Recherche de Pharmacogènes associés aux effets indésirables des Inhibiteurs de la Calcineurine : développement d'approches bio-informatiques adaptées aux petits échantillons

Claire-Cécile Barrot

Thèse de doctorat

Université de Limoges
École Doctorale n°615 Sciences Biologique et Santé – SBS
Laboratoire IPPRITT – INSERM UMR1248

Thèse pour obtenir le grade de
Docteur de l'Université de Limoges

Sciences de la Vie et de la Santé / Génétique

Présentée et soutenue par
Claire-Cécile Barrot

Le 10 décembre 2019

Recherche de Pharmacogènes associés aux effets indésirables des Inhibiteurs de la Calcineurine

Développement d'approches bio-informatiques adaptées aux petits échantillons

Thèse dirigée par Pr. Nicolas PICARD et Dr. Jean-Baptiste Woillard

JURY :

Rapporteurs
Céline Verstuyft, Professeur, CHU Bicêtre
Nicolas Sevenet, Professeur, Université de Bordeaux

Examinateurs
Anne-Sophie Lia, MCU-PH, CHU Limoges
Fréderic Libert, MCU-PH, CHU Clermont-Ferrand

Remerciements

Merci à toute l'équipe IPPRIT,

à Nicolas et Jean-Baptiste pour m'avoir encadrée,

à Hélène pour toute son aide et ses explications,

et à Benjamin pour avoir considérablement étendu mon vocabulaire.

Merci à ma famille pour leur soutient.

Droits d'auteurs

Sommaire

Index des figures

Index des tableaux

Introduction

Le terme pharmacogénétique a été inventé dans les années 1950s pour faire références à la variation phénotypique du métabolisme et la réponse à certains médicaments[1]. Il s'agissait alors d'un phénomène couramment observé et documenté, mais ce n'est qu'a partir des années 1980s que ses origines génétiques ont commencées a être comprises[2]. Le terme fait depuis référence à l'étude des variabilités interindividuelles de séquences ADN en relation à une variabilité de réponse à certains médicaments. Ce domaine de la pharmacologie a notamment joué un rôle dans le développement de la médecine personnalisée.

L'évolution des techniques de séquençages depuis les années 2000 a conduit à l'émergence de la pharmacogénomique, un nouveau domaine de la pharmacologie lié à la pharmacogénétique.

La pharmacogénomique (PGx) est l'étude de la relation entre les variations du génome au sens large (ADN, ARN) et les médicaments et leurs réponses. Cette variabilité peut avoir lieu au niveau génomique (variations des séquences de nucléotides), transcriptomique (variations de l'expression des gènes), ou épigénomique (modifications réversibles de l'ADN). Les effets possibles sur la réponse au médicament peuvent être de deux type :

- pharmacocinétique : modification de l'activité des transporteurs ou des enzymes de métabolisme qui peuvent avoir des répercussions sur la concentration du médicament.
- pharmacodynamique : variations d'activités de la cible des principes actifs, avec des répercussions sur l'efficacité ou les effets indésirables du principe actif[3].

Concernant l'analyse bioinformatiques des données PGx l'efficacité des méthodes de séquençage de nouvelle génération et la réduction de leur coût ont conduit à l'utilisation de nouvelles méthodes, basées sur le concept de l'analyse de *big data*, appliqué aux données génomiques.

Le pipeline d'analyse des *big data* est standardisé, comprenant la collecte de données, leur traitement, l'analyse puis l'interprétation clinique. Dans le cadre d'études génomiques, les données étudiées peuvent être :

- génomiques ou exomiques, concernant l'étude de variations dans les séquences de nucléotides
- transcriptomiques, concernant l'étude des variation de l'expression des gènes
- épigénomiques, concernant l'étude des modification réversibles des séquences d'ADN

La PGx est une sous-discipline de la génomique et en utilise de ce fait les outils et les méthodes d'analyse, avec cependant quelques différences liées à la focalisation de la discipline sur la réponse aux médicaments.

Les scores de pathogénicité des variants de séquences d'ADN sont par exemple très utilisés en génomique, mais sont peu pertinent en PGx. De même, les études génomiques considèrent souvent qu'un variant est plus susceptible d'être d'intérêt lorsqu'il est situé sur une zone très conservée du génome, ou bien s'il est particulièrement rare dans la population générale. Cela est inexact en PGx puisque les pharmacogènes présentent souvent de multiples versions fonctionnelles, ainsi les outils tels que les scores de conservation ou les Minor Allele Frequency (MAF) y sont peu utilisés.

Les études PGx récentes se concentrent surtout sur les analyses sans *a priori*, basées sur des principes statistiques ou mathématiques. Les études d'association GWAS (Genome Wise Association Study) couramment utilisées en génomiques sont utiles en PGx, puisqu'elles permettent de lier des variants à un phénotype par exemple, ou bien des variants les uns aux autres. Elles restent malgré tout comparativement peu utilisées. En revanche, les modèles prédictifs à base de *machine learning*, permettant notamment de prédire les concentrations circulantes d'un médicament ou d'éventuels effets secondaires, sont de plus en plus utilisés.

Cependant, parmi toutes les méthodes disponibles pour les analyses génomiques, qu'elles soient statistiques, mathématiques ou informatiques, une partie seulement est couramment utilisée en PGx à ce jour.

Dans le cadre de cette thèse, une revue des analyses big data utilisées en PGx a été effectuée, avec une attention particulière sur les aspects bioinformatiques.

Cette revue a donné lieu à la publication suivante :

C.-C. Barrot, J.-B. Woillard, N. Picard, Big data in pharmacogenomics: current applications, perspectives and pitfalls, Pharmacogenomics. 20 (2019) 609–620. doi:10.2217/pgs-2018-0184.

Bien qu'ayant prouvées leur efficacité, les analyses « big data » sont limitées par la nécessité d'avoir de grandes cohortes de patients ainsi que de nombreuses données pour chaque patient. Cela nécessite pour chaque étude un coût financier important, limitant ainsi la réalisation de ces études.

Une solution à ce problème est d'étudier un panel choisi de gènes sur une grande cohorte. Traditionnellement, le panel de gènes à analyser pour une étude donnée est déterminé par les connaissances préalables sur le sujet : famille élargie d'un gène connu pour être d'intérêt, gènes liés à une voie métabolique spécifique, gènes mis en avant dans des études précédentes, *etc*. En PGx, il peut simplement s'agir de pharmacogènes connus.

En effet, la PGx porte un intérêt particulier aux gènes codants pour des protéines régulant les concentrations circulantes des médicaments dans l'organisme (absorption, distribution, métabolisation, élimination) ainsi que les gènes des protéines cibles des médicaments ou des protéines intervenant dans les systèmes de signalisation de leurs effets.

Parmi les plus importants pharmacogènes se trouvent les enzymes de métabolisation des xénobiotiques (EMX) comme la super-famille des Cytochromes P450 (CYP) ou celle des UDP-glucuronosyltransférases, ainsi que les récepteurs nucléaires partenaires des EMX (*ex.* POR, le partenaire des CYP par exemple), et les transporteurs des superfamilles SLC (SoLute Carriers) ou ABC (ATP-Binding Casette).

L'approche traditionnelle de sélection de panels de gènes reste cependant limitée par les connaissances préalables sur le sujet étudié. Dans le cadre de cette thèse, j'ai cherché à déterminer s'il était possible d'utiliser des méthodes sans a priori pour sélectionner de tels gènes d'intérêt. En d'autres termes, est-il possible de développer des approches bioinformatiques qui, à partir de petites cohortes, permettent la sélection des pharmacogènes d'intérêt ?

J'ai développé de telles approches mises en pratique dans des études PGx de génomes, transcriptomes et épigénomes liés aux immunosuppresseurs par inhibition de calcineurine.

Inhibiteurs de la calcineurine

Les immunosuppresseurs sont des médicaments ayant pour but d'inhiber l'activation du système immunitaire, pour traiter les maladies auto-immunes ou bien pour prévenir le rejet de greffe après une transplantation. Ils peuvent cibler directement les lymphocytes T (LT) ou ou B, les cytokines ou les chimiokines, ou encore des cibles cellulaires multiples[4]. Les immunosuppresseurs les plus utilisés sont répartis en quatre catégories principales :

- Les glucocorticoïdes, qui diminuent l'immunité cellulaire en inhibant les cytokines pro-inflammatoires (IL-1 à IL-8 et TNF-γ) : leur effet anti-rejet étant non-spécifique, ils

sont surtout utilisés pour leur effet anti-inflammatoire.

- Les cytostatiques, qui inhibent la division cellulaire, les rendant particulièrement utiles sur les lymphocytes B et T qui sont des cellules à division rapide.
- Les anticorps, qu'ils soient poly-clonaux ou monoclonaux, notamment contre les récepteurs IL-2 et des cellules T, sont utilisés en début de greffe ou pour la prévention des rejets aigus.
- Enfin, les inhibiteurs de la calcineurine (ICN), actuellement les médicaments d'entretien de la greffe de référence après transplantation solide d'organe[56].

Dans le cadre de cette thèse, nous nous sommes concentrés sur les immunosuppresseurs agissant sur la calcineurine.

La calcineurine, ou protéine phosphatase 2B (PP2B), est une protéine phosphatase dépendante du calcium au sein des LT. Elle est composée d'une sous-unité catalytique d'activité phosphatase (calcineurine A) et d'une sous-unité régulatrice maintenant l'enzyme inactive (calcineurine B).

L'activation des récepteurs des cellules T (TCR, T cell receptor) provoque une libération de calcium à l'intérieur des LT depuis les organites intracellulaires, et une entrée de calcium extra-cellulaire via des canaux membranaires spécifiques.

L'augmentation du calcium active la calcineurine ainsi que la calmoduline, qui stimule la calcineurine en se fixant sur la calcineurine B, inhibant ainsi son effet inhibiteur permettant l'activation de l'activité phosphatase de la calcineurine A.

Le complexe protéique calcineurine/calmoduline active le facteur nucléaire des lymphocytes T activés (NFAT, Nuclear Factor of activated T cells) par déphosphorylation, ce qui permet sa translocation nucléaire et l'activation de la transcription de plusieurs gènes impliqués dans l'activation lymphocytaire :

- L'interleukine 2 (IL-2) et les sous unités de son récepteur (α, β et γ)[4]. Cette cytokine sécrétée par les lymphocytes T entraîne la stimulation de la prolifération lymphocytaire (LT compris) et aide les LT auxiliaires (LTCD4) à reconnaître les cellules étrangères, jouant ainsi un rôle clé dans l'organisation de la réponse immunitaire contre un greffon.
- L'interféron gamma (IFNγ) : une cytokine principalement synthétisées par les LT, qui parmi ses fonctions aide à la reconnaissance des cellules étrangères et à la coordination de la réponse immunitaire[789].

Le facteur stimulant les colonies de granulocytes et de macrophages (GM-CSF, Granulocyte-

macrophage colony-stimulating factor) : une protéine agissant comme facteur de croissance des leucocytes[10], qui agit sur diverses cellules du système immunitaire[11].

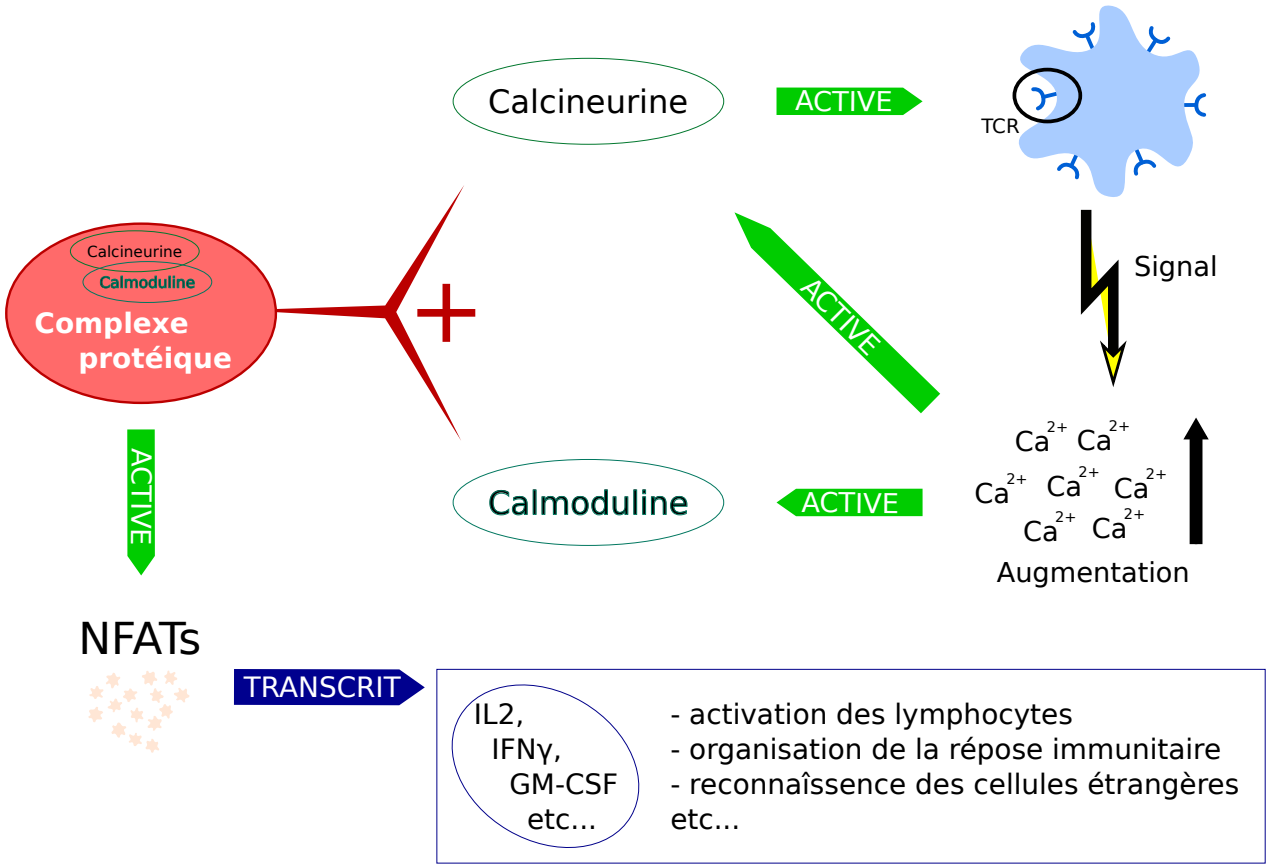

Figure 1: Voie de la Calcineurine. La calcineurine active les récepteurs TCR, provoquant une augmentation du calcium, ce qui active à la fois la Calcineurine et la Calmoduline. Le complexe protéique Calcineurine/Calmoduline active les NFATs, provoquant la transcriptions de divers gènes liés au système immunitaire.

Deux molécules agissent comme ICN en se fixant à une protéine récepteur cytoplasmique : la ciclosporine qui se fixe à la cyclophiline, et le tacrolimus qui se fixe à la FKBP12. Dans les deux cas, le complexe protéique se lie à la calcineurine et l'inhibe, ce qui bloque la translocation nucléaire de NFAT[12], et donc la prolifération des LT.

Ciclosporine et tacrolimus sont utilisés en première intention pour prévenir et traiter les rejets de greffes. Le tacrolimus est particulièrement indiqué après des greffes rénales, hépatiques et cardiaques, alors que la ciclosporine est indiquée après tout type de transplantations d'organes solides, greffes de moelle osseuse, ou greffes de cellules souches. Dans le cadre de thérapies immunosuppressives en transplantation, ils sont usuellement associés à des glucocorticoïdes immédiatement après la greffe et à un anti-métabolite (MMF ou azathioprine) ou à un inhibiteur de mTOR (évérolimus ou sirolimus). Les deux médicaments sont également utilisés dans d'autres indications telles que les maladies auto-immunes ou inflammatoires chroniques[13].

Les ICN présentent des effets indésirables parmi lesquels une néphrotoxicité ou l'apparition de cancers liés à la sur-immunosuppression, pouvant être responsable de la perte du greffon voire du décès du patient[14]. D'autres effets indésirables communs sont moins graves mais pouvant nécessiter une hospitalisation ou une diminution de la dose (avec alors un risque d'inefficacité). Leurs incidences sont variables comptes tenus des facteurs de risques spécifiques associés à chaque pathologie, ainsi que de facteurs génétiques (ethnie, antécédents familiaux, etc), mais également de l'ICN prescrit . La toxicité neurologique est plus courante lors de la prise de tacrolimus que de ciclosporine par exemple, tout comme les anomalies glucidiques. A l'inverse, les maladies cardiovasculaires et des infections bactériennes, fongiques ou virales sont plus courantes en cas de prise de ciclosporine (Tableau 1).

Tableau 1 : effets indésirables des ICN, et leur fréquences[13]

Effet indésirable	Ciclosporine	Tacrolimus
Met potentiellement en jeu le pronostic vital		
Toxicité rénale	5 – 21 %	10 – 35 %
Cancers et syndromes lymphoprolifératifs (Tumeurs cutanées, syndromes lymphoprolifératifs, sarcome de kaposi, etc)	2 %	2 %
Hospitalisation potentiellement nécessaire		
Toxicité neurologique	12 – 26 %	20 – 35 %
Trouble glucidique (Hyperglycémie, diabète)	7 %	12 - 16 %
Trouble cardiovasculaires (HTA, arythmies, insuffisance coronarienne)	3 – 40 %	1 – 37 %
Infections bactériennes, fongiques et virales	10 – 50 %	18 – 44 %
Hospitalisation exceptionnelle mais potentiel arrêt du traitement (même temporaire)		
Hyperplasie gingivale	6 %	1 %
Désordres digestifs (Constipation, douleurs abdominales, diarrhée, nausée, vomissements)	8 - 35 %	12 - 31 %
Désordres biologiques (Hyperkaliémie, anémie, leucopénie, acidose)	10 – 15 %	10 – 35 %
Hirsutisme, acné	10 %	3 %

L'étude et la prédiction de la réponse médicamenteuse a longtemps été un objectif principal de la recherche pharmacologique. Parmi les facteurs de variabilité, l'influence des polymorphismes génétiques est étudiée dans le cadre de la pharmacogénétique. Dans les dernières décennies, l'effet des médicaments sur les gènes a également commencé à être étudié dans le cadre d'études PGx.

Recherche de pharmacogènes liés aux inhibiteurs de calcineurine

Le premier chapitre de cette thèse porte sur deux études exomiques (utilisant un design cas-témoins) de patients transplantés rénaux, recherchant des gènes potentiellement liés à deux effets indésirables peu fréquents mais cliniquement pertinents : le syndrome lymphoprolifératif post-transplantation, et le diabète d'apparition tardive post-transplantation.

Le deuxième chapitre concerne une étude de la régulation des gènes (transcriptome et exome méthylé) chez des souris traitées par ICN durant laquelle nous avons développé une approche par recherche de valeurs aberrantes.

Enfin, le troisième chapitre présente l'outil informatique que nous avons développé pour détecter les haplotypes connus des pharmacogènes utilisant les bases de données publiques.

