Analysis (IPA) functional analysis for the other 26 important signatures in our model. However, visual inspection of these signatures identified the CH25H gene as highly modulated by strong vectors. Since this gene has been recently described as playing a role in DC maturation [26], we analyzed its network of connected genes with IPA (Fig 5B). This network was also globally more modulated by "Strong" rather than "Weak" vectors, and comprised genes implicated in DC function such as MYD88, DUSP5 and ABCG1.

Discussion

Understanding and predicting innate immune response to vector platforms is primordial for fast and effective production of new vaccination or gene therapy protocols. Systems biology tools efficiently extract information from large datasets in computing predictive models and have already played a major role in recent discoveries in this field [5,27]. In this paper, we initially focused on early transcriptomic changes of DCs since these are first-line players in the innate immune response and directly contribute to the triggering of the adaptive response. Our aim was to identify transcriptomic signatures predictive of the late CD8+ CTL responses to the LCMV gp33-41 model antigen conveyed by a variety of vaccine vectors.

Based on molecular signatures extracted using the non-supervised ICA method [20,22], we produced and validated a prediction model taking into account 19 available datasets generated with different vector platforms. We chose the random forest learning algorithm for its reported efficiency among classification methodologies [28–30]. The originality of our strategy was the use of signatures rather than genes to classify samples. Our results showed that this model consistently predicts both "Weak" and "Strong" vectors, with greater confidence for the latter. This suggests that there are shared gene expression modifications induced by "Strong" vectors, while changes induced by "Weak" vectors are more diverse. Consistent with this, Li et al. recently reported that different types of vaccine lead to different transcriptomic modifications in humans 3 days after vaccination [31], with vaccines inducing high transcriptomic modifications being those that induce robust antibody responses.

Among the 27 signatures selected for their importance in the RFM model, one (Sig1; see S4 Table) is related to immune components, including "viral infection", "role of RIG1-like receptors in antiviral innate immunity" and "interferon signalling" pathways. Previous studies have characterized gene expression modifications in the early stages of vaccination consistent with Sig1 annotation. Querec et al. investigated the transcriptome of patient PBMCs at days 0, 1, 3, 7 and 10 after vaccination with yellow fever vaccine [9]. Of 65 regulated genes, 26 were related in part to interferon and the antiviral response, including MX1, IFIT1, IFIT2, IFIT3, OAS1, OAS2, OAS3 and OASL, and 7 were related to signal transduction, including STAT1 and IRF7. Similarly, Zak et al. [25] applied the modular transcriptome analysis framework described in

PLOS | COMPUTATIONAL BIOLOGY

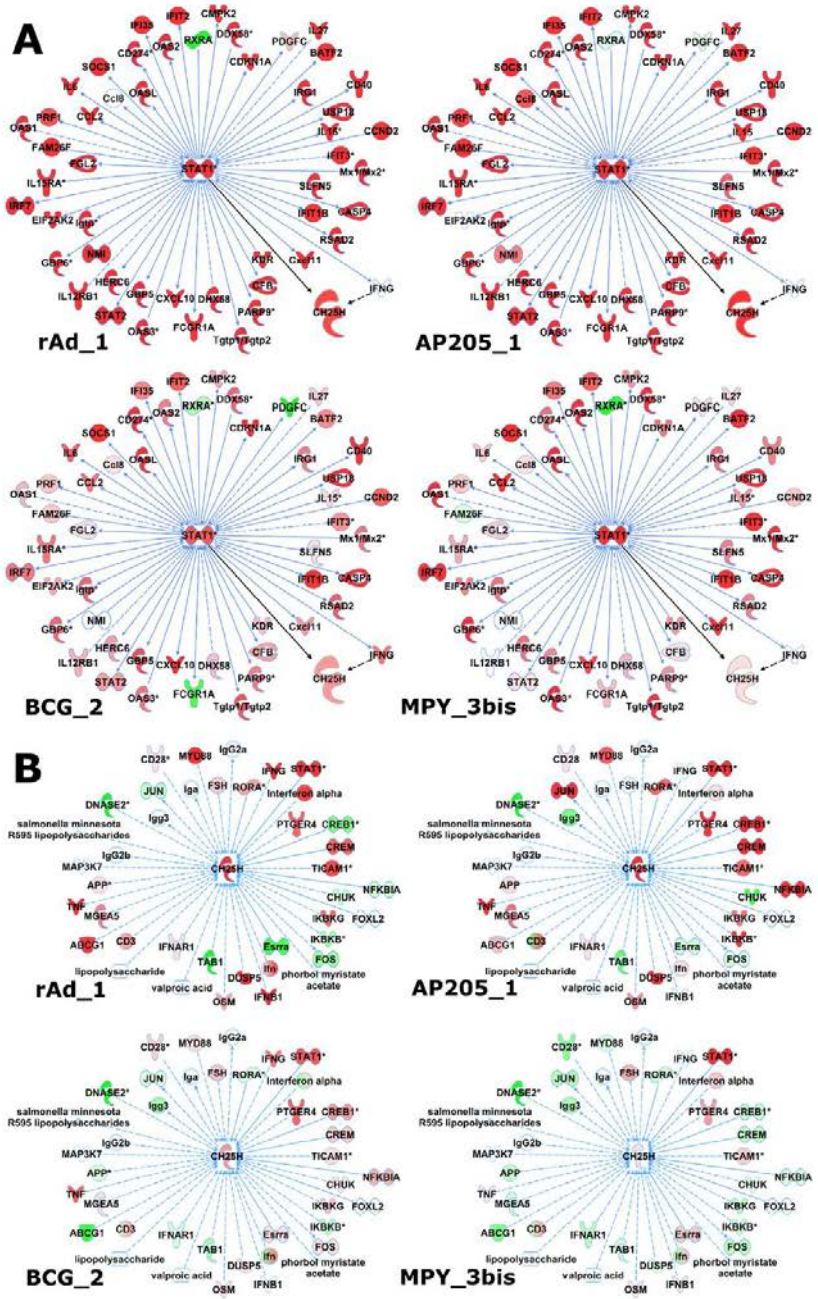


Fig 5. Gene network analysis in "Weak" and "Strong" vectors. A: Selection of STAT-1 related genes derived from Sig1 of the RFM model were targeted on Ingenuity Pathway Analysis (IPA). B: CH25H, a gene selected in one of the other 26 important signatures of the model, was targeted as the key gene on IPA. The grow functionality was used to display all known direct and indirect interactions with CH25H, except miRNA. The biological interactions of CH25H are displayed on A (black arrows). Colors depend on statistical

analyses (red: upregulated, green: downregulated) performed on rAd_1, AP205_1, MPY_3bis and BCG_2 vector datasets; color intensities were set to be in the same range in all experiments.

doi:10.1371/journal.pcbi.1004801.g005

Chaussabel et al. [32] to study the innate immune response to MRKAd5/HIV in PBMCs 6, 24, 72 and 168 hours after patient vaccination. They identified genes highly regulated at 6 and 24 hours, including STAT1, STAT2, IFITs, MXs and OASs (also identified in Querec et al.). Strikingly, all these genes are also part of Sig1, emphasizing further their key role in the early response to the vaccine. Furthermore, DDX60, a newly described antiviral factor that induces Rig-1-like receptor-mediated signaling [33], present in Sig1, was reported by Querec et al. as well [9]. Interestingly, Sig1 is upregulated in vaccinated samples compared to control group, but to a lesser extent in "Weak" vs. "Strong" vectors (see Figs 3 and 5).

Our cross-analysis of Zak et al.'s microarray data on Merck Ad5/HIV-vaccinated human PBMC samples, which yield good predictions for the 24- and 72-hour time points, demonstrates that our prediction model, solely based on mouse DC-sorted transcriptome data, efficiently predicts human transcriptome data. This can be explained by the high similarity of gene expression in immunological cell lineages between mice and humans [34], although the kinetics of the immune response to vaccine is different.

No specific molecular pathway was clearly identified by IPA annotations for the other 26 important signatures in our model. This is somewhat surprising since these signatures have been selected by the model to best distinguish "Strong" and "Weak" vectors and are therefore expected to represent differentially regulated biological pathways. In this line, none of the 27 signatures corresponds to a peculiar behavior of a vector but they rather reveal similar behavior within "Strong" or "Weak" groups (Fig 3). Moreover, the identified signatures were extracted from 13 out of 19 different vector datasets (9 "Strong" and 4 "Weak" vectors). We believe that these signatures are unlikely artifactual but related to yet undefined biological processes. Indeed, the constant improvement of annotation databases can reveal secondary or additional functions of genes. For example, CH25H, a gene found in one of the 26 signatures and clearly upregulated in "Strong" vectors, is primarily involved in cholesterol metabolism, but has recently been shown to play a role in the early stage of DC maturation [26]. Fig 5 shows how the expression of this gene is related to dendritic cell through direct or indirect interactions with STAT-1 or IFNγ, both members of Sig1, and with several genes known to be important in early dendritic cell activation: for example, MYD88 is a gene involved in toll-like receptor signaling [35], DUSP5 is known to be upregulated during dendritic cell maturation [36], and ABCG1 is a gene playing a role in adaptive immune responses [37]. The comparative analysis of gene expression modulation of this interaction network shown in Fig 5 reveals a similar pattern of differential expression for "Strong" vectors (rAd_1, AP205_1) different than that observed for "Weak" vector datasets (BCG_2, MPY_3bis). This again points at a significant difference in early dendritic cell activation-related gene behavior in "Strong" vs. "Weak" vectors.

Altogether, our results underline the relevance of the CompuVac initiative that consisted in producing, in a standardized manner, immunological and transcriptome data related to vaccine candidates in order to predict their capacity to elicit strong antigen-specific responses. Our model was based on transcriptome data from sorted spleen DCs of mice vaccinated with various "Strong" and "Weak" T-cell inducer vectors. This prediction model accurately predicted the behavior of these and other candidate vaccines only 6 hours after injection. The model was powerful enough to produce a relevant vector classification even when using whole mouse spleen and PBMCs, or even human PBMCs (Fig 6), and across 3 microarray platforms (CodeLink, Illumina and Affymetrix). The accuracy and sensitivity of the model are likely high because it is built with very different vaccine platforms therefore representative of possible

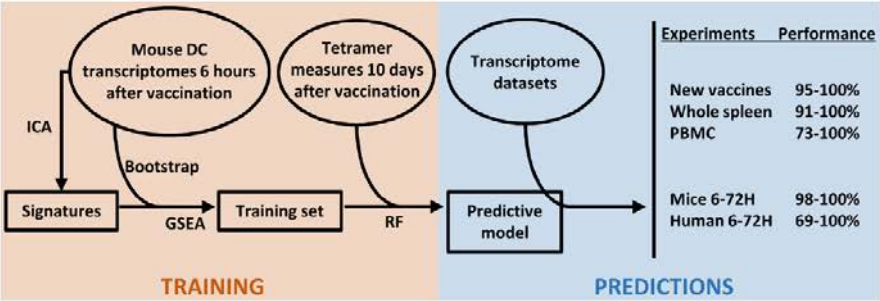


Fig 6. Strategy followed in this study for a high-performance predictive model of transcriptome datasets obtained from new vector platforms, cells from whole organs (spleen, blood). RF: random forest.

doi:10.1371/journal.pcbi.1004801.g006

vector behaviors in triggering the early immune response. This study further supports the potential of systems immunology approaches in facilitating the development and characterization of vaccines, offering robust *in silico* solutions to study the early events of the immune response to vaccines.

Material and Methods

Ethics statement

Experimental protocols complied with French law (Décret: 2001–464 29/05/01) and EEC regulations (86/609/CEE) for the care and use of laboratory animals and were carried out under Authorization for Experimentation on Laboratory Animals Number 75-673-R. Our animal protocol (Ce5/2009/042) was approved by the "Charles Darwin" Ethics Committee for Animal Experimentation (CNREEA 05) and performed in the licensed animal facility A75-13-08.

Vector platforms

Recombinant adenovirus- and MVA-derived viral vectors, BCG-derived bacterial vector, AP205 [10] or Qb [11] bacteriophage-, MPT- and MPY- [12] or MLV-derived [13] VLPs used as an antigenic platform and DNA vaccines were included in this study. According to the CompuVac evaluation scheme, each vaccine platform was engineered to display / express the LCMV gp33-41 model antigen [15] in order to measure the vaccine-induced T-cell specific responses and dendritic cell transcriptome changes (see following sections). The sequence IITSIKAVYNFATCGILAL corresponding to the GP33-41 epitope flanked upstream and downstream by 5 of its natively neighboring amino acids was used. The 53 vectors considered in this paper (S1 Table) are displayed in 13 vector platforms 7 of which were used for a training set (rAd, MVA, AP205, MPT, MPY, MLV and BCG) and 2 for prediction of new platforms (LV and Qb).

Evaluation of antigen-specific T-cell responses

Groups of three to five 7-week-old female C57BL/6 mice (Charles River, France and Germany) were immunized with a controlled quantity of vector particles as defined in CompuVac assay protocols (www.compuvac.eu). For monitoring T-cell responses, each vector was injected with its "best" route of administration: subcutaneously for VLP vectors; intramuscularly for recombinant antigen-expressing vectors and by intra-dermally by gene gun for DNA vaccines. Control mice were injected with 100 μL of phosphate buffered saline solution (PBS). For each

vector (n = 41), the T-cell immune response measurement was performed independently one to three times. T-cell immune responses induced against the LCMV gp33-41 model antigen were measured by MHC-I gp33-41/H-2Db tetramer (ProImmune, UK) staining of PBMCs at 5, 7 and 10 days after injection. The highest measure was kept for each mouse and the mean value was then calculated for the group. Values were normalized against measures monitored in parallel in mice immunized with the rAd_1 control vector.

Microarray data

Experimental groups comprised of 3 to 6 mice immunized with vaccine candidates by the intravenous route. Mice were sacrificed 6 hours after immunization. Spleen DCs were purified with CD11c+-conjugated MACS magnetic beads (Miltenyi Biotec) according to the manufacturer's instructions. After incubation for 20 minutes at 4°C, cells were washed and passed over a MACS column. Purity was checked routinely by FACS and found to be greater than 96±2%. $2x10^6$ CD11c+ cells were used for total RNA extraction using Nucleospin RNAII (Macherey Nagel). For test dataset generation, whole PBMCs and/or whole splenocytes and/or sorted spleen DCs were collected at 6 hours, and at 48- and 72-hour time points for the kinetic follow-up. RNA was checked for quality using gel electrophoresis and for quantity using a Nanodrop spectrophotometer (Thermo Scientific). Microarrays were performed using either Applied Microarrays (CodeLink Mouse Whole Genome Bioarray) or Illumina (WG6 Mouse BeadArray) technologies (S1 Table). The MessageAmp II aRNA Amplification Kit (Ambion) was used for cDNA and cRNA production from 1 μg of total RNA. 10 μg of amplified cRNA was subsequently fragmented and hybridized for 20 hours using the Applied Microarrays hybridization and washing buffer kit. Slides were scanned using the GenePix Personal 4100A scanner for CodeLink array or the Illumina BeadArray 500GX Reader for Illumina array. Hybridization and raw data extraction were performed using either GenePix Pro 6.0 (for CodeLink array) or BeadStudio (for Illumina array) software, respectively (GEO accession GSE66991).

Each tested vector dataset comprised "vector-immunized" and corresponding PBS control samples. Quantile normalization was performed with the limma package [38] on R software [39], and then a log2 transformation was applied. Probes with a detection p-value above 0.05 in all samples in a dataset were discarded.

Signature database and enrichment analysis

Following our two-step ICA→GSEA signature discovery strategy [20], signatures were extracted using the fastICA algorithm R package [40] following modifications in [22]. Parameters were set as default, except for the unmixing matrix A^{-1} convergence threshold set to 10^{-6}. Ranked gene lists were calculated using the limma modified t-test. ES were calculated using GSEA [41] with the pre-ranked gene list protocol. Normalized ES are then calculated based on the permutation performed on gene sets collection, allowing comparison between experiments. The ICA-extracted signature database was complemented with the MsigDB C2 (curated gene sets of biological pathways) and C5 (Gene Ontology gene sets) databases (www.broad.mit.edu/gsea) in order to increase universe of genes available for permutation of gene sets. Signatures with fewer than 7 detected genes were ignored.

Random forest classification and validation

For each model produced in the Results section, classification was performed on a matrix of fastICA extracted signature NES values (see above section) calculated for bootstrapped vector datasets (100 bootstraps per vector dataset), using the random forest algorithm implemented in the randomForest R package to produce a forest of 2000 trees [42]. The number of randomly

selected signatures used at each of the 2000 runs was set according to the *mtry* function implemented in the randomForest package. The class prediction of the new dataset was deduced by the probability to be "Weak" or "Strong" > 0.5. The overall vector class was then obtained as the majority of "Weak" or "Strong" class assignments over the 100 bootstraps.

For classification model validation, we implemented the leave-one-out methodology consisting in creating models with n-1 datasets, where n is the total number of datasets, and classifying the dataset left out. In addition, we implemented a "multi-model" methodology based on the classification of bootstrapped datasets over 100 models created as above. Each model was computed on an NES matrix of a random selection of one representative vector dataset of each of the 7 represented vector platforms (see Vector platforms section and S1 Table). Vector mean probabilities were calculated as the average probability of being "Weak" or "Strong" over the 100 bootstrapped vector datasets, and their distribution over the 100 models was analyzed.

Signature annotation

For biological insight evaluation of the signatures, microarray data were analyzed through the use of QIAGEN's Ingenuity Pathway Analysis (IPA, QIAGEN Redwood City, www.qiagen.com/ingenuity).

Supporting Information

S1 Table. List of studied vectors.
(PDF)

S2 Table. Classification of vectors based on the 9-vector model.
(PDF)

S3 Table. Predictions of dendritic cell transcriptome data derived from mice vaccinated with different vectors in the leave-one-out validation step.
(PDF)

S4 Table. List of genes in Sig1.
(PDF)

Acknowledgments

We thank Bruno Gouritin and Fabien Pitoiset for sample preparation, Wassila Carpentier for microarray datasets generated on the P3S platform (Groupe Hospitalier Pitié-Salpétrière, Paris), and Wahiba Chaara and Ioannis Drakos for proofreading. The authors wish to thank all the CompuVac consortium members (www.compuvac.eu).

Author Contributions

Conceived and designed the experiments: ND BB PHP DKl AS. Performed the experiments: ND AS. Analyzed the data: ND BB DKl AS. Contributed reagents/materials/analysis tools: CH BB PM ET DKa. Wrote the paper: ND BB DKl AS.