Chapitre I. Étude cas-témoins d'exomes

Le syndrome lymphoprolifératif post-transplantation et le diabète d'apparition tardive post-transplantation sont deux pathologies pouvant survenir à la suite de greffes (notamment rénales) pour lesquels la prise d'ICN a été identifiée comme facteur de risque, avec des prévalences dépendantes de facteurs génétiques[1516]. Ces pathologies peuvent entraîner la perte de greffon et mettre en jeu du pronostic vital, il est donc impératif de les diagnostiquer le plus tôt possible.

Nous avons cherché à déterminer les facteurs génétiques favorisants pour les deux pathologies, dans le but ultime de prédire leur risque d'apparition en fonction de la prise ou non d'ICN, ce qui permettrait à terme un meilleur suivi ou éventuellement la prescription de molécules alternatives.

Les cohortes de patients utilisées pour les études étaient volontairement petites (n=16 et n=42). Dans ce contexte, les fréquences attendues de variants possiblement impliqués sont trop faibles pour que leur présence soit différenciée de faux positifs en analysant de tels effectifs. Pour cette raison, nous avons étudié les gènes dans leur globalité plutôt que recherché des variants spécifiques, en introduisant la notion de degré de variation des gènes.

I.1. Projet PTLD

Le syndrome lymphoprolifératif post-transplantation (PTLD, Post-transplant lymphoproliferative disorder) est caractérisé par une prolifération anormale de lymphocytes, entraînant des complications allant d'une hyperplasie bénigne au développement de lymphomes malins. Avec une incidence entre 2 et 20 %[17], il s'agit de la plus grande cause de perte de greffon liée au cancer[18]. Malgré l'introduction du traitement monoclonal rituximab/anti-CD20 augmentant le taux de survie des patients, celui-ci n'est que de 50 à 60 % trois ans après la greffe[15].

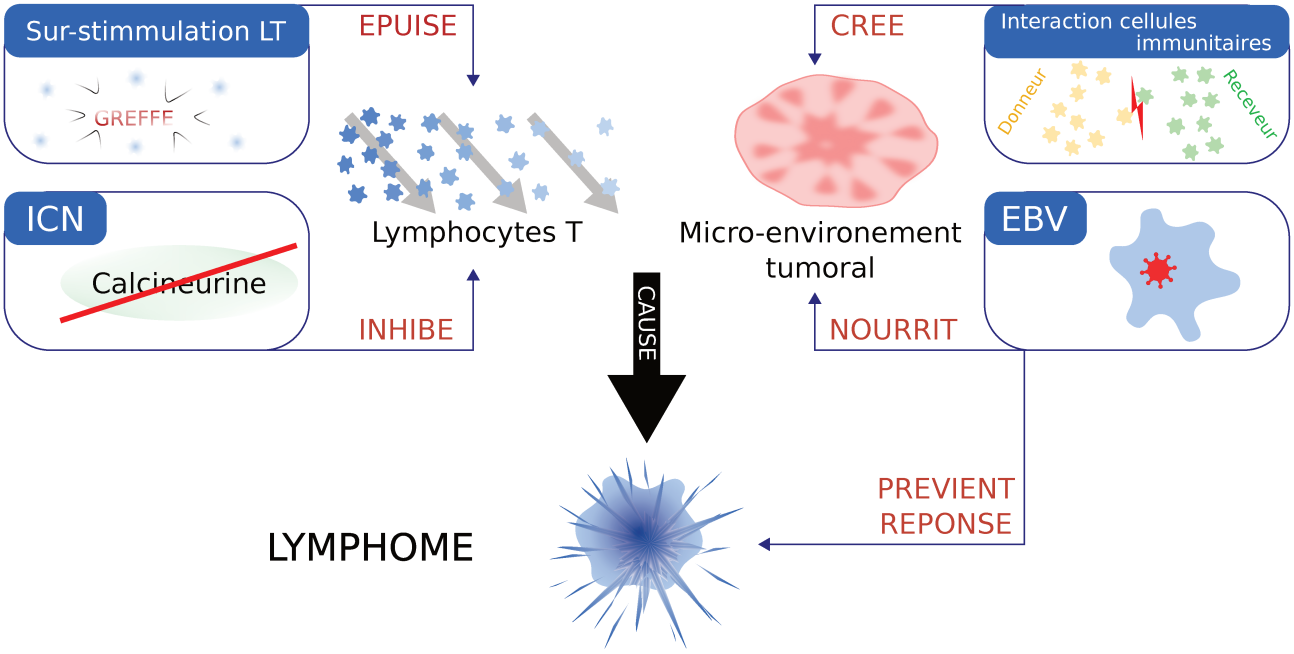

Figure 2: Facteurs causaux de PTLD, Les LT sont épuisés par la sur-stimulation liée à la greffe et inhibés par le traitement ICN. L'interaction des cellules immunitaires du donneur et du receveur crée le TME qui est nourrit par EBV, Ces facteurs entraînent l'apparition de lymphomes, dont EBV prévient la réponse.

Les ICN sont reconnus comme l'une des causes de PTLD[15], puisqu'ils aggravent l'épuisement des LT dû à la stimulation continue du système immunitaire par l'organe greffé. En plus des ICN et de la sur-stimulation des LT, deux autres cofacteurs de PTLD ont été identifiés : les interactions entre les cellules immunitaires du donneur et du receveur, créant un micro-environnement tumoral (TME) contenant tous les constituants cellulaires malins tumoraux[19] et le virus d'Epstein-Barr qui est présent dans 80 % des PTLD[15].

Le virus d'Epstein-Barr (EBV) est l'un des virus de la famille des herpès particulièrement commun, avec une prévalence mondiale au-delà de 90 %[20]. L'infection a généralement lieu dans l'enfance où elle est asymptomatique, ou du moins bénigne. L'EBV est cependant une des causes les plus courantes de mononucléose infectieuse lorsque contracté à l'adolescence ou à l'âge adulte[2122]. Après l'infection initiale, le virus infecte les cellules

épithéliales et pénètre dans le lymphocyte B en circulation où il persiste toute la vie à l'état latent[23].

L'EBV peut cependant être réactivé par un traitement immunosuppresseur tels que les ICN[24]. Dans de tels cas, la réplication lytique d'EBV libère dans le TME des facteurs solubles qui ont pour résultat de (i) promouvoir la croissance de ces lymphocytes B, (ii) inhiber les réponses immunitaires antitumorales, et (iii) stimuler la néo-angiogenèse[25].

Les quatre cofacteurs de PTLD (immunosuppresseurs, TME, sur-stimulation des LT, et EBV) sont inter-dépendants et agissent en synergie, ce qui explique en partie la grande variabilité de l'incidence du syndrome et de ses complications. Une autre cause d'hétérogénéité identifiée est la présence de variations génétiques, notamment sur les gènes de cytokines (interleukines et interféron gamma), sur les des antigènes des leucocytes humains (A et B), ou des gènes liés au facteur de nécrose tumorale[26].

Cependant, les causes génétiques de l'hétérogénéité n'ont pas encore été complètement expliquées.

Notre équipe a constitué une collection biologique adossée à des données cliniques exhaustives pour 101 cas de PTLD (survenus chez des transplantés rénaux) et 101 témoins appariés. Elle a étudié le rôle de variants génétiques fréquents et de faible pénétrance dans la survenue de PTLD. Le projet PTLD intervient dans la continuité de cette étude.

Dans le cadre de cette thèse, l'étude des exomes d'une cohorte plus petite (16 patients) nichée dans PTLD a pour but de rechercher des gènes potentiellement impliquée dans PTLD.

A terme, une étude plus approfondie de ces gènes visera à rechercher l'existence de variants rares (voire *de novo*) non décrits dans la littérature ou dans les bases de génétique qui pourraient favoriser la survenue des formes les plus sévères de PTLD.

I.1.1. Méthode

Cette recherche a été réalisée chez 8 cas de PTLD précoces, apparus au cours de la première année suivant la transplantation (notion de « phénotype extrême », ayant la plus grande probabilité d'être associé à un variant causal rare), receveurs de greffe rénale et 8 témoins appariés en particulier sur le statut virologique Epstein-Barr Virus pré-greffe (facteur de risque connu de PTLD) m ais aussi le sexe, l'âge au moment de la transplantation, la date de la transplantation, le nombre de transplantations.

La conception, la collecte des données, la collecte des échantillons et les analyses ont été approuvés par le comité éthique du centre hospitalier universitaire de Limoges. Tous les

patients ont fourni un consentement éclairé écrit pour la collecte et le stockage d'ADN dans une bio-banque autorisée par les autorités compétentes (DC-2008-604) et pour la bio-analyse de leur échantillon, y compris les enquêtes pharmacogénétiques. Le comité institutionnel de contrôle de l'hôpital universitaire de Limoges a approuvé l'étude et la CNIL a autorisé le traitement des données à caractère personnel (N°910216).

I.1.1.1. Séquençage

L'ADN a été obtenu à partir de sang total, de sérum et de tissus non malins contenus dans des fixatifs prévenant la dégradation avec des kits d'extraction dédiés à chaque matrice selon le protocole du fabricant.

La préparation des librairies a été effectuée à partir de 100ng d'ADN avec le kit de préparation de la bibliothèque Ion AmpliSeq™ Exome RDY (Thermo Fisher Scientific, États-Unis) selon le protocole du fabricant. La qualité et la quantité des bibliothèques ont été déterminées par le système de bio-analyse Agilent 2100 avec le kit High Sensitivity DNA. L'enrichissement clonal des échantillons a été réalisé à l'aide du système ionique OneTouch™ II (Thermo Fisher Scientific, USA). Un ou deux échantillons de librairies ont été chargés sur une puce Ion P1 et soumis à des cycles de séquençage sur le séquenceur Proton Ion (Thermo Fisher Scientific, USA).

Les séquences lues par NGS ont été alignées sur le génome de référence humain hg19 (ou GRCh37) avec l'outil BWA[27], générant des fichiers de séquences alignées au format BAM (Binary Alignment Map).

Le pourcentage de couverture a été calculé comme la taille des zones couvertes avec une profondeur de lecture supérieurs à 20 séquences par rapport à l'exome complet (zones cibles).

I.1.1.2. Appel de variants

Pour chaque patient, un appel de variant exploratoire a été effectué à l'aide de l'outil HaplotypeCaller (Broad Institute's GATK library), qui détecte simultanément les SNP et les indels de la lignée germinale via l'assemblage local *de novo* d'haplotypes dans chaque région présentant des variations par rapport au génome de référence.

L'outil bedtools[28] a été utilisé pour calculer la profondeur de lecture à la position de chaque variant, c'est à dire le nombre de séquences lue par NGS alignées à cette position. Pour chaque variant, la profondeur a été vérifiée pour chaque patient de la cohorte, et seuls les variants avec une profondeur de lecture ≥20 pour chaque patient ont été conservés. Cela nous a permis d'avoir la certitude que la non détection chez un patient d'un variant présent

chez le patient apparié serait due à l'effective absence de ce variant, plutôt que d'être causée par un manque de couverture.

Les variants détectés ont été stockés dans des fichiers VCF comprenant pour chacun: la position génomique, le nombre de lectures avec le nucléotide de référence et le nombre de lectures avec le nucléotide alternatif.

L'annotation des variants a été effectuée avec ANNOVAR[29] et la base de données Ensembl ![30].

L'annotation des gènes a été effectuée grâce aux bases de données GeneOntology[31], Kyoto Encyclopedia of Genes and Genomess[32] ainsi que le package R ReactomePA[33].

Script 1 : Appel de variants

```
GATK=GenomeAnalysisTK.jar
snpEff=/media/disque/Outils/java/snpEff_latest_core/snpEff/snpEff.jar
for file in ./*bam
	do
	echo $file
	#1 - Appel de variants exploratoire
	v=${file%.bam}'_GATK.vcf'
	java -jar $GATK -T HaplotypeCaller -R hg19.fasta -I $file -o $v
	#2 - Calculer les couvertures
	cov=$v'_cov'
	bedtools coverage -a $bed -b $file -counts > $cov
	#3 - Annoter
	out=$v'.annot'
	java -jar $snpEff GRCh37.75 -geneId -canon -onlyProtein -o vcf $v > $out
done;
```

I.1.1.3. Recherche de gènes d'intérêt

I.1.1.3.1. Sélection des gènes

Les analyses de cette section ont été effectuées avec le langage R[34].

Degré de variation des gènes Nous avons établi une mesure prenant en compte le degré de variation d'un gène pour un patient spécifique, défini comme le nombre moyen de variants dans ce gène par séquence de 100 nucléotides. Les données brutes concernent des exomes, mais aussi les zones introniques très proches. Lors du calcul de degré de variation, les variants ont été pondérés selon leur localisation : 1 pour exonique ; 0,1 dans le cas contraire.

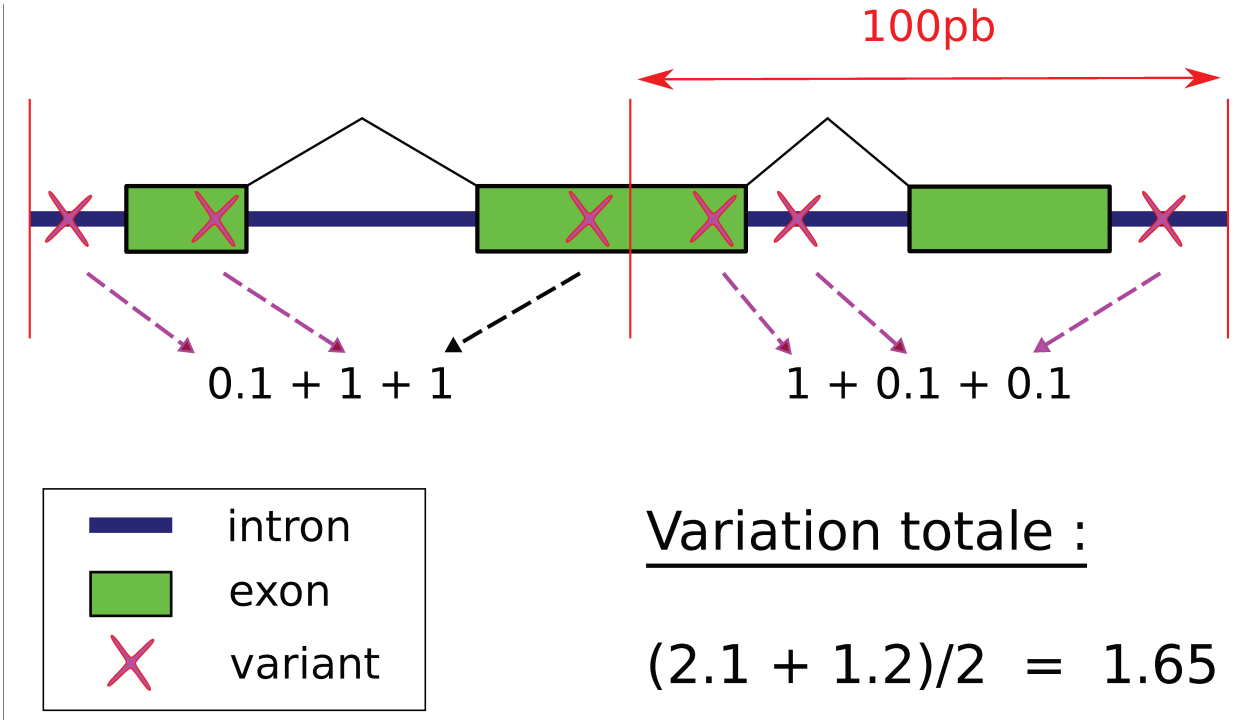

Figure 3: Le degré de variation des gènes représente la moyenne du nombre de variants pour 100pb, pondérés selon leur localisation

Nous avons ainsi reporté les degrés de variation par gènes sur deux matrices, l'une pour les cas et l'autre pour les témoins.

Script 2 : Calcul de la variabilité des gènes à partir de fichiers vcf

```
for (vcf in listVcf){
	data = read.table(vcf)
	data = data[data$AD/data$DP >= 0.1,]
	data$poids = ifelse(data$protein == '', 0.1, 1)
	data = aggregate(data$poids, by = list(data$ensembl), FUN = 'sum')
	colnames(data) = c('ensembl', 'var')
	write.table(data, paste('variabilite', vcf), row = F)
}
```

I.1.1.3.2. Classification

L'algorithme de classification non hiérarchique k-means a été utilisé sur les matrices de variation pour regrouper les gènes dont les degrés de variation sont similaires. Il consiste à affecter de manière aléatoire les centres d'un nombre défini de classes, affecter chaque élément (ici des gènes) à la classe la plus proche, puis recalculer les centres de classe et réaffecter les éléments, en boucle jusqu'à atteindre l'inertie.

Le calcul des distances s'est fait en utilisant des distances euclidiennes.

La méthode Elbow[35] a été utilisée pour déterminer le nombre de classes optimal (paramètre k), avec k ∈ 1:9, le centre du cluster de départ placé à 25, et 150 itérations. Cette méthode

consiste à minimiser les distances intra-classe en trouvant la valeur minimale de k pour lequel la variance ne se réduit plus significativement.

Une analyse en composante principale (ACP), utilisant la librairie factoextra[36] a réduit le nombre de dimensions projetées, passant de 8 dimensions (1 dimension pour chaque patients) à 2 dimensions (reconstituées à partir des 8 d'origine). Cela a permis la visualisation des résultats du partitionnement des données, et éventuellement de départager le nombre de classe optimal.

Dans le cluster contenant les gènes très variables, seuls ceux présentant une variabilité dans au moins 6 échantillons sur 8 (cas ou témoins) ont été sélectionnés.

Script 3 : Classification des gènes selon leur variabilité

```
var_tem = var_tem[apply(var_tem, 1, function(x){length(x[x > 0])}) > 0,]
var_cas = var_cas[apply(var_cas, 1, function(x){length(x[x > 0])}) > 0,]
#choix du nombre de cluster pour kmeans
ratio = data.frame(cluster = 1:9, ratio_tem = 1:9, ratio_cas = 1:9, ratio_ct = 1:9)
for (k in 1:9) {
  km_model = kmeans(scale(var_tem), k, nstart = 25, iter.max = 150)
  ratio$ratio_tem[k] = km_model$tot.withinss / km_model$totss
  km_model = kmeans(scale(var_cas), k, nstart = 25, iter.max = 150)
  ratio$ratio_cas[k] = km_model$tot.withinss / km_model$totss
}
#le meilleur nombre de cluster se trouve au niveaux du coude de ce scree plot
plot_centerclust_cas = ggplot(ratio, aes(cluster, ratio_cas)) + geom_line() + geom_point()
plot_centerclust_tem = ggplot(ratio, aes(cluster, ratio_tem)) + geom_line() + geom_point()
#Classification
km_model_tem3 = var_tem[,tem] %>% scale() %>% kmeans(centers = 3, nstart = 25, iter.max =
150)
km_model_tem2 = var_tem[,tem] %>% scale() %>% kmeans(centers = 2, nstart = 25, iter.max =
150)
km_model_cas3 = var_cas[,cas] %>% scale() %>% kmeans(centers = 3, nstart = 25, iter.max =
150)
km_model_cas2 = var_cas[,cas] %>% scale() %>% kmeans(centers = 2, nstart = 25, iter.max =
150)
var_tem$cluster3 = km_model_tem3$cluster
var_tem$cluster2 = km_model_tem2$cluster
var_cas$cluster3 = km_model_cas3$cluster
var_cas$cluster2 = km_model_cas2$cluster
#PCA
vizclust_tem3 = fviz_cluster(km_model_tem3, data = var_tem[, tem], show.clust.cent = TRUE,
labelsize = 0, main = "temoins k=3")
vizclust_tem2 = fviz_cluster(km_model_tem2, data = var_tem[, tem], show.clust.cent = TRUE,
labelsize = 0, main = "temoins k=2")
```

```
vizclust_cas3 = fviz_cluster(km_model_cas3, data = var_cas[, cas], show.clust.cent = TRUE,
labelsize = 0, main = "cas k=3")

vizclust_cas2 = fviz_cluster(km_model_cas2, data = var_cas[, cas], show.clust.cent = TRUE,
labelsize = 0, main = "cas k=2")
```

I.1.1.3.3. Cohérence des résultats

L'approche par classification vise à déterminer les gènes particulièrement variable. Afin de déterminer si cette forte variabilité est due au hasard, deux méthodes ont été utilisées. Pour commencer, nous avons cherché à déterminer si les variants présents dans les gènes sélectionnés permettaient de prédire le statut PTLD des patients. Puis nous avons vérifié si les gènes sélectionnés pressentaient une différence d'expression dans des tissus tumoraux et dans des tissus sains.

Variants utilisables comme modèle de risque

Les variants provenant des gènes sélectionnés par classification ont été utilisés pour créer des modèles de risque PTLD.

Nous avons établi une valeur de présence de chaque variant pour chaque patient, calculée à partir des informations présentes dans les vcf : AD donnant pour chaque allèle du variant (sauvage et alternatives) le nombre de séquences lues, et DP donnant la profondeur totale à la position. La valeur de présence calculée correspond au ratio du nombre de séquence contenant le variant divisé par la profondeur totale. Elle se rapproche de 1 en cas de variant homozygote, et de 0,5 en cas d'hétérozygotie. La nature aléatoire des lectures NGS fait que le ratio est rarement exact ; nous avons donc considéré que les ratio <0,1 provenaient d'erreurs de séquençage.

Nous avons utilisés 6 couples cas-témoins pour établir des modèles et 2 couples pour les tester. Avec 8 couples au total, cela nous a permis de faire 28 modèles pour chaque groupe de gènes : gènes spécifiques aux cas, gènes spécifiques aux témoins, et l'ensemble des gènes. Ces groupes ont été créés par analyse discriminante de régression parcimonieuses des moindres carrés partiels (sPLS-DA, sparse Partial Least Squares Discriminant Analysis), méthode qui permet la sélection des variants les plus prédictives ou discriminatoires afin de classer les patients[37]. Le package R mixOmics[38] a été utilisé : la fonction *tune.splsda* pour déterminer les paramètres optimaux par validations croisées (5), la fonction splsda pour créer le modèle, et la fonction *predict* pour prédire des statuts PTLD.

Un modèle est considéré comme bon s'il permet de prédire correctement le statut PTLD des 4 patients tests (2 couples cas-témoins). La petite taille de la cohorte utilisée ne permet pas d'être certain que le modèle permettrait de bien prédire des patients hors de la cohorte. Cependant, les gènes ayant certains variants retrouvés dans plusieurs modèles jugés corrects sont ceux dont le haut degré de variabilité n'est pas lié au hasard.

Script 4 : Création et test de modèles sPLS-DA à partir des variants des gènes sélectionnés

```
#data, l'ensemble des variants
#cas, la liste des patients cas
#tem, la liste des patients témoins
comb = combn(1:8, 2)
group = c(rep('cas', 8), rep('tem', 8))
group_train = c(rep('cas', 6), rep('tem', 6))
group_test = c(rep('cas', 2), rep('tem', 2))
list.keepX = c(1:30,  seq(30, dim(data)[1]/2, 10))
for(i in comp){
  cas_train = cas[!(cas %in% cas[i])]
  cas_test = cas[!(cas %in% cas_train)]
  tem_train = tem[!(tem %in% tem[i])]
  tem_test = tem[!(tem %in% tem_train)]
  data_train = t(as.matrix(data[, c(cas_train, tem_train)]))
  data_test = t(as.matrix(data[, c(cas_test, tem_test)]))
  #entrainement
  tune.splsda = tune.splsda(data_train, group_train, ncomp = ncomp, test.keepX =
list.keepX, validation = 'Mfold', dist = 'max.dist', measure = "BER", cpus = 3,
progressBar = F)
  select.keepX = tune.splsda$choice.keepX[1:ncomp]
 splsda.train = splsda(data_train, group_train, ncomp = ncomp, keep = select.keepX)
  #test
  test.predict = predict(splsda.train, data_test, dist = "max.dist")
  predicted = test.predict$predict[,,1]*splsda.train$explained_variance$X[1]
  for (nc in 2:ncomp){
 predicted = predicted + test.predict$predict[,,nc] *
splsda.train$explained_variance$X[nc]
  }
  predicted = abs(predicted)
  predicted = ifelse(predicted[, 'cas'] > predicted[, 'tem'], 'cas', 'tem')
  #Selection
  mat_test = get.confusion_matrix(truth = group_test, predicted = predicted)
  ok = mat_test[1, 1] + mat_test[2, 2]
  faux = mat_test[1, 2] + mat_test[2, 1]
  if (ok == 4 & faux == 0){
    n = n + 1
    pred[[n]] = list(i = i, pred = splsda.train, ok = ok, faux = faux)
  }
}
```

Expression génique dans des cancers similaires

Le serveur GEPIA (Gene Expression Profiling Interactive Analysis)[39], utilisant les données du portail TCGA (The Cancer Atlas Program), permet d'accéder à des analyses ANOVA comparant l'expression des gènes selon le type de tissus :

- 337 tissus sains
- 47 tissus tumoraux provenant de lymphomes diffus à grandes cellules B (DLBC) dont PTLD fait parti
- 118 tissus tumoraux provenant de thymomes (THYM), ou tumeurs épithéliales thymiques liés au système immunitaire bien que généralement bénins

Le seuil de significativité de la p-value a été placé à 0,05.

Nous avons utilisé les résultats du serveur GEPIA pour évaluer la différence d'expression entre tissus tumoraux et sains pour les gènes que nous avons identifié avec un haut degré de variation. Seules les expressions des gènes dont la variation n'est pas due au hasard (c'est-à-dire utilisés dans les modèles prédictifs) ont été analysées.

I.1.2. Résultats

I.1.2.1. Séquençages

Les exomes des 16 patients ont été séquencés avec des couvertures allant de 80,4 % à 97,4 % de l'exome, et des profondeurs moyennes allant de 98 à 195 lectures.

Tableau 2 : Couverture des exomes : Profondeur < 20, Profondeur entre 20 et 50, Couverture et Profondeur moyenne sur l'exome.

Patient	Prof <20	Prof 20-50	Prof >50	Couverture	Moyenne
Cas 1	11,6%	18,36 %	70,04 %	88,4%%	109
Témoins 1	10,8%	15,75 %	73,44 %	89,19%%	132
Cas 2	5,7%	11,53 %	82,75 %	94,28%%	152
Témoins 2	8,8%	15,47 %	75,74 %	91,21%	178
Cas 3	7,2%	15,50 %	77,28 %	92,78%	105
Témoins 3	12,9%	20,26 %	66,80 %	87,06%	98
Cas 4	4,3%	9,72 %	85,97 %	95,69%	132
Témoins 4	9,0%	14,38 %	76,58 %	90,96%	136
Cas 5	10,5%	16,93 %	72,56 %	89,49%	115
Témoins 5	9,0%	13,54 %	77,51 %	91,05%	158
Cas 6	19,6%	21,36 %	59,00 %	80,36%	108
Témoins 6	9,9%	13,43 %	76,64 %	90,07%	167

Cas 7	10,7%	13,76 %	75,57 %	89,33%	172
Témoins 7	4,8%	12,48 %	82,70 %	95,18%	121
Cas 8	2,6%	5,31 %	92,12 %	97,43%	195
Témoins 8	7,2%	13,88 %	78,95 %	92,83%	107

L'appel de variants a détecté :

- 529 292 variants répartis sur 26 926 gènes pour les cas
- 585 438 variants répartis sur 27 487 gènes pour les témoins

Deux échantillons de cas contenaient environ 1 000 variants, alors que tous les autres cas et témoins comptaient entre 13 000 et 23 000 variants. Cependant, la distribution des variants pour ces deux cas étant cohérente avec celles des autres échantillons et la qualité globale étant supérieure aux seuils prédéfinis, ces 2 échantillons ont été conservés pour les analyses.

I.1.2.2. Recherche de gènes d'intérêt

I.1.2.2.1. Sélection des gènes

La matrice de variabilité pour les cas a été classifiée en deux groupes : 19 740 gènes peu variables, et 394 gènes avec une forte variabilité. La matrice de variabilité pour les témoins a été classifiée en deux groupes : 20 498 gènes peu variables, et 301 gènes avec une forte variabilité.

124 gènes ont été classifiés comme de haute variabilité à la fois pour les cas et les témoins, laissant :

- 270 gènes à haute variabilité chez les cas seulement, dont 12 gènes se montraient variable chez au moins 6 des 8 cas (Tableau 3).
- 177 gènes à haute variabilité chez les témoins seulement, dont 46 gènes se montraient variable chez au moins 6 des 8 témoins (Tableau 4).Tableau 3 : 12 gènes de la classe de haute variabilité chez les cas mais non chez les témoins, et avec une variabilité non nulle chez au moins 6 des 8 cas

Gène	**Identifiant Ensembl !**
TESK1	ENSG00000107140
HRC	ENSG00000130528
KREMEN2	ENSG00000131650
CYSLTR2	ENSG00000152207
CCDC8	ENSG00000169515
C9orf131	ENSG00000174038
SLC16A11	ENSG00000174326
GP5	ENSG00000178732

BBS10	ENSG00000179941
EMILIN3	ENSG00000183798
TSPYL1	ENSG00000189241
GPAA1	ENSG00000197858

Tableau 4 : 46 gènes de la classe de haute variabilité chez les témoins mais non chez les cas, et avec une variabilité non nulle chez au moins 6 des 8 témoins

Gènes	**Identifiant Ensembl !**
ZNRF4	ENSG00000105428
MYF6	ENSG00000111046
TPSG1	ENSG00000116176
GPR31	ENSG00000120436
TAS2R8	ENSG00000121314
ADAM30	ENSG00000134249
SIT1	ENSG00000137078
CABS1	ENSG00000145309
C10orf12	ENSG00000155640
CCDC116	ENSG00000161180
SOX14	ENSG00000168875
ADRB2	ENSG00000169252
VPREB1	ENSG00000169575
OR7G2	ENSG00000170923
OR1B1	ENSG00000171484
OR5B12	ENSG00000172362
OR1S1	ENSG00000172774
TRAM1L1	ENSG00000174599
OR10P1	ENSG00000175398
AURKAIP1	ENSG00000175756
OR10H4	ENSG00000176231
OR51B6	ENSG00000176239
OR4S1	ENSG00000176555
OR51G1	ENSG00000176879
SPATA31E1	ENSG00000177992
CALML5	ENSG00000178372
OR1E1	ENSG00000180016
OR4C6	ENSG00000181903
GLTPD2	ENSG00000182327
OR5P2	ENSG00000183303
OR2T10	ENSG00000184022
TACSTD2	ENSG00000184292
DDX53	ENSG00000184735
SOWAHB	ENSG00000186212
OR10J1	ENSG00000196184
OR14A16	ENSG00000196772
OR2T4	ENSG00000196944
DXO	ENSG00000204348
CNN2P9	ENSG00000213149
C1GALT1C1L	ENSG00000223658
ZNF469	ENSG00000225614
OR5H6	ENSG00000230301

RNF148	ENSG00000235631
OR2AE1	ENSG00000244623
PTX4	ENSG00000251692
AC008742.1	ENSG00000267650

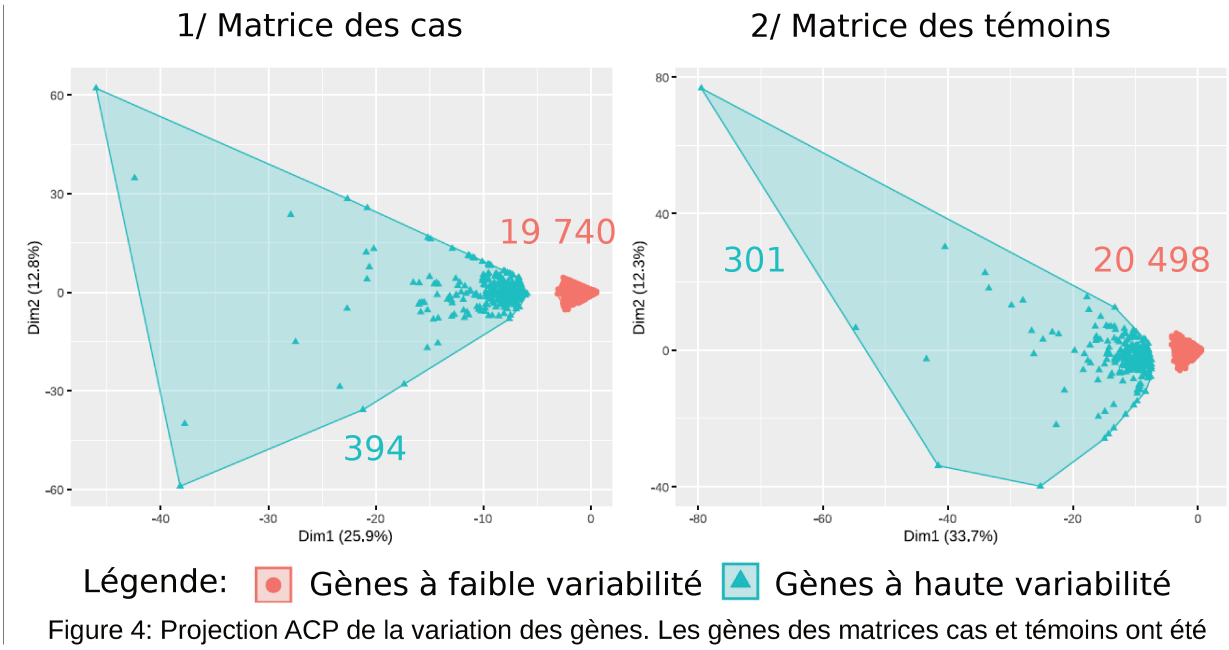

Figure 4: Projection ACP de la variation des gènes. Les gènes des matrices cas et témoins ont été regroupés en classe selon leur variabilité, 1/ Matrice de variation des gènes chez les cas PTLD, 2/ Matrice de variation des gènes chez les témoins.

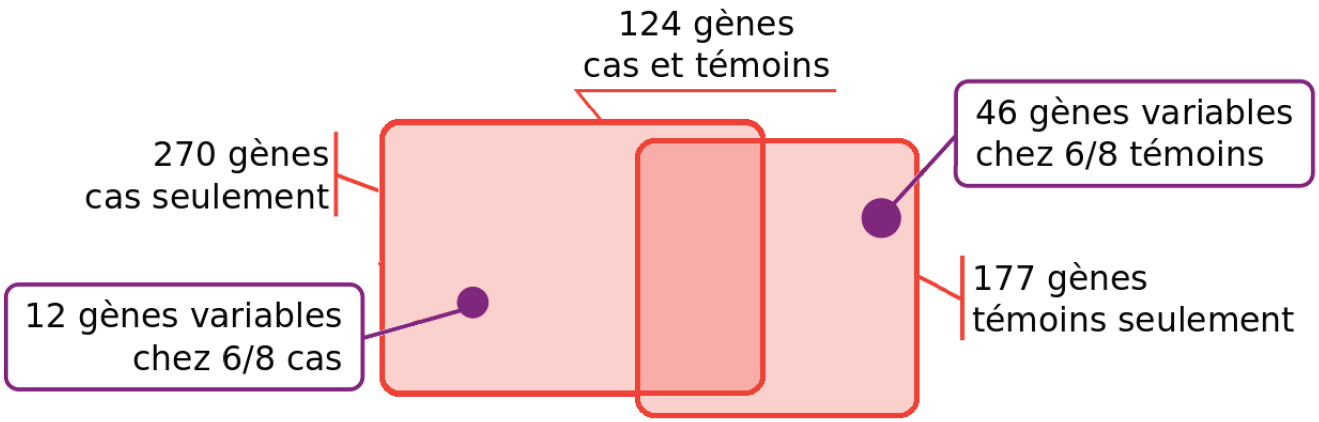

Figure 5: Diagramme de Venn des gènes sélectionnés par classification. Seuls les gènes variables chez au moins 6 des 8 patients du groupe sont considérés

I.1.2.2.2. Cohérence des résultats

Variants utilisables comme modèle de risque

Aucun des 28 modèles créés en utilisant les 46 gènes à haute variabilité chez les témoins n'ont pu prédire correctement le statut PTLD des 4 patients test.

Les modèles créés en utilisant les 58 gènes à haute variabilité ont donné des résultats

similaires à ceux n'utilisant que les 12 gènes à haute variabilité chez les cas.

Parmi les 28 modèles créés en utilisant les 12 gènes à haute variabilité chez les cas, 18 ont permis des prédictions correctes du statut PTLD des 4 patients test. Pour créer ces modèles, les variants de 7 des 12 gènes ont été utilisés : BBS10, C9orf131, CCDC8, GPAA1, HRC, KREMEN2 et SLC16A11.

Expression génique dans des cancers similaires

Les gènes GPAA1, KREMEN2, TESK et TSPYL1 sont significativement surexprimés dans les tissus tumoraux de DLBC et THYM par rapport aux tissus sains.

Les gènes BBS10, CCDC8 et SLC16A11 sont significativement surexprimés dans les tissus tumoraux de THYM par rapport aux tissus sains, mais ne présentent pas de différence significative d'expression entre les tissus tumoraux DLBC et les tissus sains.

Les gènes C9orf131, CYSLTR2, EMILIN3, GP5 et HRC ne présentent pas de différences significatives d'expression entre les tissus tumoraux (DLBC ou THYM) et les tissus sains.