References

1. Nabel GJ, Fauci AS. Induction of unnatural immunity: prospects for a broadly protective universal influenza vaccine. Nat Med. 2010 Dec; 16(12):1389–91. doi: 10.1038/nm1210-1389 PMID: 21135852
2. Souza APD, Haut L, Reyes-Sandoval A, Pinto AR. Recombinant viruses as vaccines against viral diseases. Braz J Med Biol Res. 2005 Apr; 38(4):509–22. PMID: 15962176

3. Douek DC, Nabel GJ. Vaccines. Immunol Rev. 2011 Jan; 239(1):5–7. doi: 10.1111/j.1600-065X.2010.00986.x PMID: 21198661

4. Bellier B, Six A, Thomas-Vaslin V, Klatzmann D. Predicting immune responses to viral vectors and transgenes in gene therapy and vaccination: the coming of systems biology. In: The Clinibook: Clinical gene transfert state of art. EDK, groupe EDP Sciences, Paris. Cohen-Haguenauer, O.;

5. Nakaya HI, Pulendran B. Systems vaccinology: its promise and challenge for HIV vaccine development. Curr Opin HIV AIDS. 2012 Jan; 7(1):24–31. doi: 10.1097/COH.0b013e32834dc37b PMID: 22156839

6. Mooney M, McWeeney S, Sékaly R- P. Systems immunogenetics of vaccines. Semin Immunol. 2013 Apr; 25(2):124–9. doi: 10.1016/j.smim.2013.06.003 PMID: 23886894

7. Vahey MT, Wang Z, Kester KE, Cummings J, Heppner DG Jr, Nau ME, et al. Expression of genes associated with immunoproteasome processing of major histocompatibility complex peptides is indicative of protection with adjuvanted RTS,S malaria vaccine. J Infect Dis. 2010 Feb 15; 201(4):580–9. doi: 10.1086/650310 PMID: 20078211

8. Legutki JB, Johnston SA. Immunosignatures can predict vaccine efficacy. Proc Natl Acad Sci USA. 2013 Nov 12; 110(46):18614–9. doi: 10.1073/pnas.1309390110 PMID: 24167296

9. Querec TD, Akondy RS, Lee EK, Cao W, Nakaya HI, Teuwen D, et al. Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans. Nat Immunol. 2009 Jan; 10(1):116–25. doi: 10.1038/ni.1688 PMID: 19029902

10. Tissot AC, Renhofa R, Schmitz N, Cielens I, Meijerink E, Ose V, et al. Versatile virus-like particle carrier for epitope based vaccines. PLoS ONE. 2010; 5(3):e9809. doi: 10.1371/journal.pone.0009809 PMID: 20352110

11. Maurer P, Jennings GT, Willers J, Rohner F, Lindman Y, Roubicek K, et al. A therapeutic vaccine for nicotine dependence: preclinical efficacy, and Phase I safety and immunogenicity. Eur J Immunol. 2005 Jul; 35(7):2031–40. PMID: 15971275

12. Franzén AV, Tegerstedt K, Holländerova D, Forstová J, Ramqvist T, Dalianis T. Murine polyomavirus-VP1 virus-like particles immunize against some polyomavirus-induced tumours. In Vivo. 2005 Apr; 19 (2):323–6. PMID: 15796193

13. Bellier B, Dalba C, Clerc B, Desjardins D, Drury R, Cosset F-L, et al. DNA vaccines encoding retrovirus-based virus-like particles induce efficient immune responses without adjuvant. Vaccine. 2006 Mar 24; 24(14):2643–55. PMID: 16481074

14. Bellier B, Huret C, Miyalou M, Desjardins D, Frenkiel M- P, Despres P, et al. DNA vaccines expressing retrovirus-like particles are efficient immunogens to induce neutralizing antibodies. Vaccine. 2009 Sep 25; 27(42):5772–80. doi: 10.1016/j.vaccine.2009.07.059 PMID: 19656495

15. Gallimore A, Glithero A, Godkin A, Tissot AC, Plückthun A, Elliott T, et al. Induction and exhaustion of lymphocytic choriomeningitis virus-specific cytotoxic T lymphocytes visualized using soluble tetrameric major histocompatibility complex class I-peptide complexes. J Exp Med. 1998 May 4; 187(9):1383–93. PMID: 9565631

16. Shiver JW, Fu T-M, Chen L, Casimiro DR, Davies M-E, Evans RK, et al. Replication-incompetent adenoviral vaccine vector elicits effective anti-immunodeficiency-virus immunity. Nature. 2002 Jan 17; 415 (6869):331–5. PMID: 11797011

17. Tatsis N, Lin S-W, Harris-McCoy K, Garber DA, Feinberg MB, Ertl HCJ. Multiple immunizations with adenovirus and MVA vectors improve CD8+ T cell functionality and mucosal homing. Virology. 2007 Oct 10; 367(1):156–67. PMID: 17590405

18. Di Nunzio F, Félix T, Arhel NJ, Nisole S, Charneau P, Beignon A-S. HIV-derived vectors for therapy and vaccination against HIV. Vaccine. 2012 Mar 28; 30(15):2499–509. doi: 10.1016/j.vaccine.2012.01.089 PMID: 22342915

19. Bessa J, Schmitz N, Hinton HJ, Schwarz K, Jegerlehner A, Bachmann MF. Efficient induction of mucosal and systemic immune responses by virus-like particles administered intranasally: implications for vaccine design. Eur J Immunol. 2008 Jan; 38(1):114–26. PMID: 18081037

20. Pham HP, Dérian N, Chaara W, Bellier B, Klatzmann D, Six A. A novel strategy for molecular signature discovery based on independent component analysis. Int J Data Min Bioinform. 2014; 9(3):277. PMID: 25163169

21. Kong W, Vanderburg C, Gunshin H, Rogers J, Huang X. A review of independent component analysis application to microarray gene expression data. BioTechniques. 2008 Nov; 45(5):501–20. doi: 10.2144/000112950 PMID: 19007336

22. Chiappetta P, Roubaud MC, Torrésani B. Blind source separation and the analysis of microarray data. J Comput Biol. 2004; 11(6):1090–109. PMID: 15662200

23. Efron B. Bootstrap Methods: Another Look at the Jackknife. Ann Stat. 1979 Jan; 7(1):1–26.

24. Singh S, Sedory SA. Sufficient bootstrapping. Comput Stat Data An. 2011 Apr; 55(4):1629–37.

25. Zak DE, Andersen-Nissen E, Peterson ER, Sato A, Hamilton MK, Borgerding J, et al. Merck Ad5/HIV induces broad innate immune activation that predicts CD8+ T-cell responses but is attenuated by pre-existing Ad5 immunity. Proc Natl Acad Sci USA. 2012; 109(50):E3503–12. doi: 10.1073/pnas.1208972109 PMID: 23151505

26. Park K, Scott AL. Cholesterol 25-hydroxylase production by dendritic cells and macrophages is regulated by type I interferons. J Leukocyte Biol. 2010 Aug 10; 88(6):1081–7. doi: 10.1189/jlb.0610318 PMID: 20699362

27. Buonaguro L, Wang E, Tornesello ML, Buonaguro FM, Marincola FM. Systems biology applied to vaccine and immunotherapy development. BMC Syst Biol. 2011; 5(1):146.

28. Díaz-Uriarte R, Alvarez de Andrés S. Gene selection and classification of microarray data using random forest. BMC Bioinformatics. 2006; 7:3. PMID: 16398926

29. Moorthy K, Mohamad MS. Random forest for gene selection and microarray data classification. Bioinformation. 2011; 7(3):142–6. PMID: 22125385

30. Okun O, Priisalu H. Random Forest for Gene Expression Based Cancer Classification: Overlooked Issues. In: Proceedings of the 3rd Iberian conference on Pattern Recognition and Image Analysis, Part II. Girona, Spain: Springer-Verlag; 2007. p. 483–90.

31. Li S, Rouphael N, Duraisingham S, Romero-Steiner S, Presnell S, Davis C, et al. Molecular signatures of antibody responses derived from a systems biology study of five human vaccines. Nat Immunol. 2014 Feb; 15(2):195–204. doi: 10.1038/ni.2789 PMID: 24336226

32. Chaussabel D, Quinn C, Shen J, Patel P, Glaser C, Baldwin N, et al. A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus. Immunity. 2008 Jul 18; 29(1):150–64. doi: 10.1016/j.immuni.2008.05.012 PMID: 18631455

33. Miyashita M, Oshiumi H, Matsumoto M, Seya T. DDX60, a DEXD/H box helicase, is a novel antiviral factor promoting RIG-I-like receptor-mediated signaling. Mol Cell Biol. 2011 Sep; 31(18):3802–19. doi: 10.1128/MCB.01368-10 PMID: 21791617

34. Shay T, Jojic V, Zuk O, Rothamel K, Puyraimond-Zemmour D, Feng T, et al. Conservation and divergence in the transcriptional programs of the human and mouse immune systems. Proc Natl Acad Sci USA. 2013 Feb 19; 110(8):2946–51. doi: 10.1073/pnas.1222738110 PMID: 23382184

35. Hammer GE, Ma A. Molecular control of steady-state dendritic cell maturation and immune homeostasis. Annu Rev Immunol. 2013; 31:743–91. doi: 10.1146/annurev-immunol-020711-074929 PMID: 23330953

36. Türeci Ö, Bian H, Nestle FO, Raddrizzani L, Rosinski JA, Tassis A, et al. Cascades of transcriptional induction during dendritic cell maturation revealed by genome-wide expression analysis. FASEB J. 2003; 17(8):836–47. PMID: 12724343

37. Draper DW, Gowdy KM, Madenspacher JH, Wilson RH, Whitehead GS, Nakano H, et al. ATP Binding Cassette Transporter G1 Deletion Induces IL-17-Dependent Dysregulation of Pulmonary Adaptive Immunity. J Immunol. 2012 Jun 1; 188(11):5327–36. doi: 10.4049/jimmunol.1101605 PMID: 22539789

38. Smyth GK. Limma: linear models for microarray data. In: Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W, editors. Bioinformatics and Computational Biology Solutions using R and Bioconductor. New York: Springer; 2005. p. 397–420.

39. Team RDC. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria; 2011. Available from: http://www.R-project.org

40. Marchini JL, Heaton C, Ripley BD. fastICA: FastICA Algorithms to perform ICA and Projection Pursuit [Internet]. 2012. Available from: http://CRAN.R-project.org/package=fastICA

41. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA. 2005 Oct 25; 102(43):15545–50. PMID: 16199517

42. Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002; 2(3):18–22.

ANNEXE #2

The Journal of Immunology

Regulatory T Cells Orchestrate Similar Immune Evasion of Fetuses and Tumors in Mice

Djamel Nehar-Belaid,[*,†,‡] Tristan Courau,[*,†,‡] Nicolas Dérian,[*,†,‡] Laura Florez,[*,†,‡] Maria Grazia Ruocco,[*,†,‡] and David Klatzmann[*,†,‡,§]

Embryos and tumors are both masses of dividing cells expressing foreign Ags, but they are not rejected by the immune system. We hypothesized that similar tolerogenic mechanisms prevent their rejection. Global comparison of fetal and tumor microenvironments through transcriptomics in mice revealed strikingly similar and dramatic decreases in expression of numerous immune-related pathways, including Ag presentation and T cell signaling. Unsupervised analyses highlighted the parallel kinetics and similarities of immune signature downregulation, from the very first days after tumor or embryo implantation. Besides upregulated signatures related to cell proliferation, the only significant signatures shared by the two conditions across all biological processes and all time points studied were downmodulated immune response signatures. Regulatory T cell depletion completely reverses this immune downmodulation to an immune upregulation that leads to fetal or tumor immune rejection. We propose that evolutionarily selected mechanisms that protect mammalian fetuses from immune attack are hijacked to license tumor development. ***The Journal of Immunology*****, 2016, 196: 000–000.**

There are striking similarities between fetal development and tumor development. They both rely on intense cell division, invasion of host tissues, and sustained vascularization (reviewed in Refs. 1, 2). In the physiological setting of pregnancy, trophoblast cells proliferate intensely, invade the maternal endometrial decidua, and then migrate into the uterine wall where they contribute to the blood supply. This is reminiscent of tumor development and led to the description of the trophoblast as a pseudomalignant tissue (3, 4), or even a physiological metastasis (5).

Moreover, despite the fact that fetus and tumor express foreign Ags—paternal alloantigens for fetuses and altered autoantigens for tumors (6–8)—they are not rejected by the immune system. Numerous immune cells such as NK cells, regulatory T cells (Tregs), effector T cells (Teffs), and dendritic cells (DCs) populate the maternal–fetal interface (9). Some of them, such as uterine NK cells, have a preponderant trophic function (10), whereas others have been shown to play an often redundant role in tolerance of the fetus. However, the interaction between these cells is poorly understood and there is as yet no integrated view of how the immune system is kept under control so as to prevent the elimination of the allogeneic fetus. Similarly, numerous innate and adaptive immune cells appear to participate in the tolerant environment that protects Ag-expressing tumor cells from immune destruction.

Although there are discrete similarities between tolerance to fetuses and to tumors that have intrigued scientists for decades (11), there was until recently no robust evidence of common mechanisms at work in the two settings. The discovery of Tregs, which are key players in immune tolerance, has uncovered their seemingly similar role in the two processes (12). In human pregnancy, Tregs are enriched at the maternal–fetal interface throughout healthy pregnancy (13, 14), whereas a decreased Treg pool is observed in recurrent miscarriage cases (15) and preeclampsia (16). In addition to Tregs, effector T cells may also influence the immune acceptance of the fetus. Some studies have reported the regulation of the effector immune response toward a predominant and favorable Th2 type rather than Th1 type immunity during pregnancy (17, 18). In human cancer, the situation is clearer, with the extent of tumor infiltration being most often associated with poor survival and increased infiltration of activated Teffs being favorable (19, 20).

In mice, tumor emergence as well as embryo implantation elicits a strikingly similar brisk Treg response (21). Tregs specific for self-antigens are recruited in tumor or uterine draining lymph nodes before Teffs are recruited, with the functional relevance of this being supported by the fact that Treg depletion leads to fetal (22, 23) or tumor (24) immune rejection. These observations led us to hypothesize that tumors hijack mechanisms of tolerance initially selected during evolution of the mammalian immune system to protect fetuses (25).

To substantiate this hypothesis, we aimed more broadly to compare immune tolerance mechanisms in the two settings. The immune system comprises large sets of diverse, circulating, interconnected cells, which cooperate through numerous signaling molecules and soluble factors. To capture this complexity and better understand the

*Sorbonne Universités, Université Pierre et Marie Curie, Université Paris 06, Laboratoire I3 (Immunologie–Immunopathologie–Immunothérapie), F-75013 Paris, France; †INSERM, UMR_S 959, F-75013 Paris, France; ‡CNRS, FRE3632, F-75013 Paris, France; and §Assistance-Publique Hopitaux de Paris, Groupe Hospitalier Pitié-Salpêtrière, Département de Biothérapies et Centre d'Investigation Clinique en Biothérapie, F-75013 Paris, France

ORCID: 0000-0001-6002-5021 (D.N.-B.).

Received for publication August 19, 2015. Accepted for publication November 5, 2015.

This work was supported by Institut National du Cancer Grant PLBIO2010 (to D.K.). D.N.-B.'s thesis was partly funded by the CNRS. This work was also supported by Marie Curie Research Fellowship 300901 VISTO European Union Seventh Framework Programme-PEOPLE-2011-International Incoming Fellowship (to M.G.R.).

The datasets presented in this article have been submitted to Gene Expression Omnibus under accession number GSE68454.

Address correspondence and reprint requests to Prof. David Klatzmann, Hôpital Pitié-Salpêtrière, 83 Boulevard de l'Hôpital, F-75013 Paris, France. E-mail address: david.klatzmann@upmc.fr

D.N.-B. performed all the experiments, analyzed the results, and participated in writing the manuscript; T.C., L.F., and M.G.R. participated in some experiments and N.D. in data analysis; and D.K. designed and supervised the study, analyzed the data, and wrote the first manuscript, which was reviewed by all authors.

The online version of this article contains supplemental material.

Abbreviations used in this article: DC, dendritic cell; DEG, differentially expressed gene; E, embryonic day; FDR, false discovery rate; GO, gene ontology; GSEA, gene set enrichment analysis; ICA, independent component analysis; IPA, Ingenuity Pathway Analysis; Teff, effector T cell; Treg, regulatory T cell.

global nature of immune responses at work in the tumor microenvironment and uterine tissues, systems approaches are needed. We thus investigated fetal and tumor microenvironments using whole transcriptome microarray analysis, which is now well codified and standardized and for which there are numerous mature tools for experimental design, data analysis, and graphical visualization (26–29).

We show that Tregs are not only important players, but are the conductors of the immune orchestra that plays a similar score leading to tolerance to fetuses and to tumors.

Materials and Methods

Animals

BALB/c and C57BL/6 female mice, 6–8 wk of age, were from Elevage Janvier (Le Genest St. Isle, France). All mice were treated in accordance with European Union guidelines for animal experimentation.

Tumor experiments

B16F10 melanoma cells were obtained from the American Type Culture Collection. Cells (5×10^5) were injected s.c. into the right flank of C57BL/6 mice. Mice were sacrificed at 4 and 14 d after tumor inoculation. Tumor microenvironments were collected from punch biopsies always of the same size (2.5 mm diameter) centered on the inoculation site, thus collecting both tumor and surrounding tissue. Samples were incubated overnight in RNAlater (Qiagen) at 4°C and then transferred to −80°C for storage.

Maternal–fetal experiments

Embryonic day (E)4 to E12 allopregnant C57BL/6 mice were from Elevage Janvier. The entire uterus (with fetal tissues) was collected and samples were incubated overnight in RNAlater (Qiagen) at 4°C and then transferred to −80°C for storage.

Sample size

Sample sizes were as follows: 1) for maternal–fetal experiments, uteri from nonpregnant mice (control samples), $n = 6$; E4, $n = 4$; E6, $n = 6$; E8, $n = 4$; E10, $n = 4$; E11, $n = 4$; and E12, $n = 5$; 2) for Treg depletion experiments, nonpregnant mice, $n = 5$; E12, $n = 5$; and PC61-treated mice at E12, $n = 4$; and 3) for tumor experiments, normal skin samples (control samples), $n = 7$; tumor-bearing mice at day 4 after tumor inoculation, $n = 4$; and at day 14 after tumor inoculation, $n = 5$.

Gene expression analyses

Samples from uterine tissues and tumor microenvironment were lysed and homogenized using a tissue lyser (Qiagen), and total RNA was purified using TRIzol (Invitrogen) according to the manufacturers' instructions.

RNA yield was assessed using a NanoDrop 1000 spectrophotometer (Thermo Fisher Scientific). RNA integrity was assessed using an Agilent Bioanalyzer showing a quality RNA integrity number of 8–10 (Agilent Technologies). The RNA was processed using the Illumina TotalPrep RNA amplification kit protocol according to the manufacturer's protocol. Briefly, labeled complementary RNAs were hybridized overnight to Illumina MouseWG-6 v2.0 Expression BeadChip arrays. The arrays were then washed, blocked, stained, and scanned on an Illumina BeadStation following the manufacturer's protocols. Illumina BeadStudio software was used to generate signal intensity values from the scans. Genes were filtered out from the analysis when their expression was below the detection limit ($p < 0.05$) in at least two of three samples in both microenvironment and control groups. Next, data were normalized according to the quantile method using the limma R package and then log transformed.

The limma package was used to identify differentially expressed genes (Benjamini–Hochberg corrected $p < 0.05$) at days 4 and 14 (tumor microenvironment) or E4, E6, E8, E10, and E12 (maternal–fetal tissues). The limma package is freely available at: http://www.bioconductor.org/packages/release/bioc/html/limma.html.

Dataset quality assessment

Hierarchical clustering was performed using Euclidean distance and the Ward agglomeration method. Principal variance component analysis was performed using R version 3.1.3, which is freely available at: http://www.bioconductor.org/packages/release/bioc/html/pvca.html.

All datasets were deposited in the Gene Expression Omnibus repository (reference no. GSE68454): http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE68454.

Identification of specific molecular signatures

Specific molecular signatures were generated and statistically tested using the independent component analysis (ICA) → gene set enrichment analysis (GSEA) method developed in Pham et al. (30).

Independent component analysis

The CRAN package fastICA (31) was used to carry out the ICA.

GSEA

Probe sets were formatted into the GMT file, a GSEA input file format. Probes were sorted following the score of the statistical test and used as a preranked list of RNK files. We used the weighted scoring scheme to compute the enrichment score. A false discovery rate (FDR) q value, indicating the probability of a false-positive score, was computed from the. FDR and normalized enrichment scores were calculated using the databases of signatures.

A detailed explanation of GSEA can be found in Subramanian et al. (32).

Generation of functional modules of molecular signatures

Overlaps between significantly enriched signatures from GSEA could be visualized in Cytoscape (33) (version 2.8.3) with the Enrichment Map plugin (34). The Enrichment Map produced networks whose nodes represent signatures and whose edges represent mutual overlap. This approach groups highly redundant signatures together as modules. Only gene sets with an FDR p value of at least <0.05 were selected, and the mutual overlap coefficient between signatures was set at 0.8. The Enrichment Map was used to identify the biological processes discriminating pregnant mice (E4–E12) from nonpregnant mice as controls (and tumor-injected groups from the control group).