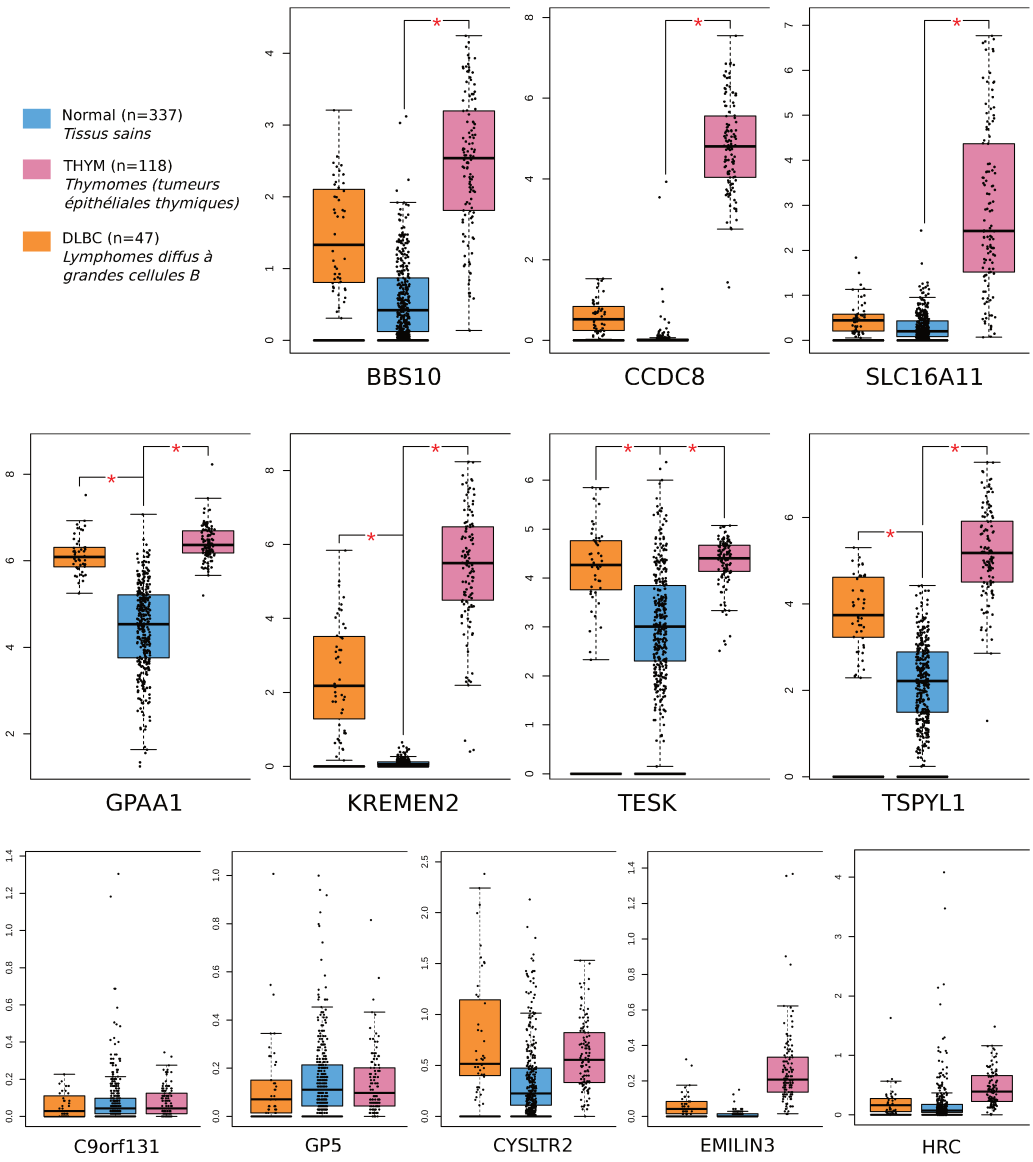

Figure 6: Expression des gènes, Comparaison par ANOVA des distribution d'expression génique de 337 tissus sains avec 118 tumeurs épithéliales et avec 47 lymphomes, Le seuil de significativité de la p-value est placé à 0,05.

I.1.2.2.3. Sélection des gènes

Parmi les gènes sélectionnés par classification, 3 n'ont été confirmés ni par les modèles ni par l'analyse d'expression : CYSLTR2 (ENSG00000152207), EMILIN3 (ENSG00000183798) et GP5 (ENSG00000178732).

Deux autres gènes n'ont pas été confirmés par les modèles, mais montrent une sur-expression significative à la fois dans THYM et DLBC. Leur implication dans les processus cancéreux ne peut pas être écartés, mais pourrait ne pas être limitée à PTLD, Il s'agit des gènes TESK1 (ENSG00000107140) et TSPYL1 (ENSG00000189241).

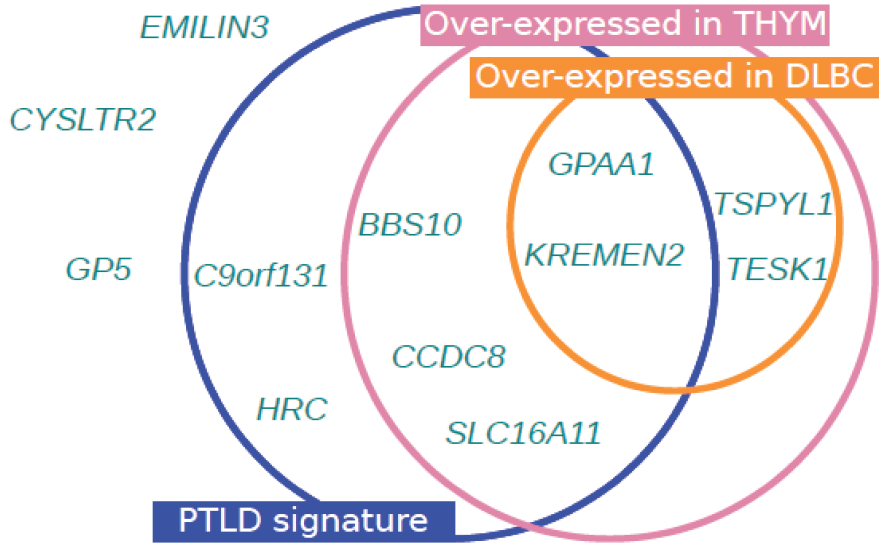

Figure 7: Sélection des gènes. Sélection parmi les 12 gènes précédemment classés comme hautement variables, en fonction des résultats des signatures PTLD et de la comparaison d'expression.

Nous avons finalement sélectionné les 7 gènes confirmés d'intérêt par les modèles, 2 d'entre eux ne montrent pas de différences significatives d'expression dans les tissus tumoraux :

- HRC (ENSG00000130528) est une protéine de liaison du calcium faisant partie des voies de signalisation du calcium et de la contraction musculaire. Son implication dans PTLD pourrait être liée à la voie des ICNs plutôt qu'à celle des lymphomes.
- C9orf131 (ENSG00000174038) codant une protéine non caractérisée à ce jour.

Les 5 autres gènes sont sur-exprimés dans des tissus tumoraux similaires à PTLD :

- Le gène BBS10 (ENSG00000179941) fait parti du complexe BBSome, impliqué dans la formation, le transport et les fonctions des cils cellulaires[40], 3 pistes pourraient expliquer l'implication du gène dans PTLD :
  - Les cils sont impliqués dans diverses voies de signalisation, incluant WNT, qui est lié aux leucémies lymphocytaires chroniques ainsi qu'aux lymphomes à cellules B[41].
  - La régulation de mTOR par les cils joue un rôle dans l'apoptose cellulaire[42], dont le dérèglement est une cause principale de tumorigenèse, tout comme leur régulation de la voie Hippo pourrait causer la tumorigenèse[4344].
  - Les cils sont aussi impliqués dans le processus d'autophagie[45], ce qui pourrait également les lier à la tumorogenèse.
- La protéine codée par CCDC8 (ENSG00000169515) est un co-facteur requis pour

l'apoptose induite par p53 à la suite de lésions de l'ADN. Il pourrait donc prévenir l'effet de EBV en provoquant la mort des cellules infectées, un mécanisme que son altération pourrait empêcher.

- GPAA1 (ENSG00000197858) est une ancre pour le Glycosylphosphatidylinositol (GPI), qui est impliqué dans le mécanisme de liaison GPI à la membrane de surface des cellules lors du transfert, en particulier sur la protéine du récepteur de l'urokinase (uPAR) :
  - uPAR est depuis longtemps associé aux métastases, et est connu pour être sur-exprimé dans la plupart des cancers humain[46].
  - Plus récemment, uPAR a été suggéré comme biomarqueur de la réponse et de la survie des patients atteints de lymphome[47].
  - uPAR pourrait également jouer un rôle mineur dans la croissance et les métastases des lymphome[48].
- KREMEN2 (ENSG00000131650) code DKK1, un récepteur transmembranaire d'un homologue à haute affinité de dickkopf. Il s'agit d'un inhibiteur sécrété par la voie Wnt/β-catenin, l'une des voies critiques régulant la transition épithélio-mésenchymateuse (TME), reconnue comme une marque caractéristique du cancer[49]. La voie contribue aussi aux TME[50], l'un des 4 cofacteurs causaux de PTLD.
- SLC16A11 (ENSG00000174326) est un transporteur du pyruvate à travers la membrane plasmique[51] ayant un rôle important dans l'angiogenèse de la croissance tumorale et des métastases[5253], en particulier dans en cas de lymphome[54].

I.1.3. Conclusion du projet PTLD

Dans cette étude, en utilisant une approche sans *a priori* (exome entier) et en comparant des patients transplantés rénaux atteints d'un PTLD précoce par rapport à des patients témoins appariés, nous avons identifié 12 gènes dont la séquence nucléotidique varie fortement dans les cas mais non dans les témoins.

7 gènes ont été sélectionnés pour confirmation dans de futures études. Le fait qu'une partie d'entre eux présentent des liens, même distants, aux processus tumoraux ou à la voie de la calcineurine est intéressant étant donné l'approche de départ sans *a priori.* Cela doit cependant être tempéré, puisque 4 des 5 gènes non sélectionnés sont aussi indirectement

liés aux processus tumoraux ou au système immunitaire.

En effet, CYSLTR2 est un récepteur pour leucotriènes cystéinyles (LT), parmi lesquels LTB4 qui est impliqué dans l'inflammation et qui permet l'activation des cellules T inhibant la prolifération des cellules B induite par EBV[55]. EMILIN3 est un interfaceur de microfibrilles d'élastine, une famille de gènes liés aux TME[56], ainsi qu'à la viabilité des cellules tumorales[57]. GP5 fait partie du système Ib-V-IX, récepteur du facteur de von Willebrand ayant été retrouvée dans des niveaux très élevés chez des patients atteints de lymphome[58] bien qu'il en existe des variations non cancéreuses. Enfin, TSPYL1 est liée à la famille des protéines d'assemblage de nucléosomes (NAP) qui a certaines associations avec le cancer, mais le gène lui-même n'y semble pas lié. Cependant, son paralogue TSPYL2 a été identifié comme promoteur et activateur de la lysine-acétyl-transférase p300, impliqué dans la régulation de l'apoptose et la tumorigenese[59].

Les gènes à haute variabilité sont moins regroupés que les gènes à faible variabilité, ce qui pourrait avoir plusieurs causes. D'une part, ces gènes montrent des degrés de variabilité très différents : tous les gènes de cette classe varient plus que le reste du génome, mais leur variabilité peut aussi bien être 2 fois plus variable que 100 fois plus. D'autre part, certains gènes ne sont pas variables sur plus d'un ou deux patients, raison pour laquelle nous n'avons considérés que les gènes présentant une variabilité sur au moins 6 des 8 patients du groupe.

L'approche par classification a montré une bonne sensibilité pour la sélection des gènes, dont nous avons voulu confirmer la variabilité liée au statut PTLD. Nous avons pour cela cherché à créer des modèles prédictifs, partant du principe qu'un gène impliqué dans PTLD présenterait un ensemble de variations chez les cas différent de chez les témoins. La création de modèles prédictifs permet de vérifier cette hypothèse.

I.2. Projet GENODAT

Le diabète d'apparition tardive après transplantation (NODAT, New-Onset Diabetes After Transplantation) est un diabète de type 2 apparaissant après une transplantation rénale[60], associé à des rejets de greffes, des infections et des maladies cardiovasculaires, entraînant une diminution de la survie globale de ces patients[6162] en addition des complications généralement observées chez les patients atteints de ce type de diabète[63].

Comme tout diabète, le NODAT se caractérise par une hyperglycémie chronique due à la résistance à l'insuline des tissus périphériques (muscle squelettique, foie, tissu adipeux) et à une sécrétion d'insuline insuffisante par les cellules bêta du pancréas.

Plusieurs facteurs à risque de NODAT ont été identifiés. Certains sont non modifiables, tels que l'âge, l'origine ethnique, des antécédents familiaux de diabète, ou certains types d'infections comme l'hépatite C ou des troubles biologiques comme l'hypomagnésémie. D'autres peuvent être modifiés, tels que l'indice de masse corporelle, ou le type immunosuppresseurs prescrit[64]. En effet, si la prévalence de NODAT est d'environ 25% chez les patients transplantés rénaux précédemment non diabétiques[65], elle varie selon le traitement. Elle est en effet bien plus importante en cas de prise de tacrolimus que de ciclosporine, à tel point que le traitement peut être modifié (arrêt du tacrolimus au profit de la ciclosporine) dès les premiers signes d'intolérance glucidique après la greffe. L'incidence moyenne de diabète chez les patients transplantés après 6 mois est similaire à celle des nouveaux diabètes chez les patients en liste d'attente, soit environ 6% par an[66], mais elle atteint 20,5 % lorsqu'ils sont traités par ICN[16].

Le facteur de risque représenté par l'origine ethnique des patients, permet de supposer que la variabilité de certains gènes encore indéterminés pourrait influencer le développement de NODAT.

Le projet GENODAT a pour objectif d'identifier de nouveaux gènes candidats impliqués dans le développement ou la susceptibilité à NODAT, par l'étude d'exomes de patients transplantés rénaux traités sous tacrolimus (21 cas extrême sélectionné sur la base de leur polymédication antidiabétique et 21 témoins appariés).

A terme, une étude plus approfondie de ces gènes visera à rechercher l'existence de variants rares (voire *de novo*) non décrits dans la littérature ou dans les bases de génétique qui pourraient favoriser la survenue des formes les plus sévères de NODAT.

I.2.1. Méthode

La conception, la collecte des données et des échantillons et les analyses ont été approuvés

par le comité d'examen éthique institutionnel de l'hôpital universitaire de Limoges. Tous les patients ont fourni un consentement éclairé écrit pour la collecte et le stockage d'ADN dans une bio-banque autorisée par les autorités compétentes (DC-2008-604) et pour la bio-analyse de leur échantillon, y compris les enquêtes pharmacogénétiques. Le comité institutionnel de contrôle de l'hôpital universitaire de Limoges a approuvé l'étude et la CNIL a autorisé le traitement des données à caractère personnel (N°912242).

Le projet GENODAT intervient dans la continuité de l'étude EPHEGREN, une étude longitudinale ayant pour but d'explorer les facteurs pharmacologiques prédictifs de l'évolution à long terme (3 ans) de la fonction rénale ainsi que la survenue de cancers, de diabète, de pathologies cardiovasculaires chez les patients transplantés rénaux[6768]. Les exomes étudiés proviennent de cette étude.

Le projet GENODAT a analysé de l'ADN extrait d'échantillons de salive de 21 cas de NODAT et de 21 témoins, tous appariés par âge, sexe, IMC (>35 ou <35), greffe depuis un donneur vivant ou décédé, date de la transplantation, et traitement ou non par corticoïde, Tous les patients ont été traités par tacrolimus, et aucun ne présentaient de diabète pré-greffe.

Le séquençage et l'appel de variant ont utilisé le même protocole que le projet PTLD.

I.2.1.1. Recherche de gènes d'intérêt

I.2.1.1.1. Sélection des gènes

Les analyses de cette partie ont été effectuées avec le langage R.

Le degré de variation des gènes a été calculé selon le protocole mis en place pour le projet PTLD. Dans le cas de GENODAT, tous les variants représentant plus de 10 % des lectures à leur position ont été pris en compte (et pas seulement ceux spécifiques aux cas ou aux témoins). Une classification des gènes a été effectuée à deux reprises, selon leur variabilité chez les cas, et leur variabilité chez les témoins.

La méthode Elbow a été utilisée pour déterminer le nombre de classes optimal (paramètre k), avec $k \in 1:9$, le centre du cluster de départ placé à 25, et 150 itérations ont été effectuées. Une analyse en composante principale (ACP) a réduit le nombre de dimensions projetées, passant de variables à 21 dimensions (1 par patient) à 2 dimensions (reconstituées à partir des 21 d'origine). Cela a permis la visualisation des résultats du partitionnement des données, et éventuellement de départager le nombre optimal de classes.

I.2.1.1.2. Cohérence des résultats

Les variants provenant des gènes sélectionnés par classification ont été utilisés pour créer

des modèles de risque NODAT selon un protocole dérivé de celui utilisé lors du projet PTLD, mais ajusté pour tenir compte de la taille de la cohorte. La valeur de présence de chaque variant pour chaque patient a été calculée à partir des informations présentes dans les vcf, proche de 1 en cas de variant homozygote, et de 0,5 en cas d'hétérozygotie. Les valeurs <0,1 ont été considérées comme provenant d'erreurs de séquençage.

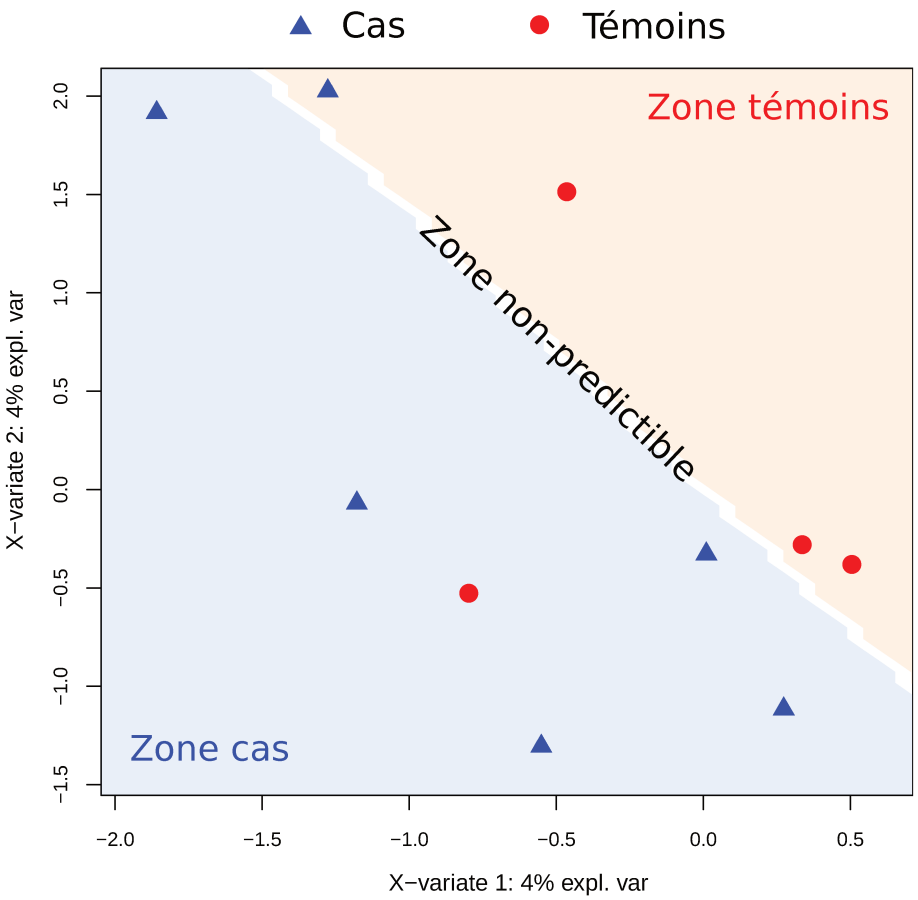

Figure 8: Exemple de modèle prédictif de variants-signature NODAT. Le modèle a tenté de prédire chacun des 7 cas et 7 témoins des couples tests, en calculant leurs valeurs pour chaque variables (ici X-variate 1 et X-variate 2). Selon la zone où ces valeurs les placent, le patient est prédit cas ou bien témoins. Dans cet exemple, 5 cas et 3 témoins ont correctement été prédits, tandis que 1 cas et 1 témoins ne l'ont pas été. Le modèle n'a pas pu prédire les 4 patients restants, car les valeurs calculées les auraient placées dans la zone floue non prédictible.

Les modèles ont été créés par analyse sPLS-DA en utilisant 14 couples cas-témoins. Avec 21 couples au total, il serait possible de créer 116 280 modèles. Nous en avons créé 11628, soit 1 sur 10, utilisant des tirages aléatoires pour choisir les couples d'entraînement.

Les modèles ont été testés en prédisant le statut NODAT des 7 couples non utilisés pour leur création, et les nombre de patients prédits correctement et incorrectement ont été comptabilisés.

Script 5 : Sélection au hasard des patients d'entraînement et de test

```
comb = combn(1:21, 7)
hasard = sample(1:dim(comb)[2], dim(comb)[2]/10, replace = F)
for(i in hasard){
  cas_train = cas[!(cas %in% cas[comb[, i]])]
  cas_test = cas[!(cas %in% cas_train)]
  tem_train = tem[!(tem %in% tem[comb[, i]])]
  tem_test = tem[!(tem %in% tem_train)]
}
```

Pour chaque modèle et chaque patients, trois résultats sont possibles (Figure 8) :

- prédiction NODAT
- prédiction non NODAT
- prédiction impossible (N/A)Seuls les modèles n'ayant aucune prédiction fausse ont été pris en compte. En pratique, ont été éliminés les modèles prédisant comme « NODAT » des patients témoins, ainsi que les modèles prédisant comme « non NODAT » des patients cas. Les modèles comportant des patients impossibles à prédire (N/A) ont cependant été acceptés. Nous avons attribué un poids aux modèles sans aucune prédiction fausse, calculé selon le nombre de prédictions correctes effectuées par le modèle : de 1 (1 patient correctement prédit et 13 patients avec prédictions impossibles) à 14 (14 patients correctement prédits et 0 prédictions N/A).

Les gènes ayant certains variants utilisés pour créer un ou plusieurs bons modèles sont ceux dont le haut degré de variabilité n'est pas complètement dû au hasard. Un score a été appliqué aux gènes eux-mêmes dans cette optique :

- L'implication d'un gène dans un modèle donné est calculée à partir du nombre de variants de ce gène utilisés dans le modèle rapportés au nombre total de variants du modèle.
- L'importance générale de chaque gène est estimée par la moyenne de leur implication par modèle, pondérée par le score du modèle.

I.2.2. Résultats

I.2.2.1. Séquençage

Tableau 5 : Couverture des exomes : Profondeur < 20. Profondeur entre 20 et 50, Couverture et Profondeur moyenne sur l'exome.

Patient	Identifiant	Prof <20	Prof 20-50	Prof >50	Couverture	Moyenne
Cas 1	EP 391	17,60 %	17,00 %	65,40 %	82,40 %	140
Témoin 1	EP 261	11,40 %	15,40 %	73,20 %	88,60 %	142
Cas 2	EP 384	13,20 %	20,60 %	66,10 %	86,80 %	101
Témoin 2	EP 357	6,00 %	9,00 %	85,00 %	94,00 %	203
Cas 3	EP 392	9,30 %	16,10 %	74,50 %	90,70 %	122
Témoin 3	EP 430	9,50 %	17,20 %	73,20 %	90,50 %	113
Cas 4	EP 328	9,90 %	16,90 %	73,20 %	90,10 %	117
Témoin 4	EP 224	12,50 %	16,10 %	71,40 %	87,50 %	122
Cas 5	EP 238	11,60 %	18,80 %	69,60 %	88,40 %	107
Témoin 5	EP 220	9,20 %	14,40 %	76,40 %	90,80 %	140
Cas 6	EP 500	15,20 %	23,50 %	61,30 %	84,80 %	86
Témoin 6	EP 267	14,30 %	23,50 %	62,20 %	85,70 %	87
Cas 7	EP 467	7,50 %	13,60 %	78,90 %	92,50 %	140
Témoin 7	EP 414	7,60 %	12,60 %	79,80 %	92,40 %	163
Cas 8	EP 262	8,90 %	14,80 %	76,30 %	91,10 %	129
Témoin 8	EP 239	8,30 %	12,80 %	78,90 %	91,70 %	164
Cas 9	EP 265	13,90 %	20,00 %	66,00 %	86,10 %	106
Témoin 9	EP 252	15,50 %	23,50 %	61,00 %	84,50 %	88
Cas 10	EP 510	9,00 %	16,00 %	75,00 %	91,00 %	120
Témoin 10	EP 375	7,50 %	12,10 %	80,30 %	92,50 %	158
Cas 11	EP 397	12,50 %	20,40 %	67,10 %	87,50 %	99
Témoin 11	EP 390	21,80 %	20,70 %	57,50 %	78,20 %	81
Cas 12	EP 326	8,70 %	13,50 %	77,80 %	91,30 %	154
Témoin 12	EP 305	11,20 %	17,40 %	71,30 %	88,80 %	120
Cas 13	EP 364	14,70 %	17,50 %	67,80 %	85,30 %	116
Témoin 13	EP 324	8,60 %	13,50 %	77,90 %	91,40 %	147
Cas 14	EP 365	7,90 %	14,40 %	77,70 %	92,10 %	132
Témoin 14	EP 338	10,80 %	19,70 %	69,40 %	89,20 %	98
Cas 15	EP 340	17,20 %	25,30 %	57,50 %	82,80 %	80
Témoin 15	EP 347	16,00 %	23,50 %	60,50 %	84,00 %	90
Cas 16	EP 288	7,70 %	11,60 %	80,70 %	92,30 %	165
Témoin 16	EP 215	13,70 %	21,30 %	65,00 %	86,30 %	94
Cas 17	EP 504	13,10 %	19,40 %	67,50 %	86,90 %	111
Témoin 17	EP 246	10,90 %	17,40 %	71,70 %	89,10 %	117
Cas 18	EP 490	9,10 %	15,20 %	75,70 %	90,90 %	132
Témoin 18	EP 454	10,40 %	16,30 %	73,40 %	89,60 %	122
Cas 19	EP 274	9,80 %	14,20 %	75,90 %	90,20 %	149
Témoin 19	EP 176	12,10 %	16,60 %	71,30 %	87,90 %	131
Cas 20	EP 361	10,20 %	13,70 %	76,10 %	89,80 %	168
Témoin 20	EP 310	13,40 %	20,60 %	66,00 %	86,60 %	101
Cas 21	EP 462	11,30 %	17,90 %	70,70 %	88,70 %	116
Témoin 21	EP 401	11,40 %	17,20 %	71,40 %	88,60 %	126

Les exomes des 42 patients ont été séquencés avec des couvertures allant de 78,21 % à 94,04 % de l'exome, et des profondeurs moyennes allant de 79 à 203 lectures.

L'appel de variants a détecté 1 655 991 variants répartis sur 17 976 gènes.

I.2.2.2. Recherche de gènes d'intérêt

I.2.2.2.1. Sélection des gènes

La méthode Elbow a permis de retenir k=3 comme nombre de classes optimal pour les 2 matrices (cas seulement ou témoins seulement). Dans les deux cas, les projections ACP ont permis de distinguer :

- Une classe bien définie comportant très peu de gènes à très haute variabilité : 15 pour les cas, 16 pour les témoins. Parmi eux :
    - 15 gènes à très haute variabilité se retrouvent à la fois chez les cas et chez les témoins, laissant supposer que ces gènes sont simplement naturellement variables.
    - Le gène OR9G1 (ENSG00000174914) est très variable chez les témoins mais pas chez les cas, il pourrait donc être d'intérêt.
- Une classe comportant un grand nombre de gènes à très faible variabilité : 13919 pour les cas (77,4 % des gènes) et 13864 pour les témoins (77,1%), Les gènes de la classe sont très regroupés, mais la frontière avec la troisième classe est mal définie.
- Une classe intermédiaire de gènes à variabilité plus haute que ceux de la seconde classe : 4029 gènes pour les cas et 4093 gènes pour les témoins. Parmi eux :
    - 3679 gènes sont communs aux classes intermédiaires des cas et des témoins, supposant là encore que ces gènes sont naturellement variables.
    - 350 gènes ne se retrouvent que dans la classe intermédiaire des cas. L'un d'entre eux est OR9G1, ce qui met en doute son intérêt.
    - 414 gènes ne se retrouvent que dans la classe intermédiaire des témoins.

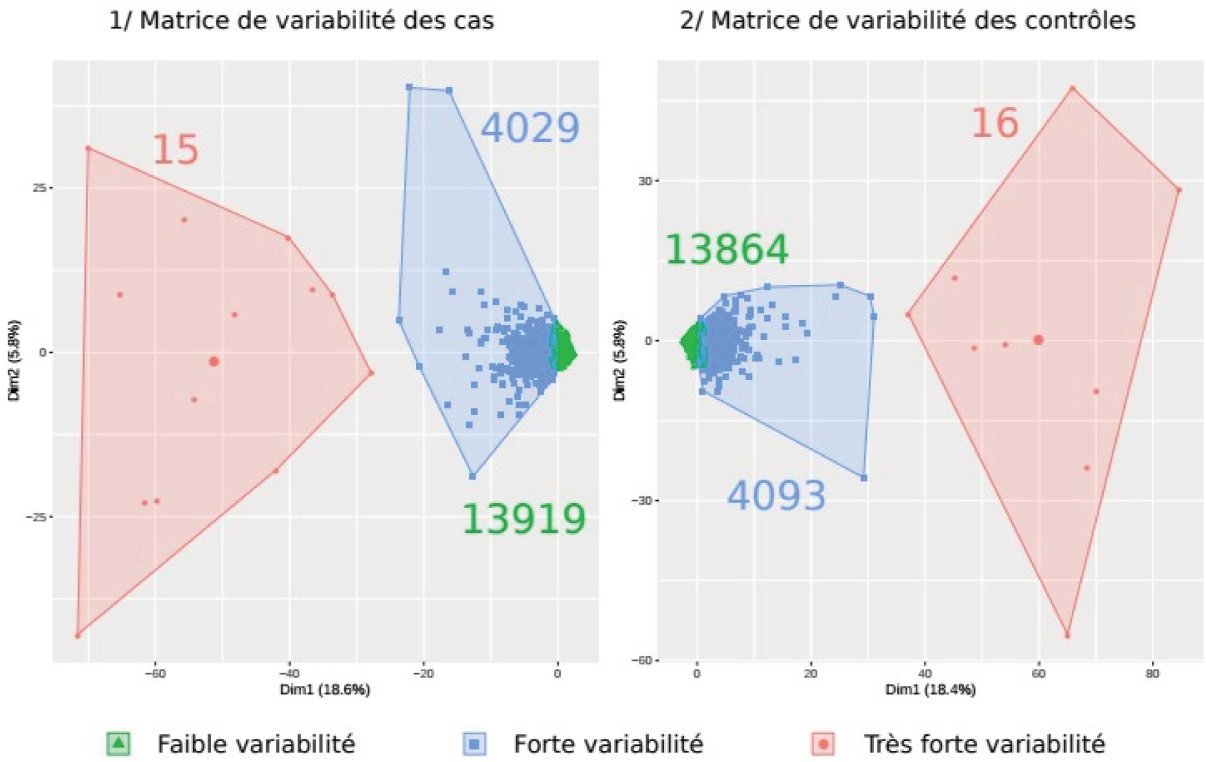

Figure 9: Projection ACP du degré de variation des gènes. Les gènes des matrices cas et témoins ont été regroupés en classe selon leur variabilité. 1/ Matrice de variation des gènes chez les cas NODAT, 2/ Matrice de variation des gènes chez les témoins.

Puisque les frontières des classes intermédiaires et à basse variabilité sont mal définies, nous avons choisi de ne prendre en compte que les gènes dont la variabilité était supérieure à la variabilité moyenne du gène le plus variable de la classe de basse variabilité, et ce pour au moins 15 des 21 patients du groupe étudié (cas ou témoin). L'algorithme de classification utilisé étant basé sur les moyennes, ce filtre est cohérent quoique particulièrement stringent.

Au final, 23 gènes ont été sélectionnés pour les cas, et 17 pour les témoins.

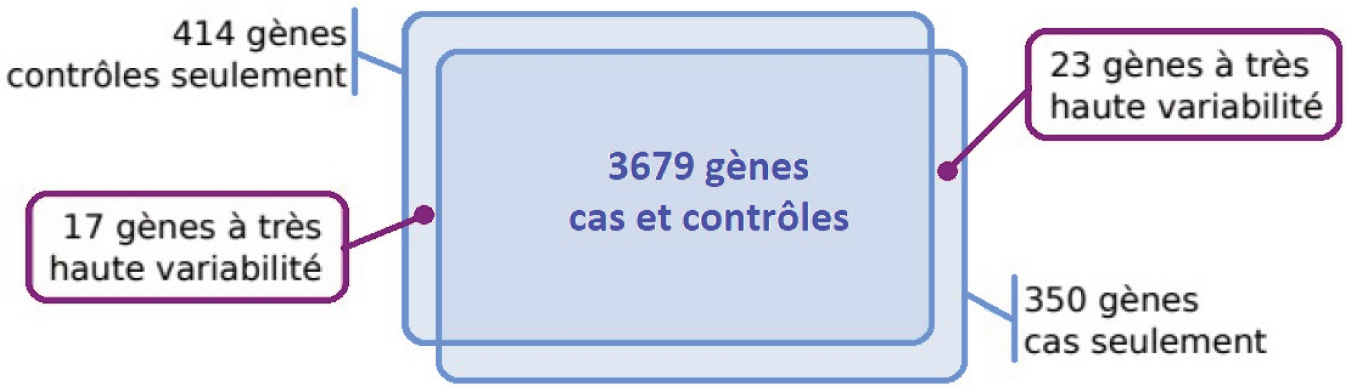

Figure 10: Diagramme de Venn des gènes sélectionnés par classification.

Tableau 6 : 23 gènes de la classe de haute variabilité chez les cas mais non chez les témoins, et avec une variabilité supérieure à la variabilité moyenne du gène le plus variable de la classe de basse variabilité chez au moins 15 des 21 cas

Gène	Identifiant Ensembl !
ADRM1	ENSG00000130706
C11orf1	ENSG00000137720
CLDN7	ENSG00000181885
FAM110A	ENSG00000125898
GOLPH3L	ENSG00000143457
HES6	ENSG00000144485
HIST1H2BC	ENSG00000180596
HIST1H4L	ENSG00000198558
HNRNPA1	ENSG00000135486
HNRNPF	ENSG00000169813
KRTAP19-4	ENSG00000186967
MAPK8IP1	ENSG00000121653
OR9G1	ENSG00000174914
PAK7	ENSG00000101349
RAI2	ENSG00000131831
RERGL	ENSG00000111404
SALL4	ENSG00000101115
SERPIND1	ENSG00000099937
SERPINH1	ENSG00000149257
SLC12A7	ENSG00000113504
SLC7A4	ENSG00000099960
TCF20	ENSG00000100207
UBA52	ENSG00000221983

Tableau 7 : 17 gènes de la classe de haute variabilité chez les témoins mais non chez les cas, et avec une variabilité supérieure à la variabilité moyenne du gène le plus variable de la classe de basse variabilité chez au moins 15 des 21 témoins

Gène	Identifiant Ensembl !
BMP6	ENSG00000153162
CLC	ENSG00000105205
EME1	ENSG00000154920
FABP9	ENSG00000205186
FHL2	ENSG00000115641
HIRIP3	ENSG00000149929
KLRC4-KLRK1	ENSG00000255819
KRTAP4-12	ENSG00000213416
OR4F15	ENSG00000182854
PCK2	ENSG00000100889
RDH8	ENSG00000080511
SELV	ENSG00000186838
SH2D3A	ENSG00000125731
SPRR1B	ENSG00000169469
SUMO4	ENSG00000177688
TPRG1	ENSG00000188001
ZNF823	ENSG00000197933

I.2.2.2.2. Cohérence des résultats

11628 modèles ont été créés par sPLS-DA, en utilisant les 532 variants situés sur les 23 gènes sélectionnés pour les cas, ainsi que les 298 variants situés sur les 17 gènes sélectionnés pour les témoins, pour un total de 830 variants.

139 modèles ont prédit au moins 1 patient correctement mais n'ont prédit aucun patient incorrectement, et sont donc considérés comme acceptables. Parmi ces modèles, les prédictions correctes sont comprises entre 1 et 9 patients. Seuls 6 modèles ont prédit au moins la moitié des patients : 2 modèles en ont prédit 7, 1 modèle en a prédit 8, et 3 modèles en ont prédit 9.

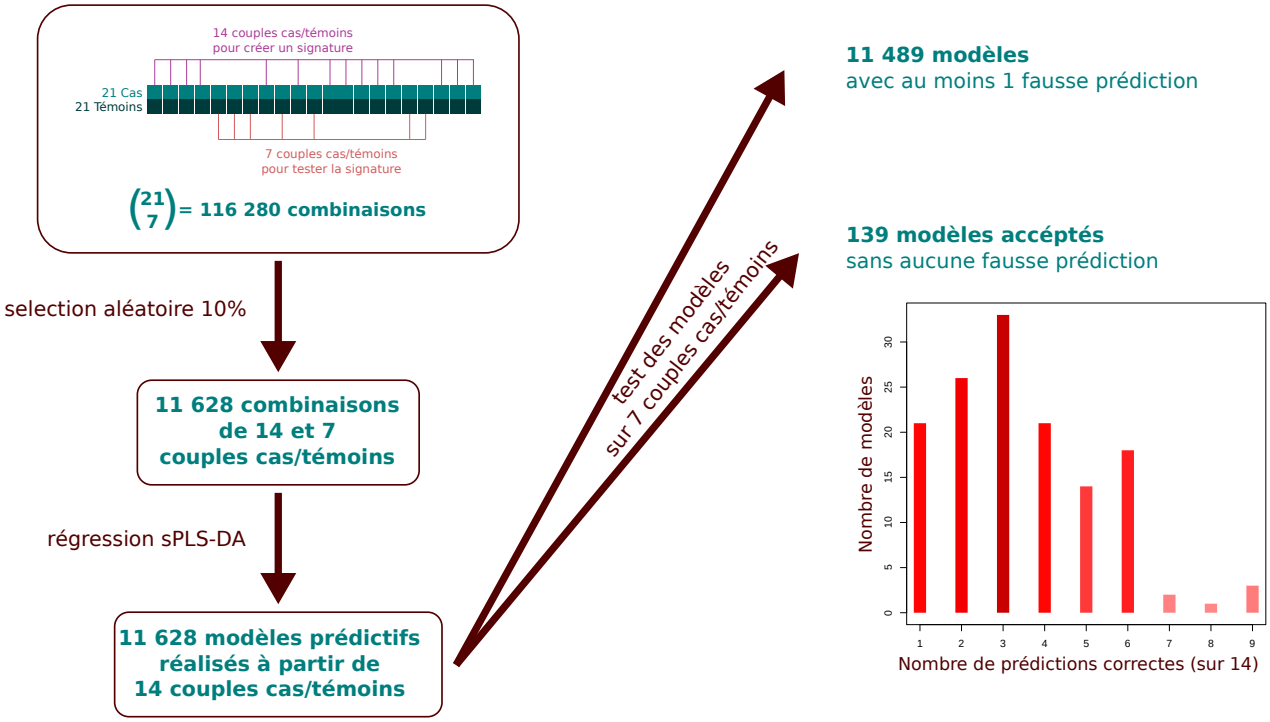

Figure 11: Il y a 116280 combinaisons de 7 et 21 couples cas/témoins. 10% sont aléatoirement sélectionnés pour réaliser des modèles par sPLS-DA à partir des 14 couples d'entraînement. Parmi ces modèles, 11489 comportent au moins 1 prédiction fausse et ne seront pas utilisés, tandis que 139 ne prédisent aucun des 14 patient tests incorrectement.