Modules of functionally related signatures were manually circled and assigned a label. The functional network was manually curated to remove modules containing fewer than three signatures, resulting in a simplified network map, as shown in Figs. 2, 4, and 5.

Similarity measure

To compare tumor microenvironment and uterine tissue–extracted signatures, we used Jaccard and GOSemSim similarity indexes. The Jaccard index (Ji) is the ratio between the intersection of two sets and its union (35). J(A,B) = [size of (A intersect B)]/[size of (A union B)].

The GOSemSim index is described in Yu et al. (36) and was released under the GNU General Public License in the Bioconductor Project and is freely available at: http://bioconductor.org/packages/2.6/bioc/html/GOSemSim.html.

In vivo depletion of $CD4^+CD25^+$ T cells

Treg in vivo depletion was performed by i.p. injection of 500 μg of an anti-CD25 mAb (PC-61.5.3 from Bio X Cell, West Lebanon, NH) at ∼2.5 d postcoitum. This induces a >80% transient depletion of $CD25^{high}$ cells for ∼3 wk in lymph nodes of normal mice. Uterine environments of rejected fetuses from these mice were then analyzed.

Results

Dynamic downregulation of immune pathways revealed by supervised transcriptome analyses of the pregnant uterus

We first analyzed global changes in the RNA expression profile of uterine tissues induced by pregnancy. We generated a set of transcriptomic data from the uterus of nonpregnant mice and from mice at 4–12 d postfertilization, that is, the E4–E12 stages of pregnancy. We verified the quality of our transcriptome dataset using principal variance component analysis, which estimates within a dataset sources of variability due to biological or technical effects such as RNA concentration, RNA quality, date, position in the array, or experimental groups. We observed that the different experimental groups (i.e., the time for a given condition) accounted for >70% of the total variability, indicating that our dataset is of high quality (Supplemental Fig. 1A). Principal component analysis of normalized gene expression datasets revealed a clear separation between mice from the nonpregnant and early (E6, E8) or late pregnancy (E10, E11, and E12) groups (Supplemental Fig. 1B).

We first performed supervised analyses looking at differentially expressed genes (DEGs) in uterine tissues. At E4, we could already

detect >2000 genes that are highly significantly up- or downregulated (FDR q value < 0.01), as compared with nonpregnant mice (Fig. 1A). This number rapidly increased with time to >6000 regulated genes at E10 and E12. These observations highlight that the pregnancy-induced gene expression changes in uterine tissues are rapid and extremely complex and cannot be reduced to one or a few pathways.

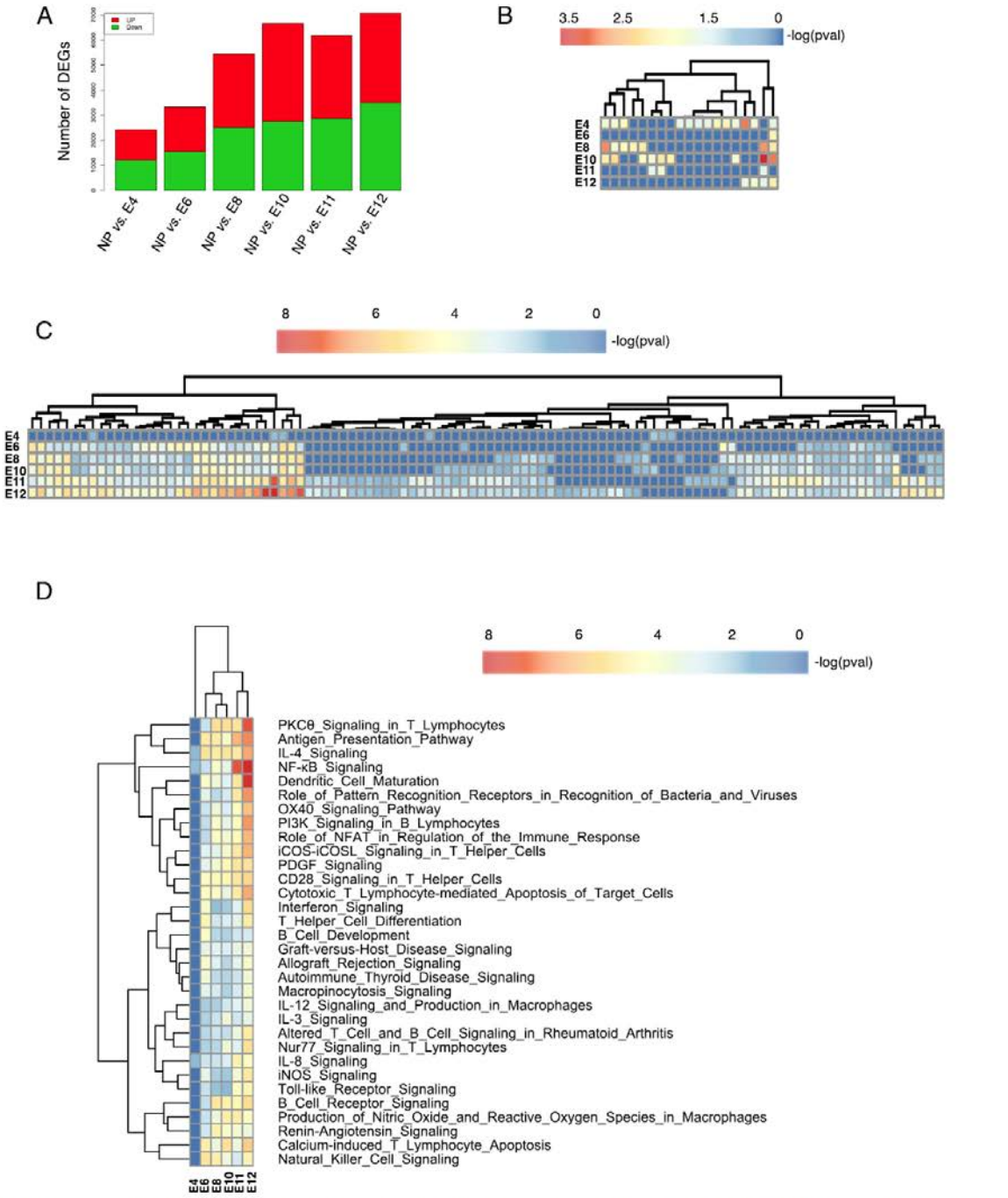


FIGURE 1. Supervised analyses of differentially expressed genes in uterine tissues during pregnancy. (**A**) Bar plot of all DEGs in pregnant mice at E4, E6, E8, E10, E11, and E12 in comparison with nonpregnant (NP) mice. The significance threshold was set at $p < 0.01$ (eBayes test). Shown in red are the upregulated genes and in green are the downregulated ones. (**B**–**D**) DEGs were annotated using IPA and depicted as heat maps according to their significance level [−log(p value)]. The upregulated immune-associated pathways are represented in (**B**) whereas downregulated ones are shown in (**C**) and (**D**). The heat map colors represent the statistical significance [−log(p value)] of each pathway. Only the pathways showing a −log(p value) > 1.3 are considered to be significantly modulated, which corresponds to a pale blue color as shown on the scale. Data are representative of n = 4–6 mice per group.

To study the regulated genes participating in immune responses during pregnancy, we annotated DEGs using Ingenuity Pathway Analysis (IPA) and analyzed them as clusters of regulated genes participating in defined immune pathways. We identified 126 immune-annotated pathways as being significantly modulated at least at one time point. Only 19 (15%) of them were upregulated (Fig. 1B). Moreover, most of these pathways were not consistently upregulated over time, except for the antiproliferative role of TOB in T cell signaling pathway (Fig. 1B, *last column*). TOB being a negative regulator of T cell proliferation (37), its upregulation should in fact lead to decreased T cell responses.

In contrast, 106 pathways (85%) were consistently downregulated over time (Fig. 1C). Downregulated pathways were subdivided into two main subclusters based on statistical significance. The first cluster contains 32 highly significantly regulated pathways ($4 < -\log(p \text{ value}) < 8$) that were all downregulated already at E6 (Fig. 1D). Of note, 25 of these immune pathways (78%) are labeled signaling. They include the IL-4, IL-12, IFN, NF-κB, and NFAT signaling pathways, which are all key to Th cell responses. The seven other downmodulated pathways are related to Ag presentation or T lymphocytes. The second cluster contains 47 pathways that are statistically less significant ($1.5 < -\log(p \text{ value}) < 3.5$) and are modulated later, starting at E10/E12 in most cases (Supplemental Fig. 1C). They include the IL-1, IL-17, CTLA4, CD40, and 4-1BB signaling pathways.

Dynamic downregulation of immune pathways revealed by unsupervised transcriptome analyses of the pregnant uterus

Supervised analyses may bias the interpretation of the results. Therefore, we further analyzed our datasets with unsupervised methodologies. We first used ICA, which blindly unravels the complexity of a dataset by identifying sources of complexity, whose variations of expression are correlated and define molecular signatures. In addition to these signatures generated from the dataset, we also supplemented our analysis with gene sets compiled from Gene Ontology (GO) (38) using GSEA (32). Altogether, the analysis identified 585–817 significantly enriched signatures from E4 to E12, and it focused on the highly discriminating ones, that is, with a Benjamini–Hochberg FDR of <0.05. Using Enrichment Map (34), a Cytoscape software (33) plugin, we represented them as networks of signatures, in green and red for the down and upregulated ones, respectively (Fig. 2A).

As early as E4, large groups of signatures categorized as related to 1) DNA and RNA synthesis, 2) ribosome, proteasome, or endoplasmic reticulum function, and 3) metabolism were upregulated. They likely reflect the intense proliferation of fetal and uterine cells. Of note, the only other signatures regulated at E4 were those related to the immune response, and they were all downregulated. The early trend for cell proliferation upregulation and immune response downregulation is then conserved from E4 to E12, with more numerous immune annotated signatures being downregulated over time (Fig. 2A).

Dynamic upregulation of Treg-related pathways in the pregnant uterus

Reductionist approaches have highlighted the role of Tregs in maternal–fetal tolerance (25). Therefore, we asked what the behavior of Treg signatures could be within our dataset. We found that a Treg signature defined by genes that are upregulated by Foxp3 were upregulated as early as E4 and remained so until E12 (Fig. 2B). In contrast, a mirror signature defining Tregs by genes that are downregulated by Foxp3 (39) was downmodulated (Supplemental Fig. 2A). Furthermore, unbiased principal component analysis, based on other described Treg gene expression signatures (40–42), showed a perfect separation of pregnant and nonpregnant groups as early as E6 (Supplemental Fig. 2B–D). Altogether, these results are in line with the early recruitment and activation of Tregs in the pregnant uterus (25), and they indicate that Foxp3 does act as a transcriptional activator and repressor in vivo.

Ag presentation and T cell activation are the main downmodulated pathways in the pregnant uterus

We then analyzed the main significantly enriched immune pathways represented in our unsupervised signatures using IPA representation (Fig. 3, Supplemental Figs. 3, 4B). The first pathway, the Ag presentation pathway, shows early (E6) downmodulation (green) of MHC molecules and genes involved in proteolysis, transport, and presentation of peptides by MHC molecules (Fig. 3, *upper panel*), which is maintained and even more pronounced at E12. Altogether, this should result in a marked decrease of Ag presentation to T cells. A larger pathway called DC maturation reveals numerous additional downmodulated genes (Supplemental Fig. 4B). These include genes encoding membrane molecules involved in DC maturation 1) by T cells (MHC class I and class II for signal 1, and CD83, CD86, and CD40 for signal 2), 2) by innate lymphocytes (TNFR, CD40, and CD1A), or 3) by microbes and cytokines (TLR2/3/4/9).

To confirm further the downmodulation of Ag presentation–related gene expression and validate the data obtained by gene expression annotation, we tested whether transcriptional changes were reflected in changes of protein expression at the cell level. Therefore, we isolated DCs from nonpregnant or pregnant uteri and analyzed their numbers and phenotype. We observed a decrease in the percentages of DCs among hematopoietic cells in pregnant uteri, which could account for part of the decreased RNA expression. Moreover, we found reduced expression of activation markers such as MHC class II or TLR2 molecules (Supplemental Fig. 4A).

The second pathway, called CD28 signaling in Th cells, shows the decreased expression of MHC and costimulatory molecules on APCs, as discussed above, together with downregulation of multiple T cell activation pathways, including NF-κB and NFAT. These early changes (E6) were also observed at E12.

Reversion from a down- to an upregulated immune microenvironment in the uterus by Treg ablation

To investigate whether changes in the maternal–fetal interface were controlled by Tregs, we mated female mice that were Treg-depleted and analyzed uterine tissues at E12 (Fig. 4). Comparison of unmanipulated pregnant and nonpregnant mice showed again major downregulation of immune response–annotated signatures and upregulation of cell proliferation–related signatures (Fig. 4A). In contrast, in Treg-depleted pregnant mice, the very same immune signatures were markedly upregulated in the uterine tissues of the rejected fetuses, compared not only with control pregnant mice (Fig. 4C) but also with nonpregnant mice (Fig. 4B). This indicates true immune activation and not just the reversion to baseline of the immune signature downmodulation observed in pregnant mice. Taken together, these results confirm that Tregs play an essential role in orchestrating tolerance in uterine tissues during pregnancy.

Remarkably similar dynamic changes in gene expression profiles in the microenvironment of tumors and fetuses

We previously showed that embryo or tumor cell implantation leads to a similar brisk recruitment of self-specific activated/memory Tregs (21, 25). To analyze the similarities between tolerance to

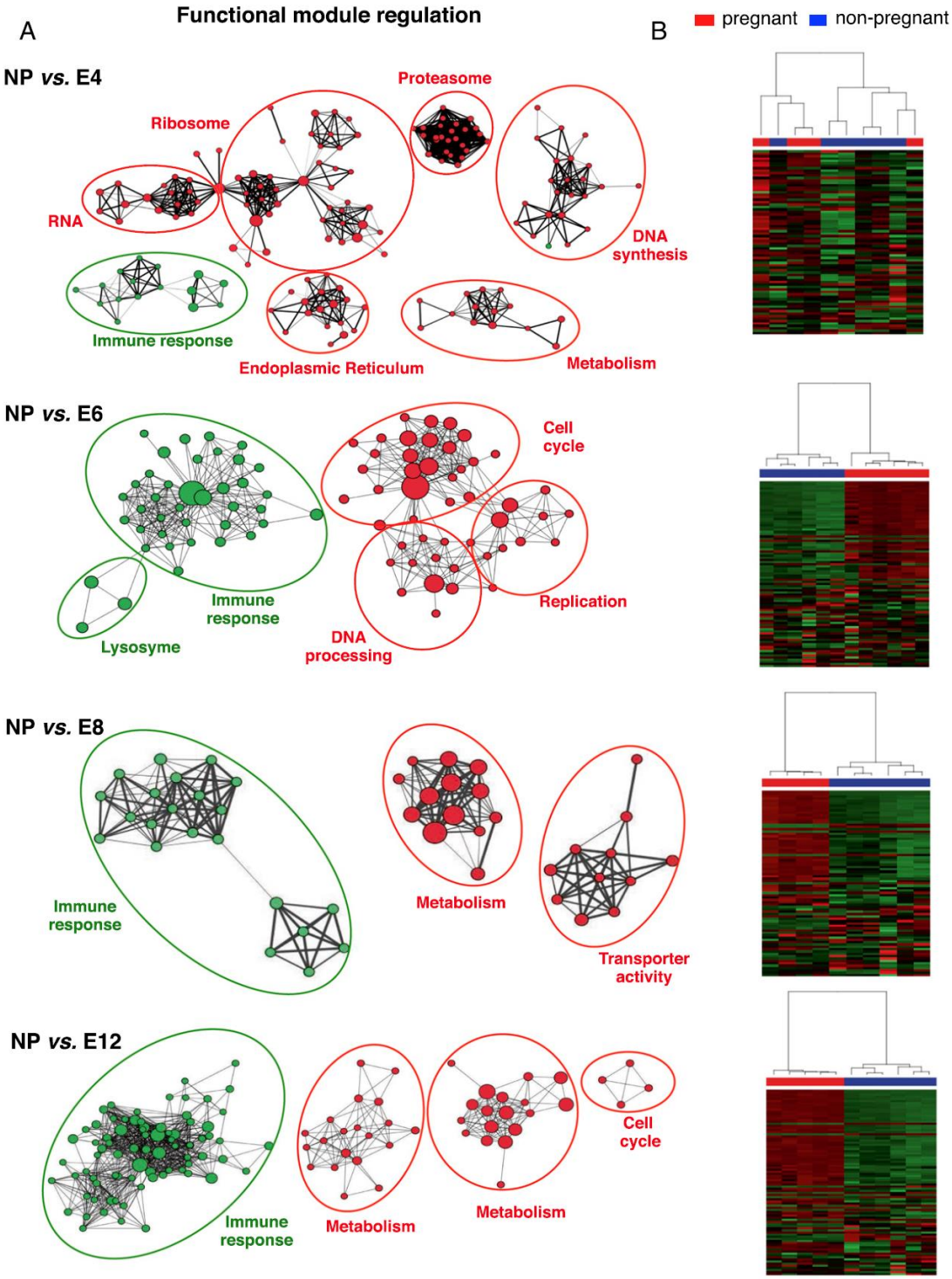


FIGURE 2. Unsupervised analyses of the dynamic regulation of pathways in uterine tissues during pregnancy. (**A**) Cytoscape representation. The enrichments of all ICA-extracted and GO-compiled signatures were tested using GSEA at E4, E6, E8, and E12 (n = 4–6), compared with nonpregnant (NP) mice (n = 6). The significantly enriched signatures (FDR $p < 0.05$) were used to generate functional modules using the Cytoscape software and its Enrichment Map plugin. Functional modules were manually circled and assigned a label based on GO annotation. Nodes represent molecular signatures, and their size is proportional to the number of genes composing the molecular signature; lines connecting nodes reflect mutual overlap between nodes. Shown in red are the upregulated signatures and in green are the downregulated ones. (**B**) Treg signature heat map showing the kinetics of a previously described Treg signature, that is, which is composed of genes that discriminate Foxp3$^+$CD25$^+$ cells from non-Tregs (49), at E4, E6, E8, and E12 (n = 4–6), compared with NP mice (n = 6). Shown in red are upregulated genes and in green are downregulated ones.