39 gènes ont été impliqué dans la création des 139 'bons' modèles. 23 ont été sélectionné par classification en utilisant la matrices de variabilité des cas, et 16 ont été sélectionné par classification en utilisant la matrices de variabilité des témoins.

L'implication de chaque gène a été calculée en fonction de leur nombre de variants, et du nombre de prédictions correctes des modèles.

Parmi ces gènes, l'implication dans les modèles varie. 1 seul gène (ADRM1) représente 12,23 %, 4 autres gènes entre 5 % et 8 %, 25 gènes entre 1 % et 5 %, et enfin 9 gènes à moins de 1 %.

Tableau 8 : 39 gènes sélectionnés, avec implications dans les modèles prédictifs

Identifiant Ensembl !	Gène	Implication	Matrice utilisée pour la classification
ENSG00000130706	ADRM1	12,23 %	Cas
ENSG00000125898	FAM110A	7,11 %	Cas
ENSG00000113504	SLC12A7	6,76 %	Cas
ENSG00000174914	OR9G1	5,63 %	Cas
ENSG00000101115	SALL4	5,06 %	Cas
ENSG00000137720	C11orf1	4,78 %	Cas
ENSG00000100207	TCF20	4,61 %	Cas
ENSG00000101349	PAK7	4,43 %	Cas
ENSG00000121653	MAPK8IP1	4,41 %	Cas
ENSG00000181885	CLDN7	4,20 %	Cas
ENSG00000180596	HIST1H2BC	3,89 %	Cas
ENSG00000111404	RERGL	3,26 %	Cas
ENSG00000149257	SERPINH1	2,90 %	Cas
ENSG00000154920	EME1	2,77 %	Témoins
ENSG00000125731	SH2D3A	2,39 %	Témoins
ENSG00000221983	UBA52	2,15 %	Cas
ENSG00000099960	SLC7A4	2,08 %	Cas
ENSG00000143457	GOLPH3L	1,68 %	Cas
ENSG00000149929	HIRIP3	1,64 %	Témoins
ENSG00000135486	HNRNPA1	1,55 %	Cas
ENSG00000186967	KRTAP19-4	1,42 %	Cas
ENSG00000099937	SERPIND1	1,42 %	Cas
ENSG00000105205	CLC	1,36 %	Témoins
ENSG00000131831	RAI2	1,32 %	Cas
ENSG00000188001	TPRG1	1,31 %	Témoins
ENSG00000255819	KLRC4-KLRK1	1,23 %	Témoins
ENSG00000100889	PCK2	1,23 %	Témoins
ENSG00000186838	SELV	1,13 %	Témoins
ENSG00000198558	HIST1H4L	1,06 %	Cas
ENSG00000197933	ZNF823	1,01 %	Témoins
ENSG00000153162	BMP6	0,95 %	Témoins
ENSG00000115641	FHL2	0,81 %	Témoins
ENSG00000080511	RDH8	0,64 %	Témoins
ENSG00000144485	HES6	0,48 %	Cas
ENSG00000169813	HNRNPF	0,41 %	Cas
ENSG00000169469	SPRR1B	0,40 %	Témoins
ENSG00000177688	SUMO4	0,15 %	Témoins
ENSG00000205186	FABP9	0,11 %	Témoins
ENSG00000213416	KRTAP4-12	0,05 %	Témoins

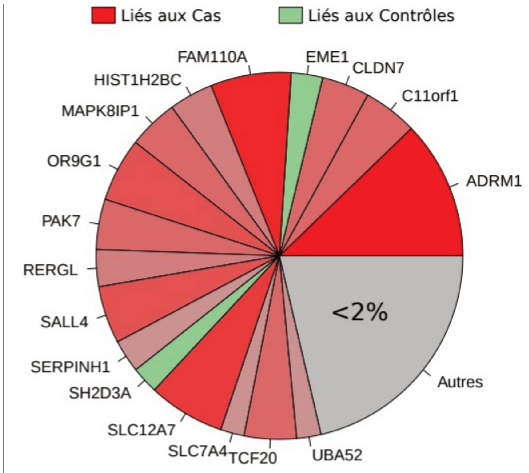

Figure 12: Gènes utilisés dans la création des modèles de risque NODAT. Les 22 gènes avec une implication <2 % du total ont été regroupés.

I.2.2.2.3. Gènes sélectionnés

- Nous nous sommes seulement intéressés aux 17 gènes représentant plus de 2 % des variants impliqués dans les modèles prédictifs. 15 d'entre eux sont liés aux cas tandis que seuls 2 sont liés aux témoins.
- ADRM1 est un récepteur du protéasome de l'ubiquitine fréquemment amplifié dans les cancer ovariens et du colon conjointement avec une modulation des cellules myéloïdes suppressives (ex monocytes ou macrophages) dans les TME des tumeurs ovariennes et causant une immunosuppression[69]. ADRM1 est la cible de la molécule RA190, un inhibiteur du protéasome utilisé en recherche pour inverser cette immunosuppression[70].
- Les fonctions de famille FAM110 à laquelle appartient FAM110A sont peu connues.
- SLC12A7 est un transporteur de K+ et Cl-, entre autre dans les reins. D'autres membres de la famille SLC12 sont impliqués dans la réabsorption du glucose, mais cela ne semble pas être le cas de SLC12A7[7172].
- OR9G1 est un récepteur olfactif.
- SALL4 est un facteur de transcription identifié comme pouvant être impliqué dans la différenciation de cellules souche productrices d'insuline au stade embryonnaire[73].
- Le gène C11orf1 est non caractérisé à ce jour.
- TCF20 est un facteur de transcription.
- PAK7 est une kinase impliquée dans la potentialisation des acides gras de la

sécrétion d'insuline[74] et la métabolisation du glucose[75], et le développement précoce de diabète de type II[76].

- MAPK81P1 est un régulateur de la fonction des cellules bêta du pancréas, supposé contribuer au diabète de type II[77].
- CLDN7 est une protéine transmembranaire connue notamment pour moduler l'homéostasie de Cl- et Na+ dans les cellules rénales[78]. Elle y régule aussi l'expression du gène WNK4[7879], qui a été associé au diabète de type II.
- HIST1H2BC est une histone.
- RERGL est une protéine liée à l'activité GTP.
- SERPINH1 est un inhibiteur de sérine protéase lié à la bio-synthèse du collagène. Le gène pourrait être impliqué dans la résistance à l'insuline[80].
- EME1 est une partie d'un complexe d'endonucléase méiotique.
- SH2D3A est une protéine liée à l'activité SH3. Il s'agit aussi d'un paralogue du gène BCAR3, qui a été lié au phénotype du diabète dans une étude GWAS[81].
- UBA52 est une ubiquitine qui a été liés au diabète dans plusieurs modèles animaux[82].
- SLC7A4 est un transporteur d'acides aminés cationiques.

I.2.3. Conclusion du projet GENODAT

Dans cette étude, en utilisant une approche sans *a priori* (exome entier) et en regroupant des patients transplantés rénaux avec un NODAT par rapport à des témoins appariés, nous avons identifié plusieurs gènes dont la séquence nucléotidique varie fortement dans les cas mais non dans les témoins.

Plusieurs de ces gènes sont liés à des voies métaboliques d'intérêt. Par exemple ADMR1 est le gène le plus impliqué dans les modèles, ainsi que la cible d'un traitement cherchant à inverser une immunosuppression induite par des TME tumoraux.

D'autres gènes sont connus pour être liés aux voix métaboliques du diabète : SALL4 est lié à l'insuline, CLDN7 est lié au diabète de type II tout comme MAPK81P1 et PAK7 aux deux. D'autres gènes ont des liens avec la physiopathologie du diabète plus incertains : SERPINH1 à l'insuline, UBA52 et SH2D3A au diabète.

Le transporteur SLC12A7 pourrait n'avoir qu'un lien faible avec le transport de glucose, alors que le transporteur SLC7A4 ne semble avoir aucun lien avec aucune voie métabolique significative concernant les NODAT[7172]. Il en est de même pour les 7 autres gènes, bien que

FAM110A soit le deuxième gène le plus impliqué dans les modèles et OR9G1 le quatrième.

Concernant la méthode, le projet GENODAT a utilisé pour la plupart le protocole de traitement bio-informatique développé pour le projet PTLD, avec quelques ajustements nécessaires dus à la taille différente de la cohorte.

D'une part, nous avons pu prendre en compte tous les variants plutôt que d'ignorer ceux présents à la fois chez les cas et chez les témoins car le plus grand nombre de patients permettait de réduire le bruit. Cela a eu pour conséquence de passer de 2 à 3 classes de variabilité, avec des gènes naturellement très variables sans liens à NODAT ou aux ICN dans une classe à part.

Secondement, les frontières entre des classes de variabilité faible et intermédiaire sont mal définies, là où elles étaient très claires pour PTLD. Cela s'explique en partie par les gènes qui variant beaucoup plus que le reste du génome chez un seul patient, qui sont regroupés avec les gènes variants légèrement plus que le reste du génome chez de multiples patients. Lors du projet PTLD, 1 seul patient très variable sur 7 suffisait pour placer ces gènes dans la classe à haute variabilité. Les 21 patients de GENODAT permettent d'atténuer cette sur-représentation, classant ainsi les gènes concernés dans la classe à faible variabilité.

La création de modèles de régression a permis de confirmer que la variabilité de certains gènes n'est pas due au hasard. Le choix de n'effectuer que 1/15 des 116 280 modèles possibles se justifie *a priori* par le grand nombre de modèles impliqués et la sélection aléatoires des couples de patients d'entraînement et de test.

Il est intéressant de constater que les gènes sélectionnés par classification de la matrice des cas sont plus utilisés par les modèles que ceux sélectionnés par classification de la matrice des témoins.

Attribuer un score aux gènes, à la fois selon le nombre de modèle les utilisant et selon leur implication dans chaque modèle, fournit une mesure permettant de pondérer la sélection des gènes. Cela aidera à choisir quels gènes étudier plus en profondeur sur une cohorte plus large, afin de confirmer les résultats.

I.3. Discussion

La petite taille des cohortes ne nous permettait pas de rechercher avec de bonne sensibilité et spécificité des variants pouvant être liés à la survenue de PTLD ou bien de NODAT, d'où le choix d'une approche par gène en calculant les degrés de variation. Cette approche a permis de prendre en compte chaque variant sans leur attribuer trop de poids.

Le degré de variation des gènes, sur lequel se base les deux études, a été mis au point lors de l'étude PTLD, initialement sans aucun poids. Cela a cependant conduit à une surreprésentation des variant introniques. Ils ont donc été pondéré, plusieurs poids ayant été testé.

Pour commencer, le poids de chaque variant a été établi en fonction de scores de pathogénicité[8384]. Cependant, de tels scores n'étaient disponibles que pour une partie des variants seulement. De plus, l'intérêt des variants en PGx n'est pas forcement lié à leur pathogénicité.

De même, la possibilité de pondérer les variants selon leur fréquence allélique MAF[85] a été écartée sur la base que de nombreux variants n'ont aucune MAF connue.

Le poids a donc été déterminé selon la localisation. Le poids des variants exoniques a été fixé à 1. Différents poids pour les variants introniques ont été testés : 0,5, 0,1, 0,01 et 0,001. Les variants introniques étaient surreprésentés lorsque pondérés à 0,5, tandis que les pondérations à 0,01 et 0,001 donnaient des résultats de classification similaires à ceux ne prenant pas les variants introniques en compte. Leur poids a donc été fixé à 0,1.

La méthode de classification a permis de regrouper les gènes en fonction de leur variabilité, en faisant l'hypothèse que les gènes inhabituellement variables au regard de l'ensemble du génome seraient relativement peu nombreux. Il est toutefois connu que certains gènes sont naturellement plus variables que d'autres, raison pour laquelle nous avons différencié les gènes variables chez les cas et ceux variables chez les témoins, en ne conservant que ceux ayant une forte variabilité dans seulement un des deux groupes. Que dans chaque cas la plus grande partie des gènes appartiennent à la classe peu variable justifie notre hypothèse.

Il serait possible de calculer un degré de variation « de base » pour les gènes sélectionnés (ou pour le génome entier) à partir des bases de données publiques (telles que gnomAD[85]). Une difficulté rencontrée serait la disparité des couvertures entre les différentes études utilisées pour créer les bases de données. Cela pourrait cependant être atténué en prenant en compte les MAF des variants. Une autre difficulté serait qu'un degré de variation de base soit calculé à partir de personnes n'ayant jamais été des transplantés rénaux traités par ICN. Il serait raisonnable de supposer que dans le cas hypothétique d'une telle transplantation,

une partie d'entre eux seraient susceptibles de développer PTLD ou NODAT, avec des prévalences similaires à celles reportées. Cela compliquerait la comparaison entre les degrés de variation tirés des bases de données publiques, et ceux des patients transplantés à la fois cas et témoins. Une méthode d'évaluation pourrait être cependant établie en utilisant les prévalence PTLD et NODAT, dont la mise au point nécessiterait un projet séparé.

Les modèles de risque variants-signatures NODAT ont été utiles pour confirmer que le haut degré de variation des gènes sélectionnés par classification n'était pas du au hasard. Cependant, les cohortes utilisées pour la régression et les tests (n=12+4 pour PTLD, n=14+28 pour GENODAT) étaient trop petites pour réellement créer des modèles prédictifs utilisables dépassent le cadre de ces études. La création de modèles par sPLS-DA pourrait cependant être utilisée sur de plus grandes cohortes afin de créer de modèles prédictifs.

Chapitre II. Régulation des gènes

Bien que les variations des séquences nucléotidiques de l'ADN soient une source i d'hétérogénéité phénotypique interindividuelle, la variabilité d'expression des gènes a également un rôle important. Elle est notamment influencée par des variables environnementales : on parle alors de régulation épigénétique, notamment par le biais de processus comme l'acétylation ou bien la méthylation.

Des enzymes de la famille des ADN (DNA) méthyl-transférases (DNMT) sont capables d'ajouter des groupements méthyles à certaines bases cytosines au début d'un dimère CG. Il s'agit d'une modification épigénétique réversible mais transmissible lors de la division cellulaire (mitose ou méiose). La relation entre méthylation de l'ADN et expression peut être complexe, mais la méthylation provoque généralement une diminution de la transcription de l'ADN en ARN messagers, et donc une baisse d'expression. Cependant, certains gènes semblent au contraire dépendre de la méthylation pour s'exprimer, et parfois la méthylation d'un gène n'influe pas sur son expression[86].

Parmi les facteurs pouvant causer des modifications épigénétiques, certains médicaments peuvent interagir avec les enzymes DNMT ou bien avec les voies de signalisation épigénétiques, provoquant la méthylation de l'ADN[87]. Cela peut être le but du médicament, on parle alors de « médicament épigénétique »[88], mais il peut aussi s'agir d'un effet secondaire, qu'il soit recherché ou indésirable[89].

L'effet des ICN sur la méthylation a jusqu'ici été peu étudiée, mais pourrait influer sur la variabilité des effets indésirables, ou bien en être le marqueur. Dans ce contexte, une modification de l'expression génique selon la prise d'ICN pourrait confirmer un effet sur la méthylation. Le projet EPIPHITe a pour but d'analyser les possibles modifications d'expression et de méthylation chez la souris après prise d'ICN.

II.1. Protocole expérimental

Nous avons pu bénéficier d'une partie des données produites au cours de la thèse de Lucie Pouché, « Variabilité d'origine génétique et épigénétique de la pharmacodynamie des ICN en transplantation rénale » (2016). Les données concernaient les niveaux d'expression et de méthylation des gènes de souris soumises aux traitements ICN. Nous avons utilisé les prélèvements restant après l'étude précédente pour lesquels il restait suffisamment de matériel génétique pour une nouvelle analyse.

II.1.1. Expérimentation animale

45 souris mâles de la lignée congénique C57BL/6J ont reçu une injection intra-péritonéale quotidienne d'ICN ou de solvant.

Les anesthésiques ont été dilués dans du sérum physiologique (Chlorure de Sodium 0,9%; Baxter) à partir de solutions mères de kétamine (Imalgene® 500, Centravet) (50mg/ml) et de xylazine (Rompun® 2%, Centravet) (20mg/ml). Les souris ont été séparées en 3 groupes selon le type d'injection :

- Groupe CSA : ciclosporin à 30mg/kg/jour, préparée à partir d'une solution mère à 50mg/ml (Sandimmune®, Injection, Novartis), d'éthanol 100%, de Cremophor EL® (Calbiochem®) (huile de ricin) et d'un soluté glucosé à 5%.
- Groupe TAC : tacrolimus à 2mg/kg/jour, préparée à partir d'une poudre (tacrolimus 99% 100mg, VWR®), d'éthanol à 100%, de Cremophor EL® et d'un soluté glucosé à 5%.
- Groupe CTR : injection d'un mélange d'éthanol 100%, de Cremophor EL® et de glucose à 5%.

A l'issue de la période de traitement, les souris ont été anesthésié avant sacrifice et prélèvement des organes.

Chaque groupe contenait 5 lots de 3 souris, sacrifiées après 1, 8, 28, 63 et 83 jours de traitement (anesthésie par solutions de kétamine à 100mg/kg et xylazine à 10mg/kg, puis sang total prélevé par ponction intracardiaque).

La rate a également été prélevée, nettoyée (1ml de PBS) puis broyée (2ml de RPMI 1640 Medium, GlutaMAX TM Supplement, Life Technologies TM) et filtrée (70µm). La solution cellulaire obtenue a été centrifugée (1500 rpm pendant 10 min à +4°C).

Le sang total et la solution cellulaire de rate ont permis d'isoler les lymphocytes T auxiliaires (LTCD4), selon le protocole détaillé en Annexe 1.

Les ADN des LTCD4 (ainsi que ceux du foie des souris) ont été extraits (Thèse de Lucie Pouché).

II.1.1.1. Séquençage d'ADN méthylé

L'ADN méthylé a été immunoprécipité par des anticorps reconnaissant la marque méthylée ou hydroxyméthylée, selon le protocole MeDip (Methylated Immunoprécipitation). L'ADN a été fragmenté mécaniquement puis incubé avec l'anticorps, avant d'être précipité à l'aide de billes magnétiques couplées à l'anticorps. L'ADN méthylé a ensuite été séquencé.

Le protocole MeDip-Seq complet, développé pour l'Ion ProtonTM System (Life Technologies) de la plateforme GénoLim puis validé par qPCR à partir de la publication de Corley[90] et de la thèse de Sengenès[91], est disponible en Annexe 2.

Le séquençage a généré des fichiers BAM, qui ont été alignés sur le génome de référence mm8 (*mus musculus*) avec l'outil BWA.

II.1.1.2. Mesures d'expression

L'amplification des ARN et leur hybridation sur des puces à ADN ont été effectuées sur un système Affymetrix Gene Atlas® avec la puce Affymetrix® Mouse Gene 2.1 ST Array suivant le protocole fournit par ThermoFisher.