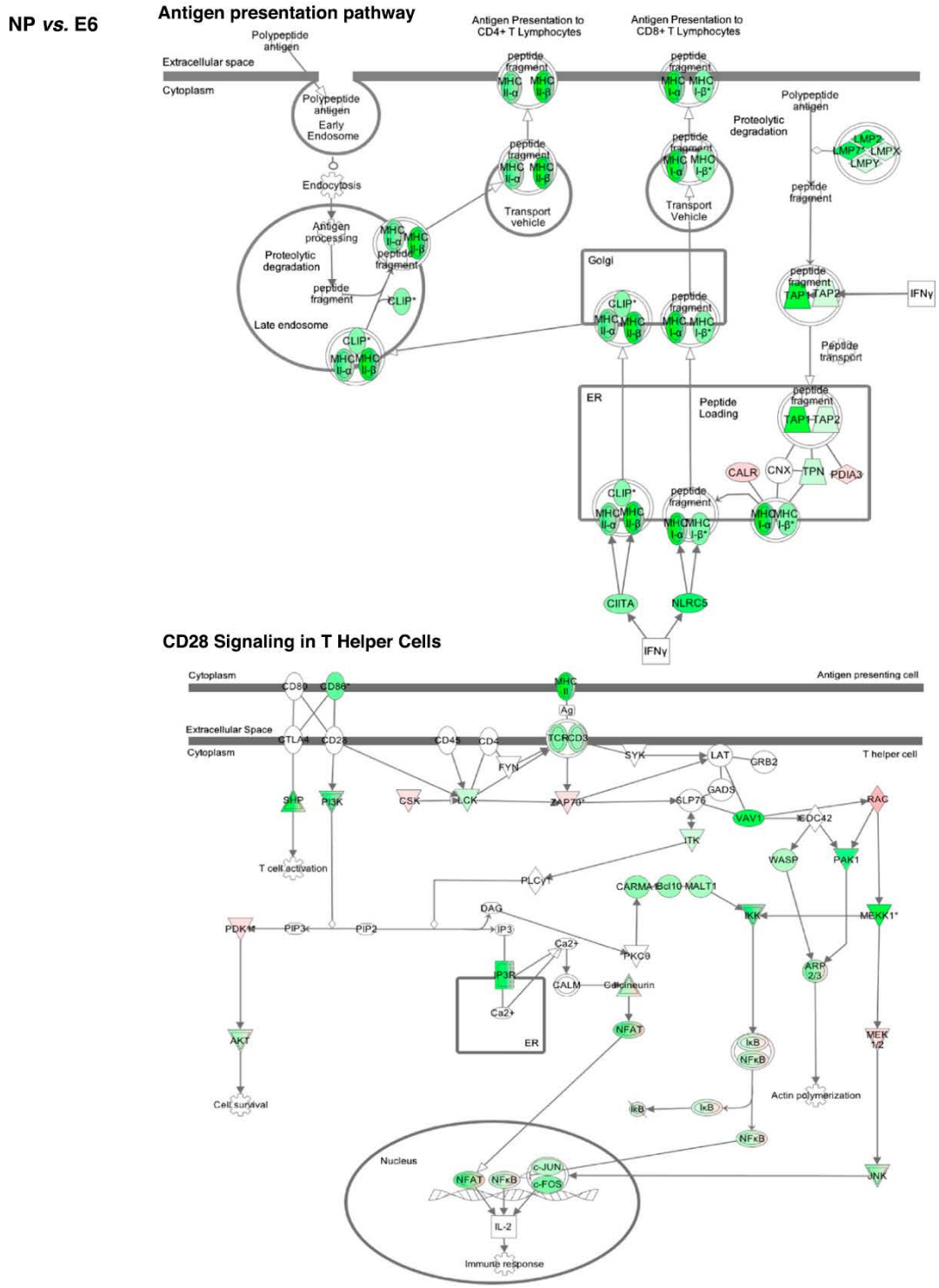


FIGURE 3. Downregulation of Ag presentation and T cell signaling pathways in uterine tissues during pregnancy. Genes identified as being regulated significantly at E6 after fetal implantation were annotated using IPA and two pathways (Ag presentation pathway and CD28 signaling in Th cells) were presented. Shown in red are upregulated genes and in green are downregulated ones. Color intensity represents statistical significance. Data are representative of $n = 4$–6 mice per group. NP, nonpregnant.

fetuses and tolerance to tumor cells on a larger scale, we compared the global transcriptomic changes occurring in the uterus during pregnancy with those changes occurring in the tumor microenvironment during tumor growth. We first analyzed the changes in the tumor microenvironment as we did for the uterus. Supervised analyses of DEGs revealed an early downregulation (as early as day 4) of immune-related pathways, including the Ag presentation pathway (Supplemental Fig. 3) and DC maturation (Supplemental

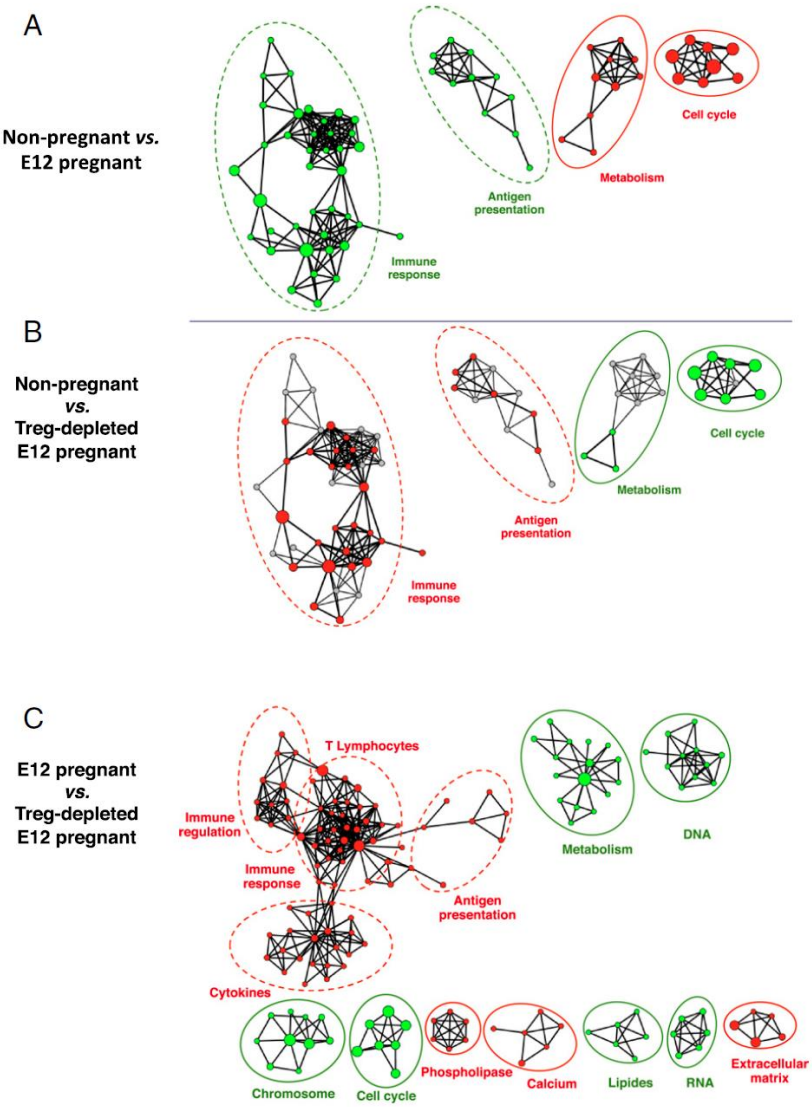


FIGURE 4. Effects of Treg depletion on the regulation of pathways in the pregnant uterine tissues. The signature modules were obtained as described in Fig. 2 and pathways regulated at E12 are shown. Shown in red are upregulated signatures and in green are the downregulated ones. Functional modules were manually circled and assigned a label based on GO annotation. Immune-associated modules are represented in dotted lines. Comparisons with nonpregnant mice of untreated (**A**) or Treg-depleted (**B**) mice are shown. (B) Shown in red are shared upregulated signatures, in green shared downregulated signatures, and in gray nonshared signatures. (**C**) Comparison between Treg-depleted and untreated pregnant mice. Data are representative of $n = 4–5$ mice per group.

Fig. 4B). Unsupervised analyses showed that cell proliferation– and metabolism-related signatures were upregulated, whereas the immune response–annotated signatures were downregulated as early as day 4 after tumor cell inoculation, with increased changes over time (Fig. 5). Thus, the temporal and functional patterns of changes are quite similar in the cancer and pregnancy settings.

We then directly compared signature modulation between the pregnancy and cancer settings, revealing a strikingly similar pattern of gene regulation (Fig. 6). Most of the immune-related signatures that were rapidly downregulated in the pregnancy setting (Fig. 2A, nonpregnant versus E6) were also consistently downregulated in the cancer setting (Fig. 6A). This was also true at later time points, when more immune response–associated signatures were downregulated (Fig. 6B). At any time points between E4 and E12, there was not even a single downregulated immune response–related signature that was upregulated in the cancer setting.

We next analyzed all significantly enriched signatures (with the Benjamini–Hochberg FDR of <0.05) from the GO and KEGG databases, whatever the biological function they represent, shared by the two conditions across all time points. We found a total of 37 common signatures (Fig. 6C). Strikingly, the only 10 upregulated signatures are all related to cell proliferation, whereas the 27 downmodulated ones are all associated with the immune response. Ten of these signatures are related to Ag presentation and six to T lymphocyte function. Among the other signatures, five are labeled as related to diseases, all being allo- or autoimmune diseases.

Overall, these studies reveal that within the complex microenvironment changes accompanying tumor and fetal development, the one and only striking overall similarity is the downmodulation of the immune response. This is true in terms of timing and of mechanisms, with decreases of Ag presentation and T cell activation, and increase of Treg activation. To strengthen the significance of these observations, we performed similarity analyses between up- and downregulated ICA extracted signatures from the two datasets. Jaccard and GOSemSim (36) analyses, which test the similarity of signatures in terms of gene composition and functional annotation, respectively, confirmed the high significance of

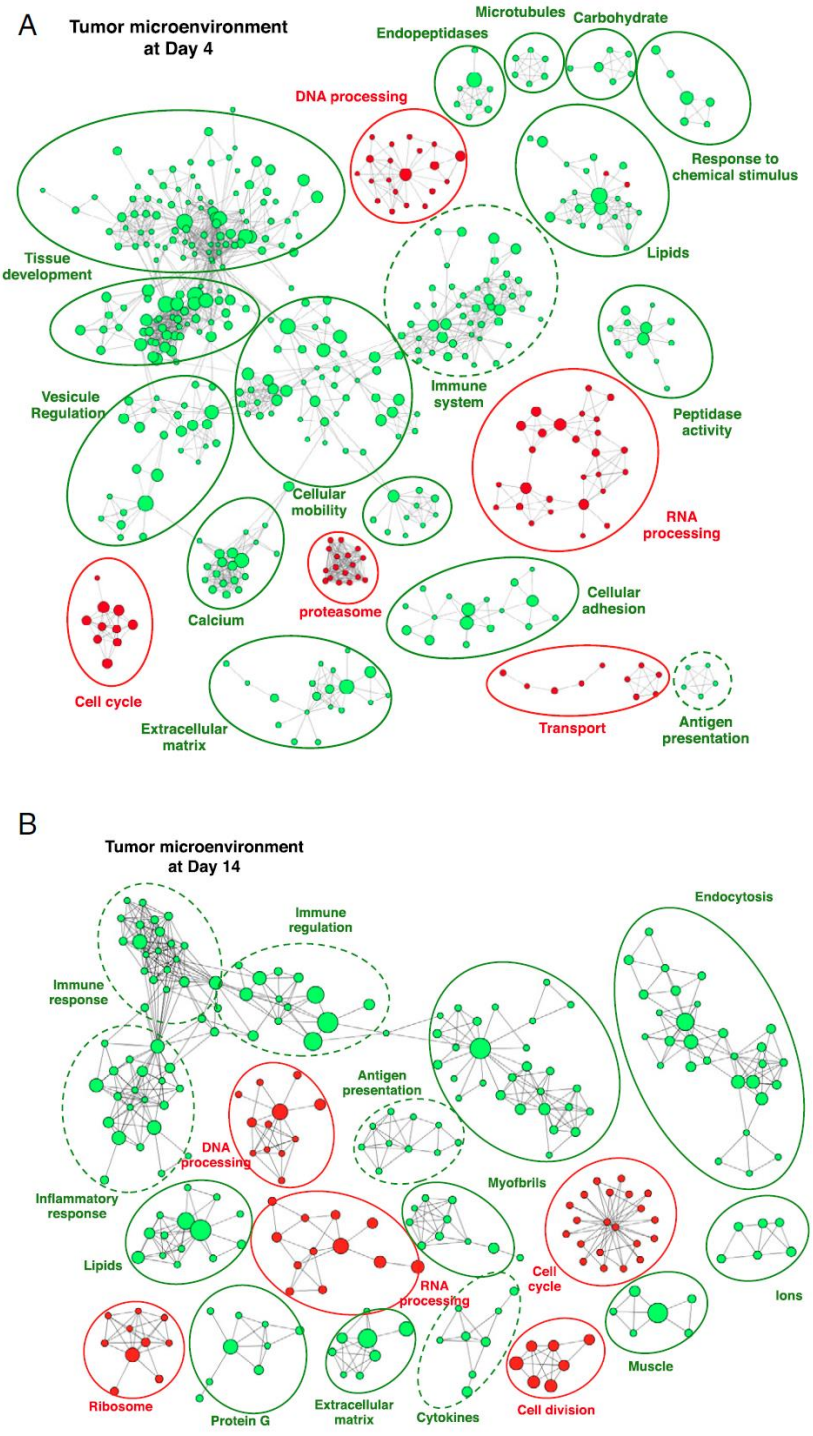
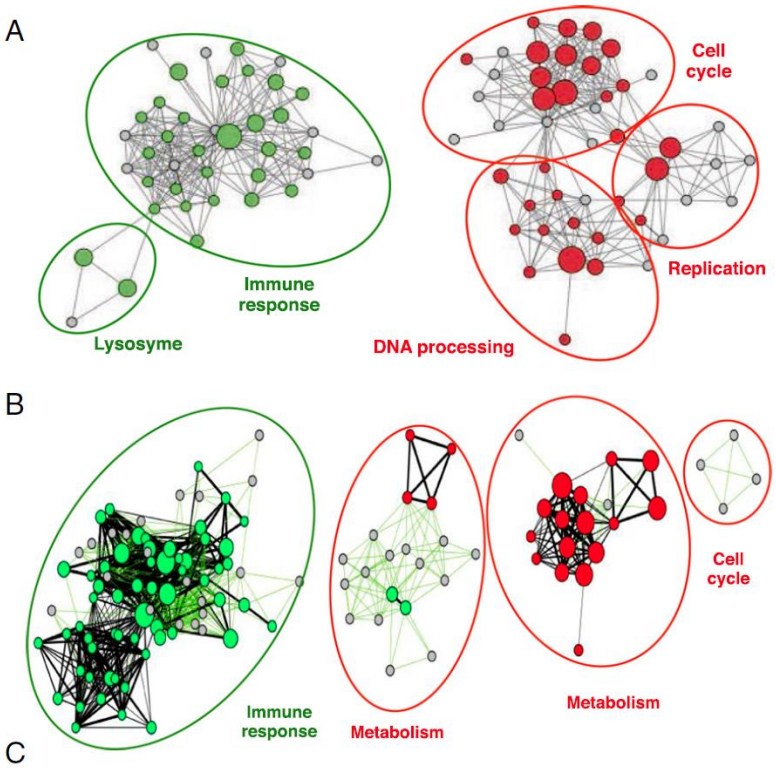


FIGURE 5. Unsupervised analyses of the dynamic regulation of functional pathways in the tumor microenvironment. The enrichments of all ICA-extracted and GO-compiled signatures were tested using GSEA. The significantly enriched signatures (FDR $p < 0.005$) were used to generate functional modules using the Cytoscape software and its Enrichment Map plugin as described in Fig. 2. Nodes represent molecular signatures, and their size is proportional to the number of genes composing the molecular signature; lines connecting nodes reflect mutual overlap between nodes. Shown in red are upregulated signatures and in green are the downregulated ones. (**A**) day 4 and (**B**) day 14 after tumor injection. Data are representative of $n = 4–7$ mice per group.

the above findings (Fig. 7A, Supplemental Fig. 2E). We then focused on the signatures that showed the highest Jaccard and GOSemSim indices, which correspond to those with the highest similarity between the cancer and pregnancy settings. Among the upregulated signatures, there were again those annotated as metabolism and cell cycle (data not shown). Among the downregulated signatures, there were again those associated with DC maturation, IFN signaling, the Ag presentation pathway, and IL-17 signaling (Fig. 7B).

Discussion

Since 1953, pregnancy has been viewed as nature's allograft (43), with the maternal immune system being in direct contact with a semiallogeneic organism, deeply engrafted and invasive, without any sign of rejection. Similarities between fetal and tumor development escaping the immune system were already alluded to >40 y ago (11). The authors concluded that an anti-inflammatory effect is a normal property of trophoblast cells and that "For a cell to become cancerous, it would have to combine a reappearance



C

Molecular signatures	Related process
Antigen_processing_and_presentation_of_peptide_antigen_via_MHC_class_II Antigen_binding Antigen_processing_and_presentation_of_peptide_or_polysaccharide_antigen_via_MHC_class_II Antigen_processing_and_presentation_of_peptide_antigen Antigen_processing_and_presentation_of_exogenous_peptide_antigen Antigen_processing_and_presentation_of_exogenous_peptide_antigen_via_MHC_class_II Antigen_processing_and_presentation Antigen_processing_and_presentation_of_exogenous_antigen MHC_protein_complex MHC_class_II_protein_complex	MHC Complex
Leukocyte_activation Leukocyte_mediated_immunity	Leukocytes
B_cell_activation PFAM__Immunoglobulin_C1-set_domain	B cells
Lymphocyte_mediated_immunity Lymphocyte_activation Lymphocyte_differentiation	Lymphocytes
Immune_response Immune_effector_process Innate_immune_response	Immune response
KEGG__Intestinal_immune_network_for_IgA_production KEGG__Autoimmune_thyroid_disease KEGG__Graft-versus-host_disease KEGG__Type_I_diabetes_mellitus KEGG__Allograft_rejection KEGG__Asthma	Diseases
Calcium_ion_transport	Calcium
mRNA_transport RNA_transport Nucleic_acid_transport Establishment_of_RNA_localization Nucleobase,_nucleoside,_nucleotide_and_nucleic_acid_transport	RNA
M_phase Nuclear_division Mitosis Nuclear_pore Organelle_fission	Cell cycle
Nucleobase,_nucleoside,_nucleotide_and_nucleic_acid_transport	Transport

FIGURE 6. Comparison of regulated pathways in the pregnant uterine tissues and tumor microenvironment. Comparison of significantly enriched signatures from uterine tissues and tumor microenvironment is shown. (**A**) E6 versus day 4; (**B**) E12 versus day 14. Shown in red are shared upregulated signatures, in green shared downregulated signatures, and in gray nonshared signatures. (**C**) Intersection of all signatures. All signatures significantly enriched (up- or downregulated) from the tumor microenvironment (at day 4 and day 14) and uterine tissues (at E4, E6, E8, E10, E11, and E12) were compared. Shown in red is the list of shared upregulated signatures and in green are the downregulated ones.

of trophoblastic properties and at least another lesion altering the regulatory circuits controlling normal division."

The following decades of studies concerning maternal–fetal or tumor cell immune tolerance have emphasized the numerous cell types and mechanisms involved in ensuring the success of pregnancy (9, 12, 44, 45) or the development of tumors (46, 47). However, as most of these studies were focused on individual molecules or cells, our understanding of maternal–fetal and cancer immune tolerance remains fragmented. Understanding the global/operational nature of an immune response, such as in immune tolerance, requires the characterization of its individual components, but also the exploration of the complex interactions and regulatory associations between these components. In this regard, our transcriptome studies revealed that >2000 genes were already significantly modulated at E4, and >5000 from E8 onward. Systems biology has the potential to tackle this

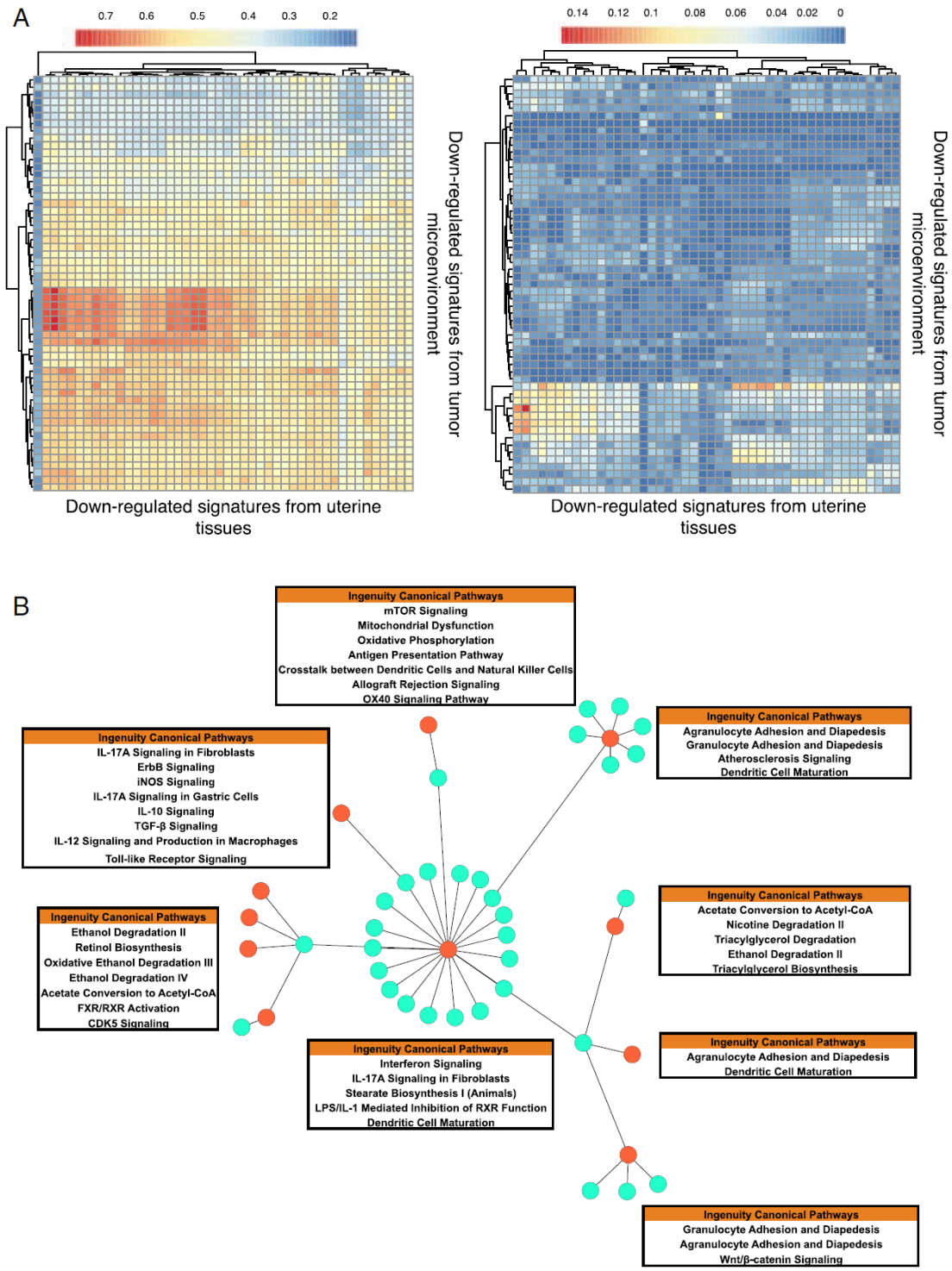


FIGURE 7. Similarity analysis of downregulated signatures from uterine tissues and the tumor microenvironment. (**A**) ICA-extracted downregulated signatures from uterine tissues (UT) and the tumor microenvironment (TM) were compared using the GOSemSim (*left panel*) and Jaccard (*right panel*) similarity indices. The heat map colors represent the similarity index as shown on the scale. The similarity indices range between 0 and 1. (**B**) Using Cytoscape software, signatures with the highest GOSemSim and Jaccard similarity indices were represented as a network. TM signatures are depicted in orange, UT signatures are shown in blue. TM signatures were then functionally annotated using IPA and the immune-related pathways were shown.

huge complexity. Furthermore, using non–hypothesis-driven studies, it has the potential to discover components beyond our current knowledge.

We analyzed the global fetal tissues and tumor microenvironment using whole transcriptome microarray and enrichment analyses with GO gene sets. The gene sets tested in our analyses

cover all the GO biological processes, which notably include developmental, cellular, metabolic, and immune system processes. The latter represent <3% of the entire GO database. It is thus remarkable that, besides genes related to the intense cell proliferation, the only genes revealed by our studies to be robustly modulated are immune response–associated genes, which were all downregulated in both settings (Figs. 2A, 5). Of note, in the cancer setting there were many more signatures that were modulated than in the physiological pregnancy setting (Fig. 5). In fact, many functions other than immune responses, such as cellular adhesion or cellular mobility, were downmodulated. This observation suggests that tumor cells may enlist further pathways to license their development.

At E8 (Fig. 2A) fewer immune-related signatures are represented as compared with both E6 or E12, whereas the total number of significantly modulated signatures is not affected. Because no variability issues were noted within the E8 group, we speculate that this could be due to a hormonal modulation of the immune system, as pregnancy-associated hormones can influence adaptive immune responses (48).

It is also remarkable that all the immune response signatures are downmodulated, whereas one would have expected upregulation of Treg-related signatures. Actually, this could not be the case, as no Treg signatures are represented in the GO database. We thus specifically analyzed the dynamic behavior of Treg signatures from the literature (40–42, 49) and observed that global immune-related signature downregulation coincides with the upregulation of previously published Treg signatures (Fig. 2B, Supplemental Fig. 2A–D). Altogether, these results reveal that immune downmodulation is a hallmark of the pregnancy and cancer microenvironments.

The detailed analysis of the immune-related pathways involved in tumor and fetal protection revealed important similarities. Among others, our studies highlight that Ag presentation, lymphocyte activation, and Treg activation are the central downmodulated pathways (Fig. 3, Supplemental Figs. 3, 4B). Furthermore, IPA annotation of highly similar ICA-extracted signatures from tumor microenvironment and uterine tissues revealed the presence of genes belonging to the DC maturation pathway in almost all shared signatures (Fig. 7B). These findings support the involvement of immature DCs in the establishment of tolerogenic microenvironments, and they are in line with the importance of the cross-talk between Tregs and DCs for maintenance of tolerance (12). Treg-mediated suppression of DC function is mediated at least in part by the interaction of CTLA4 on Tregs with CD80/CD86 on DCs. This results in downregulation of CD80/CD86 expression and defective Teff activation (50, 51), which we saw in our transcriptomic dataset and confirmed by flow cytometry analyses of uterine DCs (Supplemental Fig. 4A). These results indicate that uterine tissues are enriched in immature DCs upon embryo implantation. Such an increase in immature DCs has been described in human decidua in early pregnancy (52). In mice, the number of immature DCs also increases in normal pregnancy following implantation, whereas an increase in mature DCs was reported in pregnant female mice mated in abortion-prone conditions (53). Furthermore, a recent study reported a detailed and quantitative analysis of uterine leukocyte populations from E5.5 to E16.5 of pregnancy, the results of which are in line with our observations (54).

Although systems studies can highlight the importance of specific molecules such as CD80/CD86, our results also emphasize that it would be pointless to try to reduce maternal–fetal immune tolerance to one or a few major molecules/pathways. Actually, we think that further understanding of maternal–fetal immune tolerance will require a systems biology approach on an even larger scale, incorporating, for example, metabolic or microbiota studies, both locally and in the whole organism.

Supervised DEG analysis of uterine tissues identified two waves of immune responses over time: an early one, likely triggered by embryo implantation, and a later one during midgestation (Fig. 1D, Supplemental Fig. 1C). For the early wave, most gene expression changes started at day 6 after fertilization (E6). As embryo implantation takes place at around E4.5, this suggests that physical contact between the embryo derivatives and the uterine tissue is important in triggering immune response modulation. Nevertheless, enrichment analysis using molecular signatures pointed to downregulation of immune-related signatures as early as E4, that is, before embryo implantation (Fig. 2A). Thus, unsupervised purely statistical generation of signatures appears more sensitive than supervised methods and may reveal early changes of immune-annotated signatures that could be related to Treg activation by seminal fluid (55). A recent study demonstrated that a soluble form of CD38 released from seminal vesicles induces tolerogenic DCs and activates Tregs, thereby enhancing maternal immune tolerance and protecting the fetus from rejection (56).

Similar early downmodulation of immune response genes is also observed in the cancer setting. These results confirm the brisk Treg recruitment and activation triggered by embryo or tumor cell implantation (16, 21). This brisk response is due to the mobilization of self-specific memory Tregs, the response of which is faster than that of naive Teffs, and thus predates and pre-empts the Teff response. We also observed a significant modulation of myeloid-derived suppressor cell signatures (57) in the tumor microenvironment, suggesting the involvement of this subset in immune tolerance to tumors (data not shown).

The second wave of downmodulated immune pathways is not so strikingly different, and includes activation marker–related pathways (CD40, 4-1BB, CTLA4) and cytokine pathways (IL-1, IL-9, IL-10, IL-17). It could correspond in part to the recruitment of induced peripheral Tregs, which have also been reported to play an important role in successful pregnancy (58).

The depletion of many cell types mobilized to support normal pregnancy does not result in abortion (10). In contrast, this is the case for Tregs, both thymic Tregs (21, 22) and induced peripheral Tregs (58). This highlights the very peculiar and nonredundant role of Tregs in successful pregnancy. In our experiments, the depletion of Tregs reversed the downmodulation of immune responses genes not just to the baseline state of a nonpregnant uterus, but to an activated state where the same immune signatures that were downmodulated are now upregulated compared with the nonpregnant uterus. This emphasizes the control imposed by Tregs on effector immune responses that would develop otherwise. The combined analyses of the cellular and molecular immune response are compatible with a scenario in which the early recruitment of self-specific memory Tregs imprints a tolerogenic environment by both tuning the DCs toward a tolerogenic functional state and directly suppressing Teffs.

The first viviparous mammals were faced with the development of a sophisticated adaptive immune system (12). The development of placentation in eutherians required the co-development of immune suppression mechanisms. We and others have hypothesized that the selection of Tregs during evolution, whether natural thymic Tregs (21) or peripheral-induced Tregs (58), has been mostly driven by the need to protect the fetus. Our present study, using unsupervised transcriptome studies, confirms the similarities in downregulation of effector immune response during pregnancy and cancer, orchestrated by Tregs.

The American writer Susan Sontag had the prescience to describe her tumor as a "demonic pregnancy." "This lump is alive," "a fetus with its own will" she wrote in *Illness as Metaphor* (59). Thirty-eight years later, her literary vision has been validated by systems biology. We now think that the protection of cancer cells by Tregs became the price paid for the evolution of an efficient protection of embryos. Pregnancy and cancer represent two aspects of immune tolerance, physiological and pathological, respectively. Comparison of these settings is heuristic and identifies shared tolerogenic signatures. Further study of these signatures should now help discover targets for immune intervention to improve or to break immune tolerance. As Tregs are central to both cancer and maternal–fetal tolerance, their manipulation has therapeutic potential. Treg depletion in cancer should help break the tolerogenic environment and improve the efficacy of immunotherapies. In contrast, stimulating Tregs may help control recurrent spontaneous abortion of immune origin. We recently showed that low-dose IL-2 is a safe and specific Treg inducer in humans (60, 61) and that it could prevent recurrent spontaneous abortion in the DBA2/CBA2 model in mice (25). Thus, low-dose IL-2 warrants investigation in recurrent spontaneous abortions and infertility caused by implantation failure.

Acknowledgments

We are grateful to Wassila Carpentier for the generation of the excellent transcriptomic datasets and Wahiba Chaara and Hang-Phuong Pham for advice on statistical and bioinformatics methods.

Disclosures

The authors have no financial conflicts of interest.

References

1. Ferretti, C., L. Bruni, V. Dangles-Marie, A. P. Pecking, and D. Bellet. 2007. Molecular circuits shared by placental and cancer cells, and their implications in the proliferative, invasive and migratory capacities of trophoblasts. *Hum. Reprod. Update* 13: 121–141.
2. Holtan, S. G., D. J. Creedon, P. Haluska, and S. N. Markovic. 2009. Cancer and pregnancy: parallels in growth, invasion, and immune modulation and implications for cancer therapeutic agents. *Mayo Clin. Proc.* 84: 985–1000.
3. Ohlsson, R. 1989. Growth factors, protooncogenes and human placental development. *Cell Differ. Dev.* 28: 1–15.
4. Strickland, S., and W. G. Richards. 1992. Invasion of the trophoblasts. *Cell* 71: 355–357.
5. Genbacev, O., Y. Zhou, J. W. Ludlow, and S. J. Fisher. 1997. Regulation of human placental development by oxygen tension. *Science* 277: 1669–1672.
6. Houghton, A. N. 1994. Cancer antigens: immune recognition of self and altered self. *J. Exp. Med.* 180: 1–4.
7. Preiss, S., T. Kammertoens, C. Lampert, G. Willimsky, and T. Blankenstein. 2005. Tumor-induced antibodies resemble the response to tissue damage. *Int. J. Cancer* 115: 456–462.
8. Schumacher, T. N., and R. D. Schreiber. 2015. Neoantigens in cancer immunotherapy. *Science* 348: 69–74.
9. Arck, P. C., and K. Hecher. 2013. Fetomaternal immune cross-talk and its consequences for maternal and offspring's health. *Nat. Med.* 99: 548–556.
10. Barber, E. M., and J. W. Pollard. 2003. The uterine NK cell population requires IL-15 but these cells are not required for pregnancy nor the resolution of a *Listeria monocytogenes* infection. *J. Immunol.* 171: 37–46.
11. Fauve, R. M., B. Hevin, H. Jacob, J. A. Gaillard, and F. Jacob. 1974. Antiinflammatory effects of murine malignant cells. *Proc. Natl. Acad. Sci. USA* 71: 4052–4056.
12. Ruocco, M. G., G. Chaouat, L. Florez, A. Bensussan, and D. Klatzmann. 2014. Regulatory T-cells in pregnancy: historical perspective, state of the art, and burning questions. *Front. Immunol.* 5: 389.
13. Sasaki, Y., M. Sakai, S. Miyazaki, S. Higuma, A. Shiozaki, and S. Saito. 2004. Decidual and peripheral blood $CD4^+CD25^+$ regulatory T cells in early pregnancy subjects and spontaneous abortion cases. *Mol. Hum. Reprod.* 10: 347–353.
14. Heikkinen, J., M. Möttönen, A. Alanen, and O. Lassila. 2004. Phenotypic characterization of regulatory T cells in the human decidua. *Clin. Exp. Immunol.* 136: 373–378.
15. Jin, L.-P., Q.-Y. Chen, T. Zhang, P.-F. Guo, and D.-J. Li. 2009. The $CD4^+CD25^{bright}$ regulatory T cells and CTLA-4 expression in peripheral and decidual lymphocytes are down-regulated in human miscarriage. *Clin. Immunol.* 133: 402–410.
16. Sasaki, Y., D. Darmochwal-Kolarz, D. Suzuki, M. Sakai, M. Ito, T. Shima, A. Shiozaki, J. Rolinski, and S. Saito. 2007. Proportion of peripheral blood and decidual $CD4^+$ $CD25^{bright}$ regulatory T cells in pre-eclampsia. *Clin. Exp. Immunol.* 149: 139–145.
17. Piccinni, M. P., L. Beloni, C. Livi, E. Maggi, G. Scarselli, and S. Romagnani. 1998. Defective production of both leukemia inhibitory factor and type 2 T-helper cytokines by decidual T cells in unexplained recurrent abortions. *Nat. Med.* 4: 1020–1024.
18. Saito, S. 2000. Cytokine network at the feto-maternal interface. *J. Reprod. Immunol.* 47: 87–103.
19. Curiel, T. J., G. Coukos, L. Zou, X. Alvarez, P. Cheng, P. Mottram, M. Evdemon-Hogan, J. R. Conejo-Garcia, L. Zhang, M. Burow, et al. 2004. Specific recruitment of regulatory T cells in ovarian carcinoma fosters immune privilege and predicts reduced survival. *Nat. Med.* 10: 942–949.
20. Galon, J., A. Costes, F. Sanchez-Cabo, A. Kirilovsky, B. Mlecnik, C. Lagorce-Pagès, M. Tosolini, M. Camus, A. Berger, P. Wind, et al. 2006. Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. *Science* 313: 1960–1964.
21. Darrasse-Jèze, G., A.-S. Bergot, A. Durgeau, F. Billiard, B. L. Salomon, J. L. Cohen, B. Bellier, K. Podsypanina, and D. Klatzmann. 2009. Tumor emergence is sensed by self-specific $CD44^{hi}$ memory Tregs that create a dominant tolerogenic environment for tumors in mice. *J. Clin. Invest.* 119: 2648–2662.
22. Aluvihare, V. R., M. Kallikourdis, and A. G. Betz. 2004. Regulatory T cells mediate maternal tolerance to the fetus. *Nat. Immunol.* 5: 266–271.
23. Darrasse-Jèze, G., D. Klatzmann, F. Charlotte, B. L. Salomon, and J. L. Cohen. 2006. $CD4^+CD25^+$ regulatory/suppressor T cells prevent allogeneic fetus rejection in mice. *Immunol. Lett.* 102: 106–109.
24. Shimizu, J., S. Yamazaki, and S. Sakaguchi. 1999. Induction of tumor immunity by removing $CD25^+CD4^+$ T cells: a common basis between tumor immunity and autoimmunity. *J. Immunol.* 163: 5211–5218.
25. Chen, T., G. Darrasse-Jèze, A.-S. Bergot, T. Courau, G. Churlaud, K. Valdivia, J. L. Strominger, M. G. Ruocco, G. Chaouat, and D. Klatzmann. 2013. Self-specific memory regulatory T cells protect embryos at implantation in mice. *J. Immunol.* 191: 2273–2281.
26. Benoist, C., R. N. Germain, and D. Mathis. 2006. A plaidoyer for "systems immunology". *Immunol. Rev.* 210: 229–234.
27. Chaussabel, D., C. Quinn, J. Shen, P. Patel, C. Glaser, N. Baldwin, D. Stichweh, D. Blankenship, L. Li, I. Munagala, et al. 2008. A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus. *Immunity* 29: 150–164.
28. Querec, T. D., R. S. Akondy, E. K. Lee, W. Cao, H. I. Nakaya, D. Teuwen, A. Pirani, K. Gernert, J. Deng, B. Marzolf, et al. 2009. Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans. *Nat. Immunol.* 10: 116–125.
29. Brandes, M., F. Klauschen, S. Kuchen, and R. N. Germain. 2013. A systems analysis identifies a feedforward inflammatory circuit leading to lethal influenza infection. *Cell* 154: 197–212.
30. Pham, H.-P., N. Dérian, W. Chaara, B. Bellier, D. Klatzmann, and A. Six. 2014. A novel strategy for molecular signature discovery based on independent component analysis. *Int. J. Data Min. Bioinform.* 9: 277–304.
31. Hyvärinen, A., and E. Oja. 2000. Independent component analysis: algorithms and applications. *Neural Netw.* 13: 411–430.
32. Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* 102: 15545–15550.
33. Shannon, P., A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13: 2498–2504.
34. Merico, D., R. Isserlin, O. Stueker, A. Emili, and G. D. Bader. 2010. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS One* 5: e13984.
35. Jaccard, P. 1901. *Bull. Soc. Vaud. Sci. Nat.*
36. Yu, G., F. Li, Y. Qin, X. Bo, Y. Wu, and S. Wang. 2010. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* 26: 976–978.
37. Tzachanis, D., and V. A. Boussiotis. 2009. Tob, a member of the APRO family, regulates immunological quiescence and tumor suppression. *Cell Cycle* 8: 1019–1025.
38. Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, et al. 2000. Gene ontology: tool for the unification of biology. *Nat. Genet.* 25: 25–29.
39. Zheng, Y., S. Z. Josefowicz, A. Kas, T.-T. Chu, M. A. Gavin, and A. Y. Rudensky. 2007. Genome-wide analysis of Foxp3 target genes in developing and mature regulatory T cells. *Nature* 445: 936–940.
40. Marson, A., K. Kretschmer, G. M. Frampton, E. S. Jacobsen, J. K. Polansky, K. D. MacIsaac, S. S. Levine, E. Fraenkel, H. von Boehmer, and R. A. Young. 2007. Foxp3 occupancy and regulation of key target genes during T-cell stimulation. *Nature* 445: 931–935.
41. Karlsson, G., Y. Liu, J. Larsson, M.-J. Goumans, J.-S. Lee, S. S. Thorgeirsson, M. Ringnér, and S. Karlsson. 2005. Gene expression profiling demonstrates that TGF-β1 signals exclusively through receptor complexes involving Alk5 and identifies targets of TGF-β signaling. *Physiol. Genomics* 21: 396–403.
42. Marzec, M., K. Halasa, M. Kasprzycka, M. Wysocka, X. Liu, J. W. Tobias, D. Baldwin, Q. Zhang, N. Odum, A. H. Rook, and M. A. Wasik. 2008. Differential effects of interleukin-2 and interleukin-15 versus interleukin-21 on $CD4^+$ cutaneous T-cell lymphoma cells. *Cancer Res.* 68: 1083–1091.