La puce a généré des fichiers au format CEL.

II.1.2. Données analysées

Les données auxquelles nous avons eu accès concernent :

- L'expression en ADN des cellules LTCD4 à J28 : 3 échantillons TAC, 3 échantillons CSA, 3 échantillons CTR
- La méthylation de l'ADN des cellules LTCD4 à J28 : 3 échantillons TAC, 3 échantillons CSA, 3 échantillons CTR
- La méthylation des cellules LTCD4 à J83 : 3 échantillons TAC, 3 échantillons CSA, 3 échantillons CTR

II.2. Analyse bio-informatique

II.2.1. Normalisation des données d'expression

Les fichiers CEL ont été exploités en utilisant le package R oligo[92], et l'expression relative de chaque sonde a été extraite.

La normalisation des valeurs brutes extraites a été effectuée en deux étapes. D'abord, le bruit de fond a été corrigé avec la méthode Robust Multi-array Analysis (RMA)[93], puis l'expression de l'ensemble des sondes a été normalisé par quantile. Pour chaque profil, l'homogénéité des variances, la normalité de distribution et l'écart par rapport à la normalité des valeurs maximales et minimales (test d'Anderson-Darling) ont été vérifiés.

Une fois les données d'expression normalisée, une valeur d'expression a été calculée pour chaque gène, comme la somme des valeurs d'expression de toutes les sondes ciblant le même gène. Les sondes des puces d'expression étant dupliquées, la valeur d'expression d'une sonde a été calculée comme la moyenne des expressions de toutes les sondes identiques, pondérée par la variance entre les échantillons.

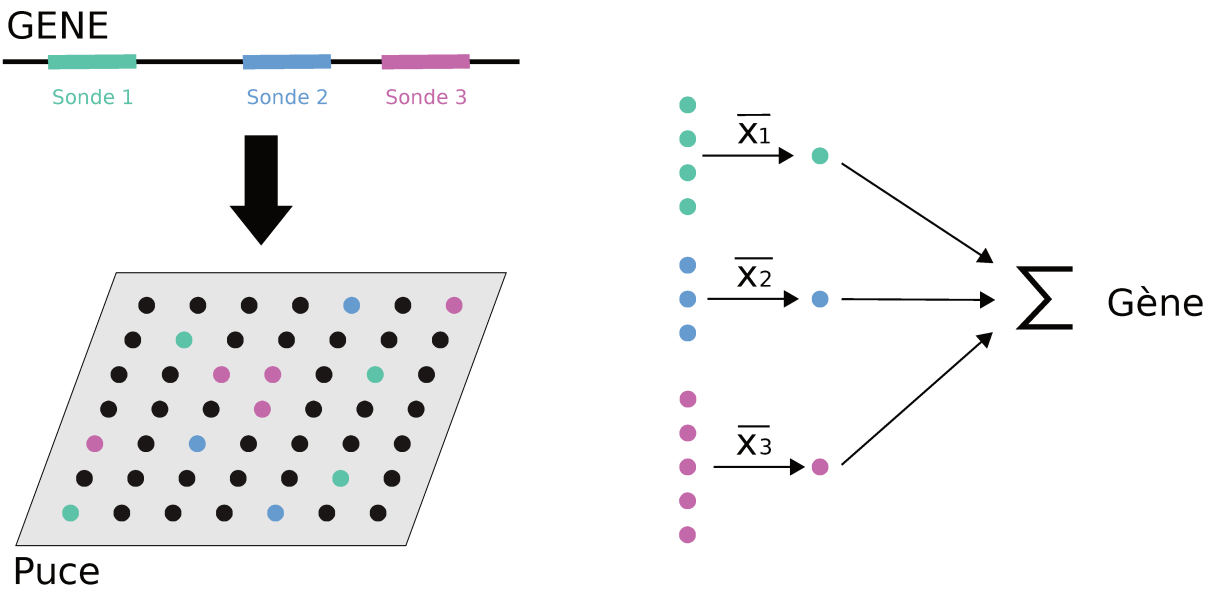

Figure 13: Puce d'expression. Plusieurs sondes ciblent le même gène. Chaque sonde est dupliquée sur la puce et rend une valeur d'expression. La valeur d'expression de la séquence ciblée par une sonde est la moyenne des expression des sondes. La valeur d'expression du gène est la somme des valeurs d'expression de toutes ses sondes.

Script 6 : Normalisation des profils d'expression

```
#lire le fichier CEL

sondrawData = read.celfiles(celFiles, experimentData = experimentData)

sonde = pm(rawData)

sonde = as.data.frame.matrix(sonde)e_bgn = sonde

#Supprimer le bruit

sonde_bgn[,c(ctr, csa, tac)] = backgroundCorrect(as.matrix(sonde[,c(ctr, csa, tac)]),
method = "rma")

#Normalisation

sonde_bgn[,c(ctr, csa, tac)] = normalize(as.matrix(sonde_bgn[,c(ctr, csa, tac)]), method =
"quantile")

#Regrouper les valeurs de chaque sonde sur plusieur echantillons de même groupe

sonde_bgn$ctr = apply(sonde_bgn[,ctr], 1, mean)

sonde_bgn$csa = apply(sonde_bgn[,csa], 1, mean)

sonde_bgn$tac = apply(sonde_bgn[,tac], 1, mean)

#Variation des valeurs des sondes sur les echantillons de même groupe

sonde_bgn$ctr_sd = apply(sonde_bgn[,ctr], 1, sd) / sonde_bgn$ctr

sonde_bgn$csa_sd = apply(sonde_bgn[,csa], 1, sd) / sonde_bgn$csa

sonde_bgn$tac_sd = apply(sonde_bgn[,tac], 1, sd) / sonde_bgn$tac

#Une seule valeur par id

res = as.data.frame(table(sonde_bgn$id))

colnames(res) = c("id", "total_probes")

#pour chaque sonde, moyenne de tout les échantillons de même groupe, pondéré par sa
variance entre échantillon

sonde_bgn$ctr_pond = ifelse(sonde_bgn$ctr_sd == 0, sonde_bgn$ctr, sonde_bgn$ctr *
sonde_bgn$ctr_sd)

sonde_bgn$csa_pond = ifelse(sonde_bgn$csa_sd == 0, sonde_bgn$csa, sonde_bgn$csa *
sonde_bgn$csa_sd)

sonde_bgn$tac_pond = ifelse(sonde_bgn$tac_sd == 0, sonde_bgn$tac, sonde_bgn$tac *
sonde_bgn$tac_sd)

  #CTR

temp = aggregate(sonde_bgn$ctr_pond, by = list(sonde_bgn[, 'id']), sum)#Pour chaque id,
somme de toute les valeurs pondérées...

temp_n = aggregate(sonde_bgn$ctr_sd, by = list(sonde_bgn[, 'id']), sum)#...et somme des
variances (qui servent de poids)

temp = merge(temp, temp_n, by = 1)

temp$ctr = temp$x.x/temp$x.y#diviser la somme des valeurs pondérées par la somme des poids

temp = temp[, c("Group.1", "ctr")]

res = merge(res, temp, by.x = 'id', by.y = 'Group.1')

    #CSA

temp = aggregate(sonde_bgn$csa_pond, by = list(sonde_bgn[, 'id']), sum)

temp_n = aggregate(sonde_bgn$csa_sd, by = list(sonde_bgn[, 'id']), sum)

temp = merge(temp, temp_n, by = 1)

temp$csa = temp$x.x/temp$x.y

temp = temp[, c("Group.1", "csa")]
```

```
res = merge(res, temp, by.x = 'id', by.y = 'Group.1')
  #TAC
temp = aggregate(sonde_bgn$tac_pond, by = list(sonde_bgn[, 'id']), sum)
temp_n = aggregate(sonde_bgn$tac_sd, by = list(sonde_bgn[, 'id']), sum)
temp = merge(temp, temp_n, by = 1)
temp$tac = temp$x.x/temp$x.y
temp = temp[, c("Group.1", "tac")]
res = merge(res, temp, by.x = 'id', by.y = 'Group.1')
```

II.2.2. Calcul des valeurs de méthylation

Pour chaque échantillon, l'exome séquencé a été divisé en zones de 10pb. Pour chaque zone, le nombre de séquences d'ADN méthylé alignées a été extrait par l'outil BEDTool. Pour chaque gène séquencé, une valeur de méthylation a été calculée comme l'aire sous la courbe du profil de nombre de séquences (Figure 14).

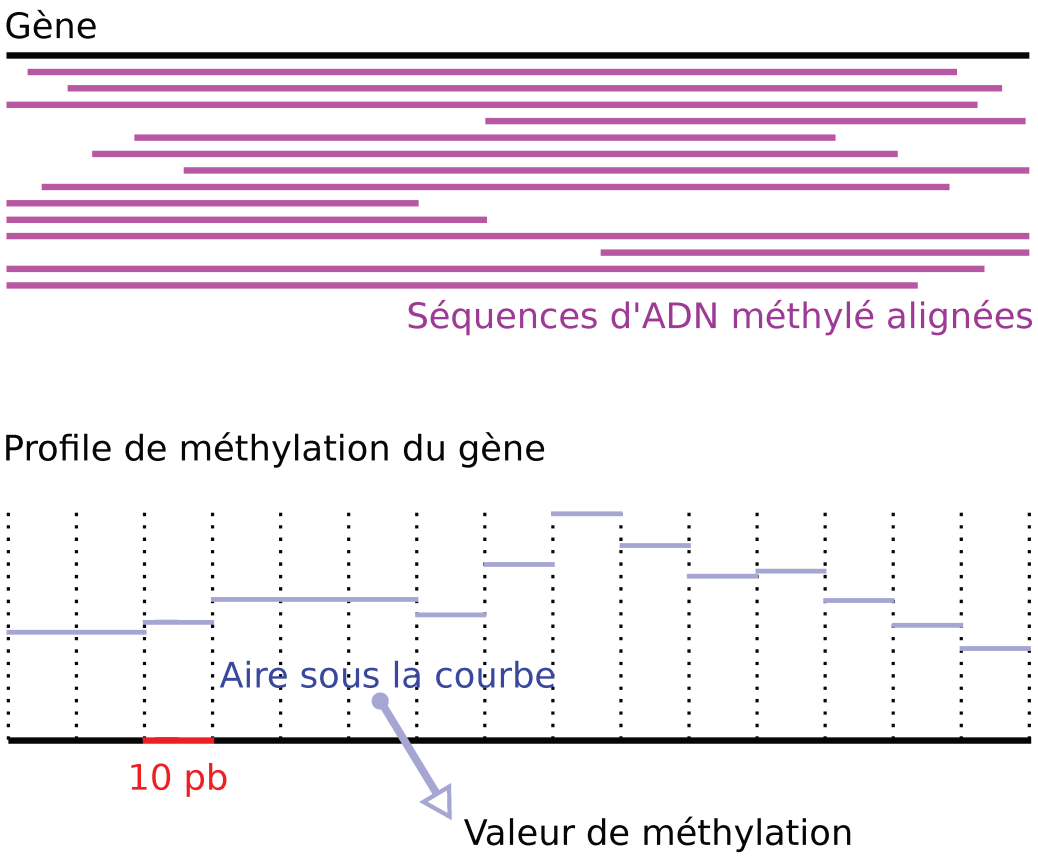

Figure 14: Calcul de la valeur de méthylation à partir du profil de profondeurs de lectures.

Pour chaque groupe de souris, le coefficient de variation de la valeur de méthylation de chaque gène a été calculé. La valeur moyenne de la méthylation d'un gène pour le groupe a été calculée lorsque le coefficient de variation était inférieur à 0,1. Dans le cas contraire, les

données étaient considérées comme non concluante et le gène non analysé.

Cela a créé pour chaque groupe un profil de méthylation, qui a été normalisé par quantile.

Script 7 : Création des profils de méthylation

```
#csa, liste des echantillons CSA
#tac, liste des echantillons TAC
#ctr, liste des echantillons CTR
exons = data.frame(exon = c(), gene = c(), transcrit = c(), taille = c(), ctr = c(), csa = c(), tac = c(), chr = c())
data = read.table(paste("raw_cov_medip", chr, sep = ''), header = T)
temp = do.call(rbind, strsplit(as.character(data$name),'_'))
data$gene = temp[, 1]
data$transcrit = temp[, 2]
data$exon = paste(temp[, 1], temp[, 2], temp[, 3], sep = '_')
#Moyenne par groupe
data$ctr = apply(data[, ctr], 1, mean)
data$csa = apply(data[, csa], 1, mean)
data$tac = apply(data[, tac], 1, mean)
#Aire sous la courbe
data$taille = data$end - data$start
data$ctr = data$ctr*data$taille
data$csa = data$csa*data$taille
data$tac = data$tac*data$taille
data_exon = data[, c('gene', 'transcrit', 'exon')]
data_exon = data_exon[!duplicated(data_exon),]
#taille
temp = aggregate(data$taille, by = list(data$exon), sum)
colnames(temp) = c('exon', 'taille')
data_exon = merge(data_exon, temp)
  #ctr
temp = aggregate(data$ctr, by = list(data$exon), sum)
colnames(temp) = c('exon', 'ctr')
data_exon = merge(data_exon, temp)
  #csa
temp = aggregate(data$csa, by = list(data$exon), sum)
colnames(temp) = c('exon', 'csa')
data_exon = merge(data_exon, temp)
  #tac
temp = aggregate(data$tac, by = list(data$exon), sum)
colnames(temp) = c('exon', 'tac')
data_exon = merge(data_exon, temp)
#Normalisation quantile
```

```
data_exon = data_exon[apply(data_exon[,c('ctr', 'csa', 'tac')], 1, min) > 0, ]
data_exon[,c('ctr', 'csa', 'tac')] = normalize.quantiles(as.matrix(data_exon[,c('ctr', 'csa', 'tac')]))
```

II.2.3. Différences de profils selon les conditions

Pour un même jour, nous avons comparés les différences entre les valeurs relatives d'expression ou de méthylation d'un groupe de souris traitées par ICN (groupe TAC ou groupe CSA) contre celles du groupe de souris CTR.

Tout d'abord, nous avons calculé ces différences, comme log2 du ratio des valeurs, nommé *delta* (Δ).

Puis nous avons cherché à savoir quels Δ étaient significatifs. Plutôt que de fixer un seuil au-delà duquel un Δ serait considéré comme « bon », nous avons cherché à déterminer si un Δ était aberrant par rapport à l'ensemble des autres *deltas*.

Pour cela, nous avons testés plusieurs outils mathématiques et statistiques.

II.2.3.1. Outils statistiques

- **Le critère de Chauvernet (William Chauvenet, 1863)**

Ce critère est la probabilité d'obtenir un delta particulier au vu de l'ensemble des autres valeurs de delta. Si le critère est < 0,5, on peut considérer la valeur comme aberrante. Il s'agit d'une approche stochastique utilisant la fonction d'erreur complémentaire (erfc). Nous lavons utilisé le critère récursivement, de *delta* en *delta* :

- Tri décroissant des *delta* en fonction de leur valeur absolue.
- Tant qu'il reste des *delta* :
  1. Calcul du critère de Chauvernet pour le *delta* le plus grand :

$$n * erfc\left(\frac{|\Delta_{max} - \bar{\Delta}|}{\sigma}\right)$$

  2. Suppression du *delta* le plus grand, dont le critère a été calculé

- **Le test de Grubbs (Frank E. *Grubbs, 1969)***

Le test se base sur la normalité de la distribution des *delta*. Tout comme pour les critère de Chauvernet, seul le *delta* le plus grand en valeur absolu est calculé, ainsi nous avons utilisé ce test récursivement.

Le test compare le résidu normalisé de Δ (noté G) à un critère utilisant la table de Student

(noté $G_{théorique}$) : $G=\frac{|\Delta-\bar{\Delta}|}{\sigma}$ et $G_{théorique}=\frac{n-1}{\sqrt{n}}*\sqrt{\frac{t^2}{n-2+t^2}}$. avec t la fonction quantile associée à la distribution de Student avec n-2 degrés de liberté et un niveau de significativité de α/n.

Si $G > G_{théorique}$, alors Δ est aberrant. Nous pouvons donc établir pour chaque Δ la p-value correspondant à son aberrance.

- **Test de Thompson (David J. Thompson, 1950)**

Une autre méthode déterminant l'aberrance du Δ à la valeur absolue la plus grande, le test de Thompson se base une fois encore sur la normalité de la distribution. Il est très similaire au test de Grubbs en cela qu'il compare le résidu normalisé du Δ le plus grand à une valeur théorique τ, calculée de la même manière que $G_{théorique.}$ Cependant, τ n'est calculé qu'une seule fois pour l'ensemble des Δ, rendant le test moins stringeant.

- **L'algorithme de Pierce (Benjamin Peirce, 1877)**

Le critère de Pierce trouve l'ensemble des Δ aberrants sans reposer sur des p-values. Il détermine comme aberrants tous les Δ_i tels que $|\Delta_i-\bar{\Delta}|>\sigma*R$, R étant une valeur tirée de la table de Pierce en fonction du nombre de valeurs aberrantes suspectées. Nous avons répété l'algorithme tant qu'il trouvait de nouveaux Δ aberrants, en retirant à chaque fois de l'ensemble les Δ aberrants précédemment détectés.

- **Le test Khi2 (Wilfrid Joseph Dixon, 1950)**

Un autre test qui calcule la p-value selon laquelle la valeur la plus éloignée de la moyenne serait aberrante, basé sur la distribution du khi carré des différences au carré entre les deltas et la moyenne. Le calcul de cette p-value a été effectué récursivement.

II.2.3.2. Calcul de scores d'aberrance

Afin de comparer les différents tests, nous les avons utilisés pour calculer des scores d'aberrance.

- Pour les tests utilisant des p-values, les score ont été calculés selon un gradient linéaire allant de 0 pour une p-value ≥ 0,05 jusqu'à 1 pour une p-value de 0.
- Le critère de Chauvernet représentant la plausibilité du Δ, son score a été calculé comme 1 – le critère.
- Le score de Pierce a été calculé à partir du nombre d'itérations nécessaires pour

qu'un Δ soit considéré comme aberrant, noté k : $1-\frac{(k-1)}{max(k)}$.

II.2.3.3. Sélection des gènes

Le score d'aberrance reflète la significativité de la différence entre l'expression (ou la méthylation) des gènes des souris du groupe CTR et du groupe traité par l'ICN (CSA ou TAC) au même jour.

Nous avons considéré comme d'intérêt les gènes dont à la fois la méthylation et l'expression à J28 sont aberrants.

II.3. Résultats

II.3.1. Normalisation des données d'expression

Les données d'expression à J28 incluaient 3 souris du groupe CTR, 2 souris du groupe CSA et 3 souris du groupe TAC.

41 345 sondes différentes hybridées entre 4 et 11 fois chacune (total de 770 069 mesures d'expression par puce) ont permis d'évaluer l'expression de 30 382 gènes. La normalisation des données pour les différentes souris et conditions est présentée en Figure 15. Elle consistait en une première étape de suppression du bruit de fond, suivie d'une normalisation quantile. Les données ont ensuite été moyennées par groupe.