43. Medawar, P. 1953. Some immunological and endocrinological problems raised by the evolution of viviparity in vertebrates. *Symp. Soc. Exp. Biol.* 7: 320–338.
44. Gobert, M., and J. J. Lafaille. 2012. Maternal-fetal immune tolerance, block by block. *Cell* 150: 7–9.
45. Erlebacher, A. 2013. Mechanisms of T cell tolerance towards the allogeneic fetus. *Nat. Rev. Immunol.* 13: 23–33.
46. Smyth, M. J., Y. Hayakawa, K. Takeda, and H. Yagita. 2002. New aspects of natural-killer-cell surveillance and therapy of cancer. *Nat. Rev. Cancer* 2: 850–861.
47. Liotta, L. A., and E. C. Kohn. 2001. The microenvironment of the tumour-host interface. *Nature* 411: 375–379.
48. Schumacher, A., S.-D. Costa, and A. C. Zenclussen. 2014. Endocrine factors modulating immune responses in pregnancy. *Front. Immunol.* 5: 196.
49. Gavin, M. A., J. P. Rasmussen, J. D. Fontenot, V. Vasta, V. C. Manganiello, J. A. Beavo, and A. Y. Rudensky. 2007. Foxp3-dependent programme of regulatory T-cell differentiation. *Nature* 445: 771–775.
50. Yamazaki, S., T. Iyoda, K. Tarbell, K. Olson, K. Velinzon, K. Inaba, and R. M. Steinman. 2003. Direct expansion of functional $CD25^+$ $CD4^+$ regulatory T cells by antigen-processing dendritic cells. *J. Exp. Med.* 198: 235–247.
51. Serra, P., A. Amrani, J. Yamanouchi, B. Han, S. Thiessen, T. Utsugi, J. Verdaguer, and P. Santamaria. 2003. CD40 ligation releases immature dendritic cells from the control of regulatory $CD4^+CD25^+$ T cells. *Immunity* 19: 877–889.
52. Gardner, L., and A. Moffett. 2003. Dendritic cells in the human decidua. *Biol. Reprod.* 69: 1438–1446.
53. Blois, S., M. Tometten, J. Kandil, E. Hagen, B. F. Klapp, R. A. Margni, and P. C. Arck. 2005. Intercellular adhesion molecule-1/LFA-1 cross talk is a proximate mediator capable of disrupting immune integration and tolerance mechanism at the feto-maternal interface in murine pregnancies. *J. Immunol.* 174: 1820–1829.
54. Habbeddine, M., P. Verbeke, S. Karaz, P. Bobé, and C. Kanellopoulos-Langevin. 2014. Leukocyte population dynamics and detection of IL-9 as a major cytokine at the mouse fetal-maternal interface. *PLoS One* 9: e107267.
55. Robertson, S. A., L. R. Guerin, J. J. Bromfield, K. M. Branson, A. C. Ahlström, and A. S. Care. 2009. Seminal fluid drives expansion of the $CD4^+CD25^+$ T regulatory cell pool and induces tolerance to paternal alloantigens in mice. *Biol. Reprod.* 80: 1036–1045.
56. Kim, B.-J., Y.-M. Choi, S.-Y. Rah, D.-R. Park, S.-A. Park, Y.-J. Chung, S.-M. Park, J. K. Park, K. Y. Jang, and U.-H. Kim. 2015. Seminal CD38 is a pivotal regulator for fetomaternal tolerance. *Proc. Natl. Acad. Sci. USA* 112: 1559–1564.
57. Youn, J.-I., M. Collazo, I. N. Shalova, S. K. Biswas, and D. I. Gabrilovich. 2012. Characterization of the nature of granulocytic myeloid-derived suppressor cells in tumor-bearing mice. *J. Leukoc. Biol.* 91: 167–181.
58. Samstein, R. M., S. Z. Josefowicz, A. Arvey, P. M. Treuting, and A. Y. Rudensky. 2012. Extrathymic generation of regulatory T cells in placental mammals mitigates maternal-fetal conflict. *Cell* 150: 29–38.
59. Sontag, S. 1978. *Illness as Metaphor.* Farrar, Straus and Giroux, New York.
60. Saadoun, D., M. Rosenzwajg, F. Joly, A. Six, F. Carrat, V. Thibault, D. Sene, P. Cacoub, and D. Klatzmann. 2011. Regulatory T-cell responses to low-dose interleukin-2 in HCV-induced vasculitis. *N. Engl. J. Med.* 365: 2067–2077.
61. Klatzmann, D., and A. K. Abbas. 2015. The promise of low-dose interleukin-2 therapy for autoimmune and inflammatory diseases. *Nat. Rev. Immunol.* 15: 283–294.

ANNEXE #3

Liste des signatures enrichies dans les profils de spécialisation des donneurs P1, P6 et P7.

ANNEXE #3

SIGNATURES SPECIFIQUES DE P1

[1] "P1_IL2"
[2] "MALEK_TREG_C2_9"
[3] "NIELSEN_MALIGNAT_FIBROUS_HISTIOCYTOMA_DN"
[4] "MALEK_SIGDN"
[5] "MCBRYAN_PUBERTAL_BREAST_3_4WK_DN"
[6] "HOLLEMAN_DAUNORUBICIN_B_ALL_DN"
[7] "MARZEC_IL2_SIGNALING_DN"
[8] "AMIT_EGF_RESPONSE_60_HELA"
[9] "ZHAN_MULTIPLE_MYELOMA_CD1_AND_CD2_UP"
[10] "BHATI_G2M_ARREST_BY_2METHOXYESTRADIOL_DN"
[11] "CHANDRAN_METASTASIS_TOP50_UP"
[12] "KEGG_COLORECTAL_CANCER"
[13] "REACTOME_TRANSCRIPTIONAL_ACTIVITY_OF_SMAD2_SMAD3_SMAD4_HETEROTRIMER"
[14] "CHIARADONNA_NEOPLASTIC_TRANSFORMATION_CDC25_UP"
[15] "ZHAN_MULTIPLE_MYELOMA_CD1_VS_CD2_DN"
[16] "GAZDA_DIAMOND_BLACKFAN_ANEMIA_ERYTHROID_UP"
[17] "KEGG_MTOR_SIGNALING_PATHWAY"
[18] "BIOCARTA_CTL_PATHWAY"
[19] "MORI_PLASMA_CELL_DN"
[20] "MAHAJAN_RESPONSE_TO_IL1A_DN"
[21] "CHOI_ATL_STAGE_PREDICTOR"
[22] "REACTOME_TRNA_AMINOACYLATION"
[23] "BERNARD_PPAPDC1B_TARGETS_UP"
[24] "REACTOME_TRANSPORT_OF_MATURE_TRANSCRIPT_TO_CYTOPLASM"
[25] "BRACHAT_RESPONSE_TO_CAMPTOTHECIN_UP"
[26] "BIOCARTA_MTOR_PATHWAY"
[27] "LEUKOCYTE DIFFERENTIATION (M160)"
[28] "BOYLAN_MULTIPLE_MYELOMA_PCA3_UP"
[29] "DE_YY1_TARGETS_DN"
[30] "KEGG_LEUKOCYTE_TRANSENDOTHELIAL_MIGRATION"
[31] "PID_LKB1_PATHWAY"
[32] "RHEIN_ALL_GLUCOCORTICOID_THERAPY_UP"
[33] "SANA_TNF_SIGNALING_DN"
[34] "GILMORE_CORE_NFKB_PATHWAY"
[35] "KIM_MYCL1_AMPLIFICATION_TARGETS_DN"

SIGNATURES SPECIFIQUES DE P6

[1] "P6_IL2"
[2] "CHIANG_LIVER_CANCER_SUBCLASS_INTERFERON_UP"
[3] "ZUCCHI_METASTASIS_UP"
[4] "HUMMEL_BURKITTS_LYMPHOMA_DN"
[5] "LEUKOCYTE ACTIVATION AND MIGRATION (M45)"

[6] "PID_MYC_ACTIVPATHWAY"

[7] "TONKS_TARGETS_OF_RUNX1_RUNX1T1_FUSION_MONOCYTE_DN"

[8] "REACTOME_SYNTHESIS_AND_INTERCONVERSION_OF_NUCLEOTIDE_DI_AND_TRIPHOSPHATES"

[9] "AKL_HTLV1_INFECTION_UP"

[10] "REACTOME_IL_3_5_AND_GM_CSF_SIGNALING"

[11] "WANG_ADIPOGENIC_GENES_REPRESSED_BY_SIRT1"

[12] "KRASNOSELSKAYA_ILF3_TARGETS_UP"

[13] "BURTON_ADIPOGENESIS_8"

[14] "REACTOME_PROTEIN_FOLDING"

[15] "RASHI_NFKB1_TARGETS"

[16] "JACKSON_DNMT1_TARGETS_UP"

[17] "TOMLINS_PROSTATE_CANCER_UP"

[18] "BLOOD COAGULATION (M11.1)"

[19] "NUNODA_RESPONSE_TO_DASATINIB_IMATINIB_UP"

SIGNATURES SPECIFIQUES DE P7

[1] "P7_IL2"

[2] "MALEK_TREG_C1_5"

[3] "MALEK_TREG_C1_10"

[4] "REACTOME_MITOTIC_G1_G1_S_PHASES"

[5] "SCHLOSSER_SERUM_RESPONSE_AUGMENTED_BY_MYC"

[6] "ONO_AML1_TARGETS_DN"

[7] "REACTOME_ASSEMBLY_OF_THE_PRE_REPLICATIVE_COMPLEX"

[8] "MARTORIATI_MDM4_TARGETS_NEUROEPITHELIUM_UP"

[9] "REACTOME_NEGATIVE_REGULATORS_OF_RIG_I_MDA5_SIGNALING"

[10]
"REACTOME_RESPIRATORY_ELECTRON_TRANSPORT_ATP_SYNTHESIS_BY_CHEMIOSMOTIC_COUPLING_AND_HEAT_PRODUCTION_BY_UNCOUPLING_PROTEINS_"

[11] "KUROZUMI_RESPONSE_TO_ONCOCYTIC_VIRUS"

[12] "BIOCARTA_CHEMICAL_PATHWAY"

[13] "REACTOME_G1_S_TRANSITION"

[14] "KEGG_LYSOSOME"

[15] "MORI_IMMATURE_B_LYMPHOCYTE_DN"

[16] "LIN_APC_TARGETS"

[17] "ROSS_ACUTE_MYELOID_LEUKEMIA_CBF"

[18] "LI_DCP2_BOUND_MRNA"

[19] "REACTOME_CROSS_PRESENTATION_OF_SOLUBLE_EXOGENOUS_ANTIGENS_ENDOSOMES"

[20] "REACTOME_SYNTHESIS_OF_DNA"

[21] "BAUS_TFF2_TARGETS_UP"

[22] "IZADPANAH_STEM_CELL_ADIPOSE_VS_BONE_UP"

[23] "TREG_PFOERTNER"

[24] "PID_IL4_2PATHWAY"

[25] "MALEK_TREG_C2_15"

[26] "REACTOME_RIG_I_MDA5_MEDIATED_INDUCTION_OF_IFN_ALPHA_BETA_PATHWAYS"

[27] "BIOCARTA_IL12_PATHWAY"
[28] "WANG_ESOPHAGUS_CANCER_VS_NORMAL_UP"
[29] "KEGG_NON_SMALL_CELL_LUNG_CANCER"
[30] "KOKKINAKIS_METHIONINE_DEPRIVATION_48HR_UP"
[31] "MOOTHA_GLUCONEOGENESIS"
[32] "KEGG_OXIDATIVE_PHOSPHORYLATION"
[33] "MELLMAN_TUT1_TARGETS_DN"
[34] "SHIN_B_CELL_LYMPHOMA_CLUSTER_8"
[35] "PUIFFE_INVASION_INHIBITED_BY_ASCITES_UP"
[36] "BIOCARTA_BIOPEPTIDES_PATHWAY"
[37] "MORI_PRE_BI_LYMPHOCYTE_UP"
[38] "KEGG_PARKINSONS_DISEASE"
[39] "RESPIRATORY ELECTRON TRANSPORT CHAIN (MITOCHONDRION) (M219)"
[40] "NUNODA_RESPONSE_TO_DASATINIB_IMATINIB_DN"
[41] "HASLINGER_B_CLL_WITH_11Q23_DELETION"
[42] "HAN_JNK_SINGALING_UP"
[43] "HOUSTIS_ROS"
[44] "REACTOME_SIGNALING_BY_EGFR_IN_CANCER"
[45] "ST_FAS_SIGNALING_PATHWAY"
[46] "PID_CERAMIDE_PATHWAY"
[47] "LUI_THYROID_CANCER_CLUSTER_4"
[48] "CUI_GLUCOSE_DEPRIVATION"
[49] "BIOCARTA_HIVNEF_PATHWAY"
[50] "PID_ECADHERIN_NASCENTAJ_PATHWAY"
[51] "SIG_BCR_SIGNALING_PATHWAY"
[52] "HOFFMANN_IMMATURE_TO_MATURE_B_LYMPHOCYTE_UP"
[53] "BIOCARTA_DEATH_PATHWAY"
[54] "DORN_ADENOVIRUS_INFECTION_12HR_DN"
[55] "YAGI_AML_RELAPSE_PROGNOSIS"
[56] "SHIN_B_CELL_LYMPHOMA_CLUSTER_3"
[57] "MALEK_TREG_C1_12"
[58] "KEGG_JAK_STAT_SIGNALING_PATHWAY"
[59] "GARGALOVIC_RESPONSE_TO_OXIDIZED_PHOSPHOLIPIDS_YELLOW_UP"
[60] "DAUER_STAT3_TARGETS_UP"
[61] "VARELA_ZMPSTE24_TARGETS_UP"
[62] "REACTOME_M_G1_TRANSITION"
[63] "HOFMANN_CELL_LYMPHOMA_UP"
[64] "TIAN_TNF_SIGNALING_VIA_NFKB"
[65] "BURTON_ADIPOGENESIS_5"
[66] "HAHTOLA_MYCOSIS_FUNGOIDES_SKIN_DN"
[67] "BIOCARTA_NO2IL12_PATHWAY"
[68] "HEDENFALK_BREAST_CANCER_HEREDITARY_VS_SPORADIC"

[69] "MISSIAGLIA_REGULATED_BY_METHYLATION_DN"
[70] "KHETCHOUMIAN_TRIM24_TARGETS_UP"
[71] "BYSTRYKH_HEMATOPOIESIS_STEM_CELL_AND_BRAIN_QTL_TRANS"
[72] "CHOI_ATL_CHRONIC_VS_ACUTE_DN"
[73] "BIOCARTA_PGC1A_PATHWAY"
[74] "PID_CASPASE_PATHWAY"
[75] "WILENSKY_RESPONSE_TO_DARAPLADIB"
[76] "SASSON_RESPONSE_TO_FORSKOLIN_UP"
[77] "PID_ALK1PATHWAY"
[78] "LEE_LIVER_CANCER_MYC_TGFA_UP"
[79] "PID_IFNGPATHWAY"
[80] "UDAYAKUMAR_MED1_TARGETS_UP"
[81] "BRACHAT_RESPONSE_TO_METHOTREXATE_DN"
[82] "CHENG_RESPONSE_TO_NICKEL_ACETATE"
[83] "PID_HIVNEFPATHWAY"
[84] "TBA (M32.5)"
[85] "KAAB_FAILED_HEART_VENTRICLE_DN"
[86] "BIOCARTA_IL10_PATHWAY"
[87] "COLLER_MYC_TARGETS_UP"
[88] "DAIRKEE_CANCER_PRONE_RESPONSE_BPA_E2"
[89] "ST_INTERLEUKIN_4_PATHWAY"
[90] "PID_RHOA_REG_PATHWAY"
[91] "REACTOME_PYRUVATE_METABOLISM_AND_CITRIC_ACID_TCA_CYCLE"
[92] "HOLLEMAN_ASPARAGINASE_RESISTANCE_ALL_UP"
[93] "REACTOME_MRNA_SPLICING_MINOR_PATHWAY"
[94] "KEGG_CHRONIC_MYELOID_LEUKEMIA"
[95] "LEE_METASTASIS_AND_RNA_PROCESSING_UP"
[96] "GERHOLD_ADIPOGENESIS_DN"
[97] "HOEBEKE_LYMPHOID_STEM_CELL_UP"
[98] "LIANG_HEMATOPOIESIS_STEM_CELL_NUMBER_LARGE_VS_TINY_DN"
[99] "REACTOME_DARPP_32_EVENTS"
[100] "ZHAN_V2_LATE_DIFFERENTIATION_GENES"
[101] "PID_NFAT_3PATHWAY"
[102] "TYPE I INTERFERON RESPONSE (M127)"
[103] "ZHANG_TLX_TARGETS_DN"
[104] "WANG_ESOPHAGUS_CANCER_VS_NORMAL_DN"
[105] "KIM_ALL_DISORDERS_DURATION_CORR_DN"
[106] "PROINFLAMMATORY DENDRITIC CELL, MYELOID CELL RESPONSE (M86.1)"
[107] "AMUNDSON_POOR_SURVIVAL_AFTER_GAMMA_RADIATION_8G"
[108] "KOKKINAKIS_METHIONINE_DEPRIVATION_48HR_DN"
[109] "YAO_TEMPORAL_RESPONSE_TO_PROGESTERONE_CLUSTER_10"
[110] "T CELL DIFFERENTIATION VIA ITK AND PKC (M18)"

[111] "REACTOME_RECYCLING_PATHWAY_OF_L1"
[112] "SEMENZA_HIF1_TARGETS"
[113] "GAVIN_FOXP3_TARGETS_CLUSTER_P6"
[114] "MULLIGAN_NTF3_SIGNALING_VIA_INSR_AND_IGF1R_UP"
[115] "BIOCARTA_IL22BP_PATHWAY"
[116] "NAKAYAMA_SOFT_TISSUE_TUMORS_PCA1_UP"
[117] "PEART_HDAC_PROLIFERATION_CLUSTER_DN"
[118] "BIOCARTA_TOB1_PATHWAY"
[119] "HOFFMANN_SMALL_PRE_BII_TO_IMMATURE_B_LYMPHOCYTE_UP"
[120] "T CELL DIFFERENTIATION (M14)"
[121] "KAMMINGA_EZH2_TARGETS"
[122] "LIU_COMMON_CANCER_GENES"
[123] "GOLUB_ALL_VS_AML_DN"
[124] "ZHAN_LATE_DIFFERENTIATION_GENES_DN"

SIGNATURES SPECIFIQUES DE P1 et P6

[1] "MALEK_TREG_C1_1"
[2] "MALEK_TREG_C2_10"
[3] "REACTOME_DOWNREGULATION_OF_TGF_BETA_RECEPTOR_SIGNALING"
[4] "YANG_MUC2_TARGETS_DUODENUM_3MO_DN"
[5] "T CELL ACTIVATION (III) (M7.4)"
[6] "HUTTMANN_B_CLL_POOR_SURVIVAL_DN"
[7] "LINDGREN_BLADDER_CANCER_CLUSTER_2A_DN"
[8] "SESTO_RESPONSE_TO_UV_C8"
[9] "NAKAYAMA_FRA2_TARGETS"
[10] "ZAMORA_NOS2_TARGETS_DN"
[11] "KEGG_ENDOMETRIAL_CANCER"
[12] "ZHENG_RESPONSE_TO_ARSENITE_UP"
[13] "GRAHAM_CML_QUIESCENT_VS_NORMAL_QUIESCENT_DN"
[14] "CAFFAREL_RESPONSE_TO_THC_UP"
[15] "BAKER_HEMATOPOIESIS_STAT3_TARGETS"
[16] "ZHAN_MULTIPLE_MYELOMA_PR_DN"
[17] "REACTOME_GLUTATHIONE_CONJUGATION"
[18] "BIOCARTA_FCER1_PATHWAY"
[19] "KEGG_PROTEIN_EXPORT"
[20] "MARIADASON_RESPONSE_TO_CURCUMIN_SULINDAC_7"
[21] "GUTIERREZ_MULTIPLE_MYELOMA_DN"

SIGNATURES SPECIFIQUES DE P1 et P7

[1] "MALEK_TREG_C1_13"
[2] "LINDSTEDT_DENDRITIC_CELL_MATURATION_D"
[3] "GENTLES_LEUKEMIC_STEM_CELL_UP"

[4] "BAKKER_FOXO3_TARGETS_UP"
[5] "GAVIN_FOXP3_TARGETS_CLUSTER_P3"
[6] "BRACHAT_RESPONSE_TO_METHOTREXATE_UP"
[7] "ST_GA13_PATHWAY"
[8] "KEGG_AMINOACYL_TRNA_BIOSYNTHESIS"
[9] "HERNANDEZ_MITOTIC_ARREST_BY_DOCETAXEL_2_UP"
[10] "KERLEY_RESPONSE_TO_CISPLATIN_UP"
[11] "ZHAN_MULTIPLE_MYELOMA_HP_DN"
[12] "KEEN_RESPONSE_TO_ROSIGLITAZONE_DN"
[13] "MISSIAGLIA_REGULATED_BY_METHYLATION_UP"
[14] "KARLSSON_TGFB1_TARGETS_UP"
[15] "ZHONG_SECRETOME_OF_LUNG_CANCER_AND_ENDOTHELIUM"
[16] "PARK_TRETINOIN_RESPONSE_AND_PML_RARA_FUSION"
[17] "SASSON_RESPONSE_TO_GONADOTROPHINS_DN"
[18] "PID_IL6_7PATHWAY"
[19] "MARSON_FOXP3_TARGETS_DN"
[20] "PID_FAK_PATHWAY"
[21] "LEE_AGING_NEOCORTEX_UP"
[22] "GALE_APL_WITH_FLT3_MUTATED_DN"
[23] "PID_VEGFR1_2_PATHWAY"
[24] "ONGUSAHA_TP53_TARGETS"
[25] "KEGG_P53_SIGNALING_PATHWAY"
[26] "BARRIER_COLON_CANCER_RECURRENCE_DN"
[27] "ODONNELL_TARGETS_OF_MYC_AND_TFRC_UP"
[28] "KIM_TIAL1_TARGETS"
[29] "JIANG_AGING_HYPOTHALAMUS_DN"
[30] "REACTOME_UNFOLDED_PROTEIN_RESPONSE"
[31] "MA_MYELOID_DIFFERENTIATION_DN"
[32] "CHIARETTI_ACUTE_LYMPHOBLASTIC_LEUKEMIA_ZAP70"
[33] "ST_B_CELL_ANTIGEN_RECEPTOR"
[34] "BIOCARTA_TCYTOTOXIC_PATHWAY"
[35] "PID_TXA2PATHWAY"
[36] "MARSHALL_VIRAL_INFECTION_RESPONSE_DN"
[37] "LEE_LIVER_CANCER_MYC_UP"
[38] "IGLESIAS_E2F_TARGETS_UP"
[39] "MACLACHLAN_BRCA1_TARGETS_UP"
[40] "PARK_APL_PATHOGENESIS_DN"
[41] "CAIRO_PML_TARGETS_BOUND_BY_MYC_UP"

SIGNATURES SPECIFIQUES DE P6 et P7

[1] "BYSTROEM_CORRELATED_WITH_IL5_UP"
[2] "PID_GMCSF_PATHWAY"

[3] "PELLICCIOTTA_HDAC_IN_ANTIGEN_PRESENTATION_UP"
[4] "FUNG_IL2_SIGNALING_1"
[5] "GREGORY_SYNTHETIC_LETHAL_WITH_IMATINIB"
[6] "JOSEPH_RESPONSE_TO_SODIUM_BUTYRATE_UP"
[7] "BENNETT_SYSTEMIC_LUPUS_ERYTHEMATOSUS"
[8] "MEINHOLD_OVARIAN_CANCER_LOW_GRADE_DN"
[9] "KIM_GLIS2_TARGETS_UP"
[10] "AMIT_EGF_RESPONSE_480_HELA"
[11] "BIOCARTA_IL2RB_PATHWAY"
[12] "YAGI_AML_SURVIVAL"
[13] "ENRICHED IN ANTIGEN PRESENTATION (I) (M71)"
[14] "REACTOME_ANTIGEN_ACTIVATES_B_CELL_RECEPTOR_LEADING_TO_GENERATION_OF_SECOND_MESSENGERS"
[15] "PLATELET ACTIVATION (II) (M32.1)"
[16] "REACTOME_NGF_SIGNALLING_VIA_TRKA_FROM_THE_PLASMA_MEMBRANE"
[17] "BIOCARTA_STATHMIN_PATHWAY"
[18] "BAELDE_DIABETIC_NEPHROPATHY_UP"
[19] "REACTOME_METABOLISM_OF_NUCLEOTIDES"
[20] "BIOCARTA_CTLA4_PATHWAY"
[21] "SCHLOSSER_SERUM_RESPONSE_UP"
[22] "KORKOLA_SEMINOMA_UP"
[23] "KEGG_PROTEASOME"
[24] "SIG_IL4RECEPTOR_IN_B_LYPHOCYTES"
[25] "REACTOME_IL_2_SIGNALING"
[26] "REACTOME_DOWNSTREAM_SIGNAL_TRANSDUCTION"
[27] "ND001_PBS_SPLEENVSTREG_C1_1"
[28] "PID_PI3KCIAKTPATHWAY"
[29] "REACTOME_SIGNALING_BY_PDGF"
[30] "REACTOME_SCF_BETA_TRCP_MEDIATED_DEGRADATION_OF_EMI1"
[31] "REACTOME_SIGNALLING_TO_RAS"
[32] "CAFFAREL_RESPONSE_TO_THC_24HR_5_DN"
[33] "KORKOLA_YOLK_SAC_TUMOR_UP"
[34] "YAO_TEMPORAL_RESPONSE_TO_PROGESTERONE_CLUSTER_12"
[35] "BIOCARTA_KERATINOCYTE_PATHWAY"
[36] "KEGG_INTESTINAL_IMMUNE_NETWORK_FOR_IGA_PRODUCTION"
[37] "REACTOME_SIGNALLING_TO_ERKS"
[38] "NUTT_GBM_VS_AO_GLIOMA_UP"
[39] "HOFFMANN_LARGE_TO_SMALL_PRE_BII_LYMPHOCYTE_UP"
[40] "ST_PHOSPHOINOSITIDE_3_KINASE_PATHWAY"
[41] "CHEN_LUNG_CANCER_SURVIVAL"
[42] "SCHUHMACHER_MYC_TARGETS_UP"
[43] "HENDRICKS_SMARCA4_TARGETS_UP"
[44] "LENAOUR_DENDRITIC_CELL_MATURATION_DN"

[45] "LINDSTEDT_DENDRITIC_CELL_MATURATION_A"

[46] "LIU_NASOPHARYNGEAL_CARCINOMA"

[47] "REACTOME_APC_C_CDH1_MEDIATED_DEGRADATION_OF_CDC20_AND_OTHER_APC_C_CDH1_TARGETED_PROTEINS_IN_LATE_MITOSIS_EARLY_G1"

[48] "BRUNO_HEMATOPOIESIS"

SIGNATURES SPECIFIQUES DE P1, P6 et P7

[1] "REACTOME_PEPTIDE_CHAIN_ELONGATION"

[2] "REACTOME_3_UTR_MEDIATED_TRANSLATIONAL_REGULATION"

[3] "KEGG_RIBOSOME"

[4] "REACTOME_TRANSLATION"

[5] "REACTOME_SRP_DEPENDENT_COTRANSLATIONAL_PROTEIN_TARGETING_TO_MEMBRANE"

[6] "REACTOME_NONSENSE_MEDIATED_DECAY_ENHANCED_BY_THE_EXON_JUNCTION_COMPLEX"

[7] "REACTOME_INFLUENZA_VIRAL_RNA_TRANSCRIPTION_AND_REPLICATION"

[8] "LEE_EARLY_T_LYMPHOCYTE_DN"

[9] "REACTOME_INFLUENZA_LIFE_CYCLE"

[10] "CHAUHAN_RESPONSE_TO_METHOXYESTRADIOL_DN"

[11] "BILANGES_SERUM_AND_RAPAMYCIN_SENSITIVE_GENES"

[12] "WIELAND_UP_BY_HBV_INFECTION"

[13] "REACTOME_ACTIVATION_OF_THE_MRNA_UPON_BINDING_OF_THE_CAP_BINDING_COMPLEX_AND_EIFS_AND_SUBSEQUENT_BINDING_TO_43S"

[14] "MALEK_TREG_C2_11"

[15] "REACTOME_INTERFERON_SIGNALING"

[16] "ALCALA_APOPTOSIS"

[17] "REACTOME_FORMATION_OF_THE_TERNARY_COMPLEX_AND_SUBSEQUENTLY_THE_43S_COMPLEX"

[18] "MALEK_TREG_C1_3"

[19] "KEGG_VIRAL_MYOCARDITIS"

[20] "DANG_MYC_TARGETS_UP"

[21] "KEGG_ANTIGEN_PROCESSING_AND_PRESENTATION"

[22] "REACTOME_INTERFERON_GAMMA_SIGNALING"

[23] "FLECHNER_PBL_KIDNEY_TRANSPLANT_OK_VS_DONOR_UP"

[24] "FLOTHO_PEDIATRIC_ALL_THERAPY_RESPONSE_UP"

[25] "KEGG_TYPE_I_DIABETES_MELLITUS"

[26] "BOSCO_ALLERGEN_INDUCED_TH2_ASSOCIATED_MODULE"

[27] "MALEK_TREG_C1_4"

[28] "ENRICHED IN T CELLS (I) (M7.0)"

[29] "MALEK_TREG_C2_1"

[30] "DER_IFN_BETA_RESPONSE_UP"

[31] "PICCALUGA_ANGIOIMMUNOBLASTIC_LYMPHOMA_DN"

[32] "BASSO_CD40_SIGNALING_UP"

[33] "CHNG_MULTIPLE_MYELOMA_HYPERPLOID_UP"

[34] "TIEN_INTESTINE_PROBIOTICS_6HR_UP"

[35] "ICHIBA_GRAFT_VERSUS_HOST_DISEASE_D7_UP"

[36] "DAZARD_RESPONSE_TO_UV_SCC_UP"

[37] "DIRMEIER_LMP1_RESPONSE_EARLY"
[38] "NEMETH_INFLAMMATORY_RESPONSE_LPS_UP"
[39] "KEGG_GRAFT_VERSUS_HOST_DISEASE"
[40] "FLECHNER_BIOPSY_KIDNEY_TRANSPLANT_REJECTED_VS_OK_UP"
[41] "SANA_RESPONSE_TO_IFNG_UP"
[42] "REGULATION OF ANTIGEN PRESENTATION AND IMMUNE RESPONSE (M5.0)"
[43] "RADAEVA_RESPONSE_TO_IFNA1_UP"
[44] "YU_MYC_TARGETS_DN"
[45] "REACTOME_TCR_SIGNALING"
[46] "HUANG_GATA2_TARGETS_UP"
[47] "HILLION_HMGA1B_TARGETS"
[48] "MUNSHI_MULTIPLE_MYELOMA_UP"
[49] "REACTOME_ANTIGEN_PRESENTATION_FOLDING_ASSEMBLY_AND_PEPTIDE_LOADING_OF_CLASS_I_MHC"
[50] "MALEK_TREG_C2_6"
[51] "SEIDEN_ONCOGENESIS_BY_MET"
[52] "GAVIN_FOXP3_TARGETS_CLUSTER_T4"
[53] "GNATENKO_PLATELET_SIGNATURE"
[54] "SCHLOSSER_MYC_TARGETS_REPRESSED_BY_SERUM"
[55] "KEGG_ALLOGRAFT_REJECTION"
[56] "ZHONG_SECRETOME_OF_LUNG_CANCER_AND_FIBROBLAST"
[57] "TAKAO_RESPONSE_TO_UVB_RADIATION_UP"
[58] "DER_IFN_ALPHA_RESPONSE_UP"
[59] "PECE_MAMMARY_STEM_CELL_DN"
[60] "CASORELLI_ACUTE_PROMYELOCYTIC_LEUKEMIA_UP"
[61] "FARMER_BREAST_CANCER_CLUSTER_1"
[62] "HONMA_DOCETAXEL_RESISTANCE"
[63] "PID_TCR_PATHWAY"
[64] "GAVIN_FOXP3_TARGETS_CLUSTER_T7"
[65] "PELLICCIOTTA_HDAC_IN_ANTIGEN_PRESENTATION_DN"
[66] "STEARMAN_LUNG_CANCER_EARLY_VS_LATE_UP"
[67] "ZHONG_SECRETOME_OF_LUNG_CANCER_AND_MACROPHAGE"
[68] "MORI_PRE_BI_LYMPHOCYTE_DN"
[69] "JOHNSTONE_PARVB_TARGETS_2_UP"
[70] "FERRANDO_T_ALL_WITH_MLL_ENL_FUSION_UP"
[71] "REACTOME_ANTIVIRAL_MECHANISM_BY_IFN_STIMULATED_GENES"
[72] "MALEK_TREG_C2_3"
[73] "BIOCARTA_TCR_PATHWAY"
[74] "HAHTOLA_MYCOSIS_FUNGOIDES_CD4_DN"
[75] "YAO_TEMPORAL_RESPONSE_TO_PROGESTERONE_CLUSTER_11"
[76] "REACTOME_PHOSPHORYLATION_OF_CD3_AND_TCR_ZETA_CHAINS"
[77] "REACTOME_SIGNALING_BY_RHO_GTPASES"
[78] "LAIHO_COLORECTAL_CANCER_SERRATED_UP"

[79] "GAZDA_DIAMOND_BLACKFAN_ANEMIA_PROGENITOR_UP"
[80] "REACTOME_ER_PHAGOSOME_PATHWAY"
[81] "MORI_SMALL_PRE_BII_LYMPHOCYTE_DN"
[82] "BOYAULT_LIVER_CANCER_SUBCLASS_G5_DN"
[83] "ZHOU_TNF_SIGNALING_4HR"
[84] "PID_CXCR4_PATHWAY"
[85] "ADDYA_ERYTHROID_DIFFERENTIATION_BY_HEMIN"
[86] "REACTOME_ANTIGEN_PROCESSING_CROSS_PRESENTATION"
[87] "JIANG_AGING_CEREBRAL_CORTEX_DN"
[88] "ZHU_CMV_24_HR_UP"
[89] "MALONEY_RESPONSE_TO_17AAG_DN"
[90] "ZHU_CMV_ALL_UP"
[91] "MALEK_TREG_C1_6"
[92] "REACTOME_GENERATION_OF_SECOND_MESSENGER_MOLECULES"
[93] "CROMER_METASTASIS_UP"
[94] "REACTOME_INTERFERON_ALPHA_BETA_SIGNALING"
[95] "MARSON_FOXP3_TARGETS_UP"
[96] "T CELL ACTIVATION (I) (M7.1)"
[97] "GEISS_RESPONSE_TO_DSRNA_UP"
[98] "SARRIO_EPITHELIAL_MESENCHYMAL_TRANSITION_DN"
[99] "LINDSTEDT_DENDRITIC_CELL_MATURATION_B"
[100] "BILANGES_RAPAMYCIN_SENSITIVE_VIA_TSC1_AND_TSC2"
[101] "MOSERLE_IFNA_RESPONSE"
[102] "REACTOME_CELL_DEATH_SIGNALLING_VIA_NRAGE_NRIF_AND_NADE"
[103] "KIM_LRRC3B_TARGETS"
[104] "MORI_LARGE_PRE_BII_LYMPHOCYTE_DN"
[105] "DER_IFN_GAMMA_RESPONSE_UP"
[106] "REACTOME_P75_NTR_RECEPTOR_MEDIATED_SIGNALLING"
[107] "BOSCO_INTERFERON_INDUCED_ANTIVIRAL_MODULE"
[108] "MARZEC_IL2_SIGNALING_UP"
[109] "BASSO_B_LYMPHOCYTE_NETWORK"
[110] "DAUER_STAT3_TARGETS_DN"
[111] "MATTIOLI_MGUS_VS_PCL"
[112] "BROCKE_APOPTOSIS_REVERSED_BY_IL6"
[113] "GALINDO_IMMUNE_RESPONSE_TO_ENTEROTOXIN"
[114] "HUANG_DASATINIB_RESISTANCE_UP"
[115] "SAKAI_CHRONIC_HEPATITIS_VS_LIVER_CANCER_UP"
[116] "MHC-TLR7-TLR8 CLUSTER (M146)"
[117] "REACTOME_SIGNALING_BY_TGF_BETA_RECEPTOR_COMPLEX"
[118] "MORI_MATURE_B_LYMPHOCYTE_UP"
[119] "GROSS_HYPOXIA_VIA_HIF1A_UP"
[120] "REACTOME_SIGNALING_BY_THE_B_CELL_RECEPTOR_BCR"

[121] "REACTOME_HOST_INTERACTIONS_OF_HIV_FACTORS"

[122] "SHAFFER_IRF4_TARGETS_IN_MYELOMA_VS_MATURE_B_LYMPHOCYTE"

[123] "BROWN_MYELOID_CELL_DEVELOPMENT_DN"

[124] "THEILGAARD_NEUTROPHIL_AT_SKIN_WOUND_UP"

[125] "HEDENFALK_BREAST_CANCER_BRCA1_VS_BRCA2"

[126] "REACTOME_TRANSLOCATION_OF_ZAP_70_TO_IMMUNOLOGICAL_SYNAPSE"

[127] "ZHOU_TNF_SIGNALING_30MIN"

[128] "T CELL ACTIVATION AND SIGNALING (M5.1)"

[129] "PID_PDGFRBPATHWAY"

[130] "VILIMAS_NOTCH1_TARGETS_UP"

[131] "HOSHIDA_LIVER_CANCER_SUBCLASS_S2"

[132] "HUMMERICH_SKIN_CANCER_PROGRESSION_UP"

[133] "AKL_HTLV1_INFECTION_DN"

[134] "REACTOME_DOWNSTREAM_TCR_SIGNALING"

[135] "MALEK_TREG_C2_5"

[136] "SAKAI_TUMOR_INFILTRATING_MONOCYTES_DN"

[137] "PENG_LEUCINE_DEPRIVATION_UP"

[138] "UEDA_PERIFERAL_CLOCK"

[139] "NING_CHRONIC_OBSTRUCTIVE_PULMONARY_DISEASE_DN"

[140] "WIERENGA_STAT5A_TARGETS_GROUP2"

[141] "DIRMEIER_LMP1_RESPONSE_LATE_UP"

[142] "LABBE_WNT3A_TARGETS_UP"

[143] "HOFFMANN_PRE_BI_TO_LARGE_PRE_BII_LYMPHOCYTE_DN"

[144] "SESTO_RESPONSE_TO_UV_C0"

[145] "TARTE_PLASMA_CELL_VS_B_LYMPHOCYTE_DN"

[146] "TAKAO_RESPONSE_TO_UVB_RADIATION_DN"

[147] "FLECHNER_PBL_KIDNEY_TRANSPLANT_REJECTED_VS_OK_DN"

[148] "PID_CD8TCRPATHWAY"

[149] "HARRIS_HYPOXIA"

[150] "NADLER_OBESITY_UP"

[151] "DIRMEIER_LMP1_RESPONSE_LATE_DN"

[152] "OUELLET_OVARIAN_CANCER_INVASIVE_VS_LMP_UP"

[153] "T CELL DIFFERENTIATION (TH2) (M19)"

[154] "KEGG_HEMATOPOIETIC_CELL_LINEAGE"

[155] "KEGG_NATURAL_KILLER_CELL_MEDIATED_CYTOTOXICITY"

[156] "REACTOME_REGULATION_OF_MRNA_STABILITY_BY_PROTEINS_THAT_BIND_AU_RICH_ELEMENTS"

[157] "NING_CHRONIC_OBSTRUCTIVE_PULMONARY_DISEASE_UP"

[158] "T CELL SURFACE SIGNATURE (S0)"

[159] "REACTOME_TGF_BETA_RECEPTOR_SIGNALING_ACTIVATES_SMADS"

[160] "T CELL ACTIVATION (II) (M7.3)"

[161] "WINTER_HYPOXIA_DN"

[162] "REACTOME_SIGNALING_BY_ILS"

[163] "KEGG_T_CELL_RECEPTOR_SIGNALING_PATHWAY"

[164] "REACTOME_IMMUNOREGULATORY_INTERACTIONS_BETWEEN_A_LYMPHOID_AND_A_NON_LYMPHOID_CELL"

[165] "TARTE_PLASMA_CELL_VS_B_LYMPHOCYTE_UP"

[166] "REACTOME_DOWNSTREAM_SIGNALING_EVENTS_OF_B_CELL_RECEPTOR_BCR"

[167] "PID_IL2_1PATHWAY"

[168] "MARTIN_INTERACT_WITH_HDAC"

[169] "REACTOME_PROCESSING_OF_CAPPED_INTRON_CONTAINING_PRE_MRNA"

[170] "KEGG_AUTOIMMUNE_THYROID_DISEASE"

[171] "BACOLOD_RESISTANCE_TO_ALKYLATING_AGENTS_DN"

[172] "ST_INTEGRIN_SIGNALING_PATHWAY"

[173] "ZHANG_PROLIFERATING_VS_QUIESCENT"

[174] "GUTIERREZ_CHRONIC_LYMPHOCYTIC_LEUKEMIA_DN"

[175] "FLECHNER_PBL_KIDNEY_TRANSPLANT_REJECTED_VS_OK_UP"

[176] "LUI_THYROID_CANCER_PAX8_PPARG_DN"

[177] "GRANDVAUX_IFN_RESPONSE_NOT_VIA_IRF3"

[178] "CHIBA_RESPONSE_TO_TSA"

[179] "YAO_TEMPORAL_RESPONSE_TO_PROGESTERONE_CLUSTER_14"

[180] "BROWNE_INTERFERON_RESPONSIVE_GENES"

[181] "TSAI_RESPONSE_TO_IONIZING_RADIATION"

[182] "MARTINEZ_RESPONSE_TO_TRABECTEDIN_UP"

[183] "SINGH_NFE2L2_TARGETS"

[184] "JIANG_AGING_HYPOTHALAMUS_UP"

[185] "ZHENG_FOXP3_TARGETS_IN_T_LYMPHOCYTE_DN"

[186] "PID_PI3KCIPATHWAY"

[187] "SWEET_KRAS_ONCOGENIC_SIGNATURE"

[188] "ZHAN_MULTIPLE_MYELOMA_LB_DN"

[189] "BIOCARTA_TCRA_PATHWAY"

[190] "LU_IL4_SIGNALING"

[191] "SHIPP_DLBCL_VS_FOLLICULAR_LYMPHOMA_DN"

[192] "FERRANDO_T_ALL_WITH_MLL_ENL_FUSION_DN"

[193] "MENSSEN_MYC_TARGETS"

[194] "REACTOME_NUCLEOTIDE_BINDING_DOMAIN_LEUCINE_RICH_REPEAT_CONTAINING_RECEPTOR_NLR_SIGNALING_PATHWAYS"

[195] "APPIERTO_RESPONSE_TO_FENRETINIDE_DN"

[196] "REACTOME_MHC_CLASS_II_ANTIGEN_PRESENTATION"

[197] "PECE_MAMMARY_STEM_CELL_UP"

[198] "BRACHAT_RESPONSE_TO_CAMPTOTHECIN_DN"

[199] "MALEK_TREG_C2_7"

[200] "DIAZ_CHRONIC_MEYLOGENOUS_LEUKEMIA_DN"

[201] "DUNNE_TARGETS_OF_AML1_MTG8_FUSION_UP"

[202] "PID_HIF1_TFPATHWAY"

[203] "LUI_TARGETS_OF_PAX8_PPARG_FUSION"

[204] "LI_LUNG_CANCER"

[205] "REACTOME_MRNA_SPLICING"
[206] "REACTOME_NRAGE_SIGNALS_DEATH_THROUGH_JNK"
[207] "LUI_THYROID_CANCER_CLUSTER_3"
[208] "BLALOCK_ALZHEIMERS_DISEASE_INCIPIENT_DN"
[209] "KEGG_ASTHMA"
[210] "MODY_HIPPOCAMPUS_PRENATAL"
[211] "KEGG_SPLICEOSOME"
[212] "PID_ERBB1_DOWNSTREAM_PATHWAY"
[213] "PID_RAC1_REG_PATHWAY"
[214] "HAHTOLA_SEZARY_SYNDROM_UP"
[215] "TRACEY_RESISTANCE_TO_IFNA2_DN"
[216] "PUJANA_BRCA_CENTERED_NETWORK"
[217] "MYELOID, DENDRITIC CELL ACTIVATION VIA NFKB (II) (M43.1)"
[218] "SESTO_RESPONSE_TO_UV_C2"
[219] "PID_RAC1_PATHWAY"
[220] "BILBAN_B_CLL_LPL_DN"
[221] "ENRICHED IN ANTIGEN PRESENTATION (II) (M95.0)"
[222] "NGUYEN_NOTCH1_TARGETS_DN"
[223] "MOREIRA_RESPONSE_TO_TSA_UP"
[224] "NATSUME_RESPONSE_TO_INTERFERON_BETA_DN"
[225] "KLEIN_PRIMARY_EFFUSION_LYMPHOMA_DN"
[226] "WIERENGA_PML_INTERACTOME"
[227] "RAHMAN_TP53_TARGETS_PHOSPHORYLATED"
[228] "EINAV_INTERFERON_SIGNATURE_IN_CANCER"
[229] "REACTOME_PD1_SIGNALING"
[230] "MALEK_TREG_C2_13"
[231] "SANA_RESPONSE_TO_IFNG_DN"
[232] "HINATA_NFKB_TARGETS_KERATINOCYTE_UP"
[233] "ICHIBA_GRAFT_VERSUS_HOST_DISEASE_35D_UP"
[234] "TOOKER_GEMCITABINE_RESISTANCE_UP"
[235] "HOLLEMAN_ASPARAGINASE_RESISTANCE_B_ALL_UP"
[236] "YAMASHITA_LIVER_CANCER_WITH_EPCAM_UP"
[237] "ENRICHED IN ANTIGEN PRESENTATION (III) (M95.1)"
[238] "KAAB_FAILED_HEART_ATRIUM_DN"
[239] "PROVENZANI_METASTASIS_DN"
[240] "REACTOME_VIF_MEDIATED_DEGRADATION_OF_APOBEC3G"
[241] "CELL ADHESION (GO) (M117)"
[242] "REACTOME_REGULATION_OF_ORNITHINE_DECARBOXYLASE_ODC"
[243] "PID_TCPTP_PATHWAY"
[244] "PID_AMB2_NEUTROPHILS_PATHWAY"
[245] "REACTOME_CYTOSOLIC_TRNA_AMINOACYLATION"
[246] "TOOKER_GEMCITABINE_RESISTANCE_DN"

[247] "REACTOME_MRNA_PROCESSING"
[248] "WATANABE_ULCERATIVE_COLITIS_WITH_CANCER_DN"
[249] "PID_IL12_2PATHWAY"
[250] "IIZUKA_LIVER_CANCER_PROGRESSION_L0_L1_DN"
[251] "ABRAMSON_INTERACT_WITH_AIRE"
[252] "RHODES_CANCER_META_SIGNATURE"
[253] "STAMBOLSKY_TARGETS_OF_MUTATED_TP53_DN"
[254] "BRACHAT_RESPONSE_TO_CISPLATIN"
[255] "LINDGREN_BLADDER_CANCER_CLUSTER_1_UP"
[256] "SASSON_RESPONSE_TO_FORSKOLIN_DN"
[257] "REACTOME_DESTABILIZATION_OF_MRNA_BY_AUF1_HNRNP_D0"
[258] "IL2, IL7, TCR NETWORK (M65)"
[259] "CAFFAREL_RESPONSE_TO_THC_24HR_5_UP"
[260] "KEGG_ALZHEIMERS_DISEASE"
[261] "BIOCARTA_PDGF_PATHWAY"
[262] "WOOD_EBV_EBNA1_TARGETS_UP"
[263] "CHIARADONNA_NEOPLASTIC_TRANSFORMATION_KRAS_UP"
[264] "MULLIGHAN_NPM1_MUTATED_SIGNATURE_2_DN"
[265] "HU_GENOTOXIN_ACTION_DIRECT_VS_INDIRECT_24HR"
[266] "BURTON_ADIPOGENESIS_PEAK_AT_8HR"
[267] "KEGG_PATHOGENIC_ESCHERICHIA_COLI_INFECTION"
[268] "TIEN_INTESTINE_PROBIOTICS_2HR_UP"
[269] "CHAUHAN_RESPONSE_TO_METHOXYESTRADIOL_UP"
[270] "BIOCARTA_EGF_PATHWAY"
[271] "REACTOME_P53_DEPENDENT_G1_DNA_DAMAGE_RESPONSE"
[272] "PID_TGFBRPATHWAY"
[273] "REACTOME_GPVI_MEDIATED_ACTIVATION_CASCADE"
[274] "BIOCARTA_PROTEASOME_PATHWAY"
[275] "WUNDER_INFLAMMATORY_RESPONSE_AND_CHOLESTEROL_UP"
[276] "NGO_MALIGNANT_GLIOMA_1P_LOH"
[277] "KEGG_FOCAL_ADHESION"
[278] "GAURNIER_PSMD4_TARGETS"
[279] "YAN_ESCAPE_FROM_ANOIKIS"
[280] "PID_BCR_5PATHWAY"
[281] "KEGG_PHOSPHATIDYLINOSITOL_SIGNALING_SYSTEM"
[282] "KEGG_GLIOMA"
[283] "REACTOME_ACTIVATION_OF_NF_KAPPAB_IN_B_CELLS"
[284] "KEGG_LEISHMANIA_INFECTION"
[285] "REACTOME_AUTODEGRADATION_OF_THE_E3_UBIQUITIN_LIGASE_COP1"
[286] "BARIS_THYROID_CANCER_DN"
[287] "YAO_TEMPORAL_RESPONSE_TO_PROGESTERONE_CLUSTER_8"
[288] "TREG_SUGIMOTO"

[289] "REACTOME_REGULATION_OF_APOPTOSIS"

[290] "PID_IL12_STAT4PATHWAY"

[291] "BYSTRYKH_HEMATOPOIESIS_STEM_CELL_QTL_CIS"

[292] "MYELOID, DENDRITIC CELL ACTIVATION VIA NFKB (I) (M43.0)"

[293] "KOKKINAKIS_METHIONINE_DEPRIVATION_96HR_DN"

[294] "CD4 T CELL SURFACE SIGNATURE TH2-STIMULATED (S7)"

[295] "MULLIGHAN_NPM1_SIGNATURE_3_DN"

[296] "HILLION_HMGA1_TARGETS"

[297] "REACTOME_COSTIMULATION_BY_THE_CD28_FAMILY"

[298] "KIM_GERMINAL_CENTER_T_HELPER_UP"

[299] "BILANGES_SERUM_RESPONSE_TRANSLATION"

[300] "ST_T_CELL_SIGNAL_TRANSDUCTION"

[301] "REACTOME_AUTODEGRADATION_OF_CDH1_BY_CDH1_APC_C"

[302] "DASU_IL6_SIGNALING_UP"

[303] "XU_RESPONSE_TO_TRETINOIN_AND_NSC682994_UP"

[304] "BIOCARTA_NDKDYNAMIN_PATHWAY"

[305] "WEIGEL_OXIDATIVE_STRESS_BY_TBH_AND_H2O2"

[306] "LIU_CMYB_TARGETS_UP"

[307] "JUBAN_TARGETS_OF_SPI1_AND_FLI1_DN"

[308] "REACTOME_CYCLIN_E_ASSOCIATED_EVENTS_DURING_G1_S_TRANSITION_"

[309] "KEGG_GLUTATHIONE_METABOLISM"

[310] "PID_CMYB_PATHWAY"

[311] "JAIN_NFKB_SIGNALING"

[312] "PID_SHP2_PATHWAY"

[313] "BURTON_ADIPOGENESIS_2"

[314] "SMIRNOV_RESPONSE_TO_IR_6HR_DN"

[315] "RODRIGUES_DCC_TARGETS_DN"

[316] "REACTOME_REGULATION_OF_MITOTIC_CELL_CYCLE"

[317] "BOHN_PRIMARY_IMMUNODEFICIENCY_SYNDROM_UP"

[318] "BOWIE_RESPONSE_TO_TAMOXIFEN"

[319] "KEGG_PROSTATE_CANCER"

[320] "BYSTROEM_CORRELATED_WITH_IL5_DN"

[321] "KEGG_B_CELL_RECEPTOR_SIGNALING_PATHWAY"

[322] "LENAOUR_DENDRITIC_CELL_MATURATION_UP"

[323] "PID_IL8CXCR2_PATHWAY"

[324] "PID_EPOPATHWAY"

[325] "REACTOME_CDK_MEDIATED_PHOSPHORYLATION_AND_REMOVAL_OF_CDC6"

[326] "REACTOME_THE_ROLE_OF_NEF_IN_HIV1_REPLICATION_AND_DISEASE_PATHOGENESIS"

[327] "MMS_MOUSE_LYMPH_HIGH_4HRS_UP"

[328] "JOHANSSON_GLIOMAGENESIS_BY_PDGFB_UP"

[329] "PID_IL2_STAT5PATHWAY"

[330] "GAVIN_IL2_RESPONSIVE_FOXP3_TARGETS_UP"

[331] "PID_CD8TCRDOWNSTREAMPATHWAY"

[332] "BIOCARTA_RHO_PATHWAY"

[333] "KEGG_SYSTEMIC_LUPUS_ERYTHEMATOSUS"

[334] "REACTOME_ORC1_REMOVAL_FROM_CHROMATIN"

[335] "KEGG_ADHERENS_JUNCTION"

[336] "CORO1A-DEF6 NETWORK (I) (M32.2)"

[337] "SANA_TNF_SIGNALING_UP"

[338] "HOLLEMAN_VINCRISTINE_RESISTANCE_ALL_DN"

[339] "ZAMORA_NOS2_TARGETS_UP"

[340] "REACTOME_CDT1_ASSOCIATION_WITH_THE_CDC6_ORC_ORIGIN_COMPLEX"

[341] "JIANG_VHL_TARGETS"

[342] "REACTOME_P53_INDEPENDENT_G1_S_DNA_DAMAGE_CHECKPOINT"

[343] "POMEROY_MEDULLOBLASTOMA_PROGNOSIS_UP"

[344] "RUIZ_TNC_TARGETS_UP"

[345] "PID_IL2_PI3KPATHWAY"

[346] "LINDSTEDT_DENDRITIC_CELL_MATURATION_C"

[347] "KEGG_APOPTOSIS"

[348] "T CELL SIGNALING AND COSTIMULATION (M44)"

[349] "FONTAINE_FOLLICULAR_THYROID_ADENOMA_DN"

[350] "WATANABE_RECTAL_CANCER_RADIOTHERAPY_RESPONSIVE_UP"

[351] "SIMBULAN_UV_RESPONSE_IMMORTALIZED_DN"

[352] "SASSON_RESPONSE_TO_GONADOTROPHINS_UP"

[353] "PID_TNFPATHWAY"

[354] "FALVELLA_SMOKERS_WITH_LUNG_CANCER"

[355] "BOWIE_RESPONSE_TO_EXTRACELLULAR_MATRIX"

[356] "SHAFFER_IRF4_TARGETS_IN_ACTIVATED_DENDRITIC_CELL"

[357] "REACTOME_CELL_CYCLE_CHECKPOINTS"

[358] "REACTOME_SCFSKP2_MEDIATED_DEGRADATION_OF_P27_P21"

[359] "ACOSTA_PROLIFERATION_INDEPENDENT_MYC_TARGETS_DN"

[360] "BIOCARTA_THELPER_PATHWAY"

[361] "SHIPP_DLBCL_VS_FOLLICULAR_LYMPHOMA_UP"

[362] "IRITANI_MAD1_TARGETS_DN"

[363] "REACTOME_TCA_CYCLE_AND_RESPIRATORY_ELECTRON_TRANSPORT"

[364] "BIOCARTA_CSK_PATHWAY"

[365] "BHATTACHARYA_EMBRYONIC_STEM_CELL"

[366] "PASQUALUCCI_LYMPHOMA_BY_GC_STAGE_DN"

[367] "KEGG_NEUROTROPHIN_SIGNALING_PATHWAY"

[368] "ELVIDGE_HYPOXIA_DN"

[369] "KEGG_ACUTE_MYELOID_LEUKEMIA"

[370] "REACTOME_S_PHASE"

[371] "SHAFFER_IRF4_TARGETS_IN_ACTIVATED_B_LYMPHOCYTE"

[372] "REACTOME_SIGNALING_BY_WNT"

[373] "GRADE_METASTASIS_DN"

[374] "BIOCARTA_ERK_PATHWAY"

[375] "DAZARD_UV_RESPONSE_CLUSTER_G1"

[376] "REACTOME_APC_C_CDC20_MEDIATED_DEGRADATION_OF_MITOTIC_PROTEINS"

[377] "KEGG_CELL_ADHESION_MOLECULES_CAMS"

[378] "GRAHAM_CML_DIVIDING_VS_NORMAL_QUIESCENT_DN"

[379] "PID_ILK_PATHWAY"

[380] "ND001_ADENO_SPLEENVSTREG_C2_2"

[381] "REACTOME_APOPTOSIS"

[382] "WATANABE_RECTAL_CANCER_RADIOTHERAPY_RESPONSIVE_DN"

[383] "PURBEY_TARGETS_OF_CTBP1_AND_SATB1_UP"

[384] "PID_FOXOPATHWAY"

[385] "MOOTHA_VOXPHOS"

[386] "KEGG_ENDOCYTOSIS"

[387] "MIKKELSEN_MCV6_LCP_WITH_H3K4ME3"

[388] "TAKEDA_TARGETS_OF_NUP98_HOXA9_FUSION_3D_UP"

[389] "PID_PI3KPLCTRKPATHWAY"

[390] "PID_TRKRPATHWAY"

[391] "AMUNDSON_POOR_SURVIVAL_AFTER_GAMMA_RADIATION_2G"

[392] "ZHANG_RESPONSE_TO_CANTHARIDIN_DN"

[393] "MCCLUNG_COCAINE_REWARD_5D"

[394] "MULLIGHAN_NPM1_MUTATED_SIGNATURE_1_DN"

[395] "MODY_HIPPOCAMPUS_POSTNATAL"

[396] "MORI_IMMATURE_B_LYMPHOCYTE_UP"

[397] "KAYO_AGING_MUSCLE_DN"

[398] "PLATELET ACTIVATION (I) (M32.0)"

[399] "ZHANG_INTERFERON_RESPONSE"

[400] "SCHLINGEMANN_SKIN_CARCINOGENESIS_TPA_UP"