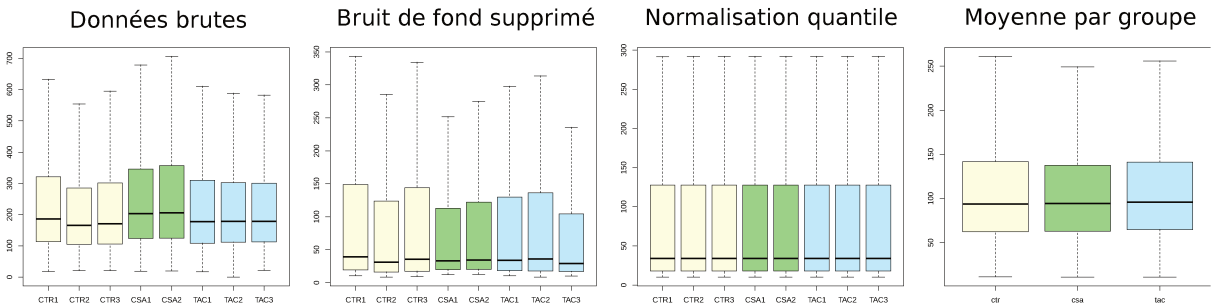

Figure 15: Normalisation des données d'expression. Répartitions des expressions relatives de chaque gène, pour chaque étape du processus de pré-traitement. Répartitions données pour chaque souris pour les trois premières étapes : données brutes, après la suppression du bruit de fond, après normalisation. Répartitions données pour chaque groupe après pré-traitement.

II.3.2. Calcul des valeurs de méthylation

Les données de méthylation à J28 incluaient les 2 souris du groupe CTR, 3 souris du groupe CSA et 3 souris du groupe TAC. Le calcul des profondeurs de lecture a permis de calculer les valeur de méthylation de 35431 gènes.

Les données de méthylation à J83 incluaient les 2 souris du groupe CTR, 2 souris du groupe CSA et 3 souris du groupe TAC. Le calcul des profondeurs de lecture a permis de calculer les valeur de méthylation de 34445 gènes.

II.3.3. Différences de profils selon les conditions

Les cinq mesures d'aberrances des delta CSA/CTR et TAC/CTR ont été calculées pour chaque gène à J28 (expression et méthylation) et à J83 (méthylation).

La comparaison de ces critères a montré que le critère le plus stringent est celui de Grubbs, suivi de Chi2, de Pierce, de Chauvernet et enfin de Thompson (Figure 16).

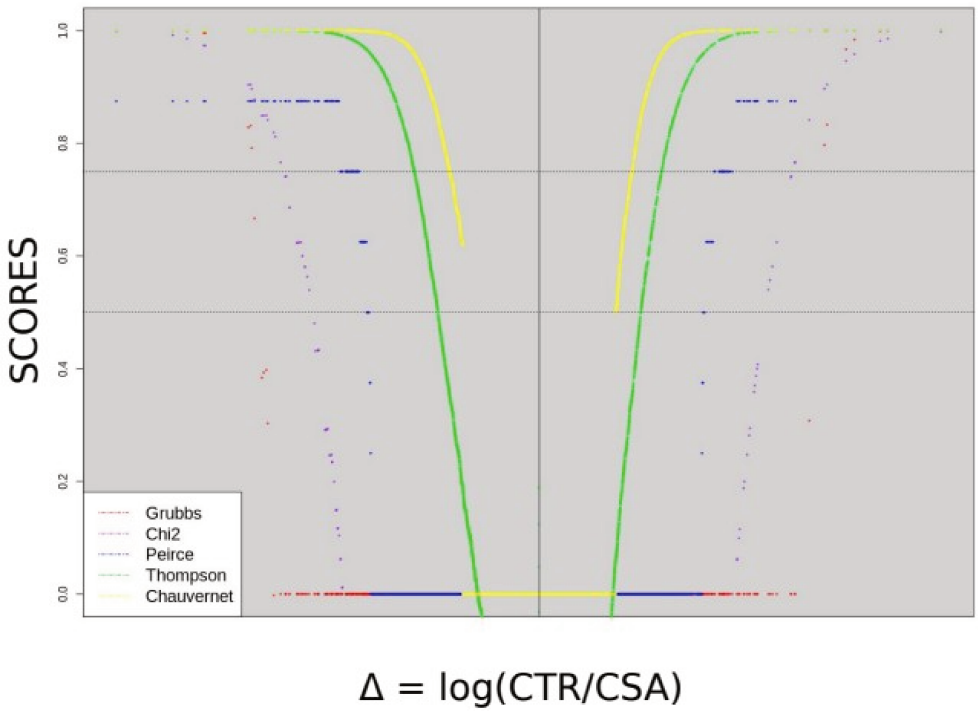

Figure 16: Comparaison des p-values des tests de Grubbs, Chi2 et Thomson ainsi que du critère de Pierce. Exemple de la différence d'expression des groupes CSA et CTR à J28.

Tous les critères suivent cependant une tendance similaire, et se corrèlent bien lorsque les *delta* augmentent. Le critère à utiliser reflétera donc le type de seuil souhaité, à l'exception possible du critère de Pierce, qui est moins précis.

Nous avons choisi d'utiliser comme critère de sélection un score de Thompson ≥ 0,9 pour l'expression à J28 et la méthylation à J83. Le critère reflète bien la distance d'un *delta* à l'ensemble des valeurs, comme l'illustrent les figures 17 et 18.

Nous n'avons pas utilisé de score pour filtrer la méthylation à J83, mais nous n'avons considéré que les gènes dont les potentielles différences de méthylation allaient dans le même sens à J28 et à J83.

II.3.3.1. Gènes d'intérêt détectés chez les souris traitées par ciclosporine

5 gènes montrent une baisse d'expression à J28 qui s'accompagne d'une hausse de méthylation à J28 poursuivie à 283 :

- l'histone Hist1h2ac
- Ighj3 se situant sur le segment J des chaînes d'immunoglobuline[94]
- Vmn2r106, un récepteur vomeronasal
- Maz, un facteur de transcription associé au doigts de zinc.

- Cd40lg est une cytokine exprimée à la surface des lymphocytes T qui régule la fonction des cellules B[95]. Il stimule la prolifération des cellules T et la production de cytokines en générant un signal de stimulation qui améliore la production d'IL4 et d'IL10[96]. Il induit également l'activation de NF-kappa-B et des kinases MAPK8 et PAK2 dans les lymphocytes T[97].

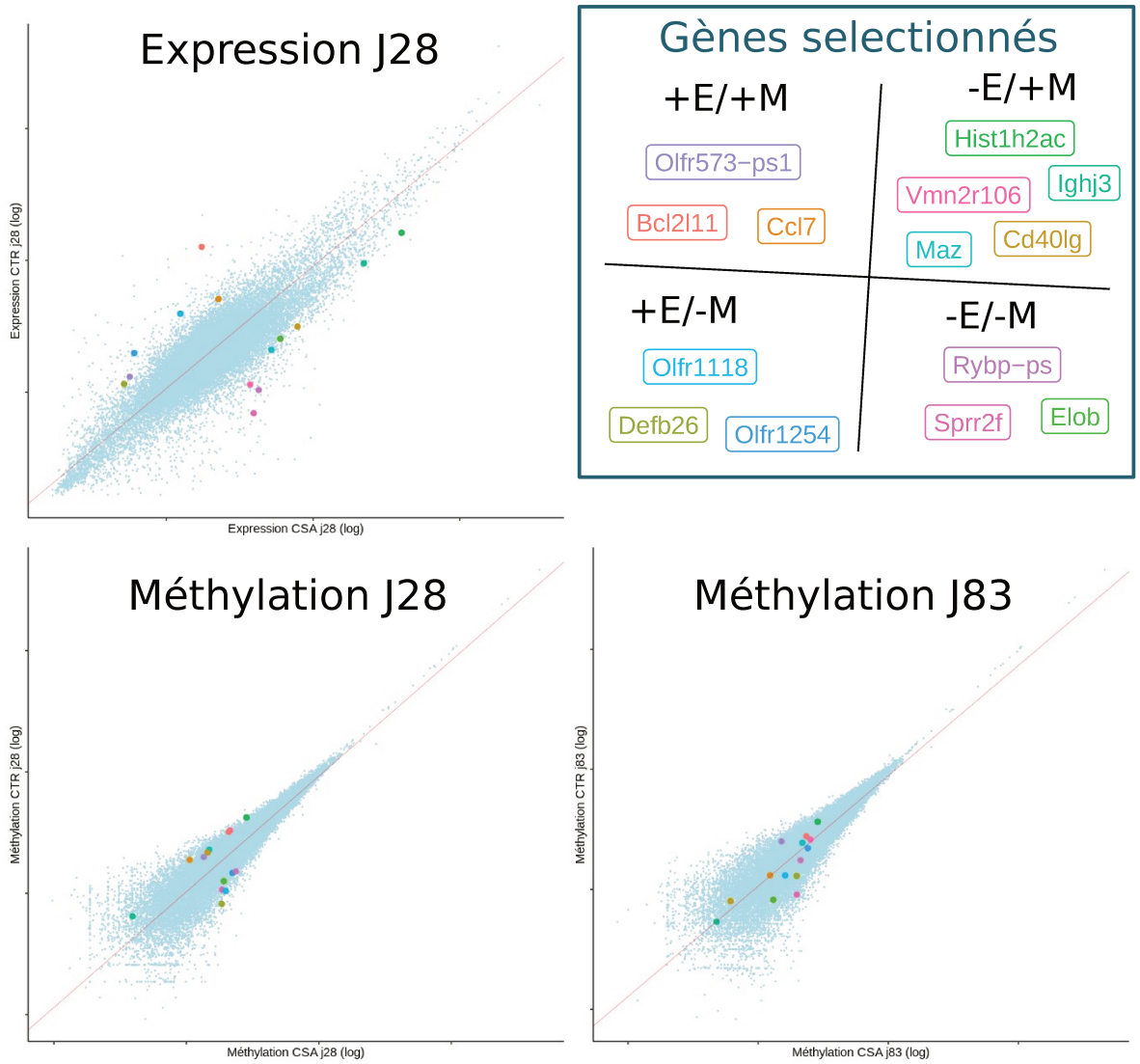

Figure 17: Valeur des profils de gènes, souris du groupe témoins par rapport aux souris traitées par ciclosporine (en log) : expression des gènes à J28, méthylation à J28, et méthylation à J83. Gènes sélectionnés en fonction des delta aberrants d'expression et de méthylation à J28, avec la différence de méthylation à J83 dans le même sens que celle à J28.

3 gènes montrent une hausse d'expression à J28 qui s'accompagne d'une baisse de méthylation à J28 poursuivie à 283 :

- Les récepteurs olfactifs Olfr1118 et Olfr1254

- La bêta défensine Defb26

3 gènes montrent une baisse d'expression à J28 qui s'accompagne d'une baisse de méthylation à J28 poursuivie à 283 :

- Sprr2f lié au développement de la kératine.
- Elob, un facteur de transcription SIII.
- Rybp-ps, un composant d'un complexe requis pour maintenir l'état de répression de la transcription de nombreux gènes via le remodelage de la chromatine et la modification des histones, spécialement H2A[98]. Il joue notamment un rôle dans la régulation de l'apoptose[99100], de la croissance tumorale et des métastases[101102].

Enfin, 3 gènes montrent une hausse d'expression à J28 qui s'accompagne d'une hausse de méthylation à J28 poursuivie à 283 :

- Le récepteur olfactif Olfr573-ps1
- Ccl7, codant la protéine chimiotactique monocyte 3, une chimiokine qui attire les macrophages lors d'inflammation et de métastases[103].

Bcl2l11, dont l'expression peut être induite par le facteur de croissance nerveuse et le facteur de transcription à la fourche FKHR-L1, suggérant un rôle dans l'apoptose neuronale et lymphocytaire. Il pourrait également contrôler l'activation des cellules T[104].

Tableau 9 : Gènes d'intérêt CSA, avec les delta, la différence d'expression ou de méthylation, et la score selon la méthode de Thompson.

Identifiant Ensembl !	Gène	Expression J28			Méthylation J28			Méthylation J83		
		Δ	↑↓	Score	Δ	↑↓	Score	Δ	↑↓	Score
Hausse de méthylation et baisse d'expression										
ENSMUSG00000031132	Cd40lg	-0,79	↓2,2	0,940	0,89	↑2,4	0,915	0,57	↑1,8	0,644
ENSMUSG00000069270	Hist1h2ac	-0,79	↓2,2	0,939	0,86	↑2,4	0,906	0,53	↑1,7	0,607
ENSMUSG00000076619	Ighj3	-0,74	↓2,1	0,911	1,09	↑3,0	0,961	0,25	↑1,3	0,316
ENSMUSG00000030678	Maz	-0,79	↓2,2	0,939	0,93	↑2,5	0,930	0,24	↑1,3	0,316
ENSMUSG00000091656	Vmn2r106	-1,02	↓2,8	0,992	0,94	↑2,6	0,931	0,07	↑1,1	0,208
Baisse de méthylation et hausse d'expression										
ENSMUSG00000074680	Defb26	0,70	↑2,0	0,901	-1,77	↓5,9	0,999	-0,92	↓2,5	0,876
ENSMUSG00000083706	Olfr1118	1,00	↑2,7	0,993	-1,37	↓4,0	0,993	-0,45	↓1,6	0,578
ENSMUSG00000075074	Olfr1254	1,03	↑2,8	0,995	-0,90	↓2,5	0,941	-0,19	↓1,2	0,268
Baisse de méthylation et d'expression										
ENSMUSG00000050635	Sprr2f	-1,51	↓4,5	1,000	-1,20	↓3,3	0,983	-1,70	↓5,5	0,993
ENSMUSG00000055763	Rybp-ps	-1,22	↓3,4	0,999	-0,96	↓2,6	0,954	-0,41	↓1,5	0,517
ENSMUSG00000055839	Elob	-0,74	↓2,1	0,916	-0,91	↓2,5	0,942	-1,02	↓2,8	0,913
Hausse de méthylation et d'expression										
ENSMUSG00000035373	Ccl7	0,70	↑2,0	0,902	1,26	↑3,5	0,981	0,13	↑1,1	0,235
ENSMUSG00000052785	Olfr573-ps1	0,73	↑2,1	0,922	0,84	↑2,3	0,901	1,10	↑3,0	0,916
ENSMUSG00000027381	Bcl2l11	1,72	↑5,6	1,000	0,93	↑2,6	0,930	0,36	↑1,4	0,436

II.3.3.2. Gènes d'intérêt détectés chez les souris traitées par tacrolimus

6 gènes montrent une baisse d'expression à J28 qui s'accompagne d'une hausse de méthylation à J28 poursuivie à J83 :

- Les histones Hist1h1b et Hist1h4k
- Igkv4-91, du segment V du domaine variable de la chaîne légère des immunoglobulines participant à la reconnaissance des antigènes[105].
- Traj9, un récepteur de cellule T reconnaissent les antigènes étrangers.
- Rab5a, un régulateur du trafic membranaire intracellulaire
- Sdhc, sous-unité d'un complexe lié au transport d'électrons

5 gènes montrent une hausse d'expression à J28 qui s'accompagne d'une baisse de méthylation à J28 poursuivie à 283 :

- Le récepteur olfactif Olfr1105
- Lcn3, un transporteur extra-cellulaire de la famille des petites protéines sécrétoires
- Fam3a, une protéine similaire aux cytokines.
- Cyp2c69, un cytochrome P450
- Tdpoz4, une protéine du domaine TD et POZ

4 gènes montrent une baisse d'expression à J28 qui s'accompagne d'une baisse de méthylation à J28 poursuivie à 283 :

- Les récepteurs olfactifs Olfr582 et Olfr694
- Ighv1-18, du segment V du domaine variable de la chaîne légère des immunoglobulines participant à la reconnaissance des antigènes[105].
- S100a4 de la famille S100 contenant 2 sites de fixation du calcium de type « main EF »

Enfin, 3 gènes montrent une hausse d'expression à J28 qui s'accompagne d'une hausse de méthylation à J28 poursuivie à 283 :

- Rpl10l, une protéine similaire au ribosome
- Trav14n-3, un récepteur des cellules T
- Krtap4-1, lié au développement de la kératine

Tableau 10 : Gènes d'intérêt TAC, avec les delta, la différence d'expression ou de méthylation, et la score selon la méthode de Thompson.

Identifiant Ensembl !	Gène	Expression J28			Méthylation J28			Méthylation J83		
		Δ	↑ ↓	Score	Δ	↑ ↓	Score	Δ	↑ ↓	Score
Hausse de méthylation et baisse d'expression										
ENSMUSG00000058773	Hist1h1b	-0,91	↓2,5	0,984	0,83	↑2,3	0,938	0,56	↑1,7	0,665
ENSMUSG00000064288	Hist1h4k	-0,69	↓2,0	0,914	1,24	↑3,5	0,992	1,29	↑3,6	0,971
ENSMUSG00000076532	Igkv4-91	-0,97	↓2,7	0,992	0,84	↑2,3	0,938	1,84	↑6,3	0,997
ENSMUSG00000017831	Rab5a	-0,71	↓2,0	0,922	0,75	↑2,1	0,909	0,50	↑1,7	0,610
ENSMUSG00000058076	Sdhc	-0,70	↓2,0	0,918	0,92	↑2,5	0,958	0,93	↑2,5	0,894
ENSMUSG00000076919	Traj9	-0,96	↓2,6	0,991	0,73	↑2,1	0,904	2,64	↑14,0	1,000
Baisse de méthylation et hausse d'expression										
ENSMUSG00000092008	Cyp2c69	0,98	↑2,7	0,991	-1,12	↓3,1	0,977	-0,24	↓1,3	0,335
ENSMUSG00000031399	Fam3a	1,01	↑2,7	0,993	-1,21	↓3,4	0,986	-0,18	↓1,2	0,247
ENSMUSG00000026936	Lcn3	1,39	↑4,0	1,000	-1,20	↓3,3	0,985	-0,19	↓1,2	0,265
ENSMUSG00000075165	Olfr1105	0,80	↑2,2	0,955	-0,91	↓2,5	0,936	-0,17	↓1,2	0,239
ENSMUSG00000060256	Tdpoz4	0,76	↑2,2	0,941	-1,02	↓2,8	0,961	-0,10	↓1,1	0,179
Baisse de méthylation et d'expression										
ENSMUSG00000076695	Ighv1-18	-0,67	↓2,0	0,902	-0,98	↓2,7	0,955	-0,22	↓1,3	0,308
ENSMUSG00000073961	Olfr582	-0,81	↓2,3	0,961	-1,09	↓3,0	0,973	-1,01	↓2,8	0,925
ENSMUSG00000064223	Olfr694	-0,87	↓2,4	0,978	-0,87	↓2,4	0,922	-0,26	↓1,3	0,347
ENSMUSG00000001020	S100a4	-1,13	↓3,1	0,998	-1,04	↓2,8	0,966	-0,53	↓1,7	0,655
Hausse de méthylation et d'expression										
ENSMUSG00000063251	Krtap4-1	0,81	↑2,2	0,957	1,09	↑3,0	0,982	0,48	↑1,6	0,589
ENSMUSG00000060499	Rpl10l	0,71	↑2,0	0,912	0,83	↑2,3	0,937	0,04	↑1,0	0,175
ENSMUSG00000076824	Trav14n-3	1,51	↑4,6	1,000	0,96	↑2,6	0,965	1,37	↑4,0	0,979

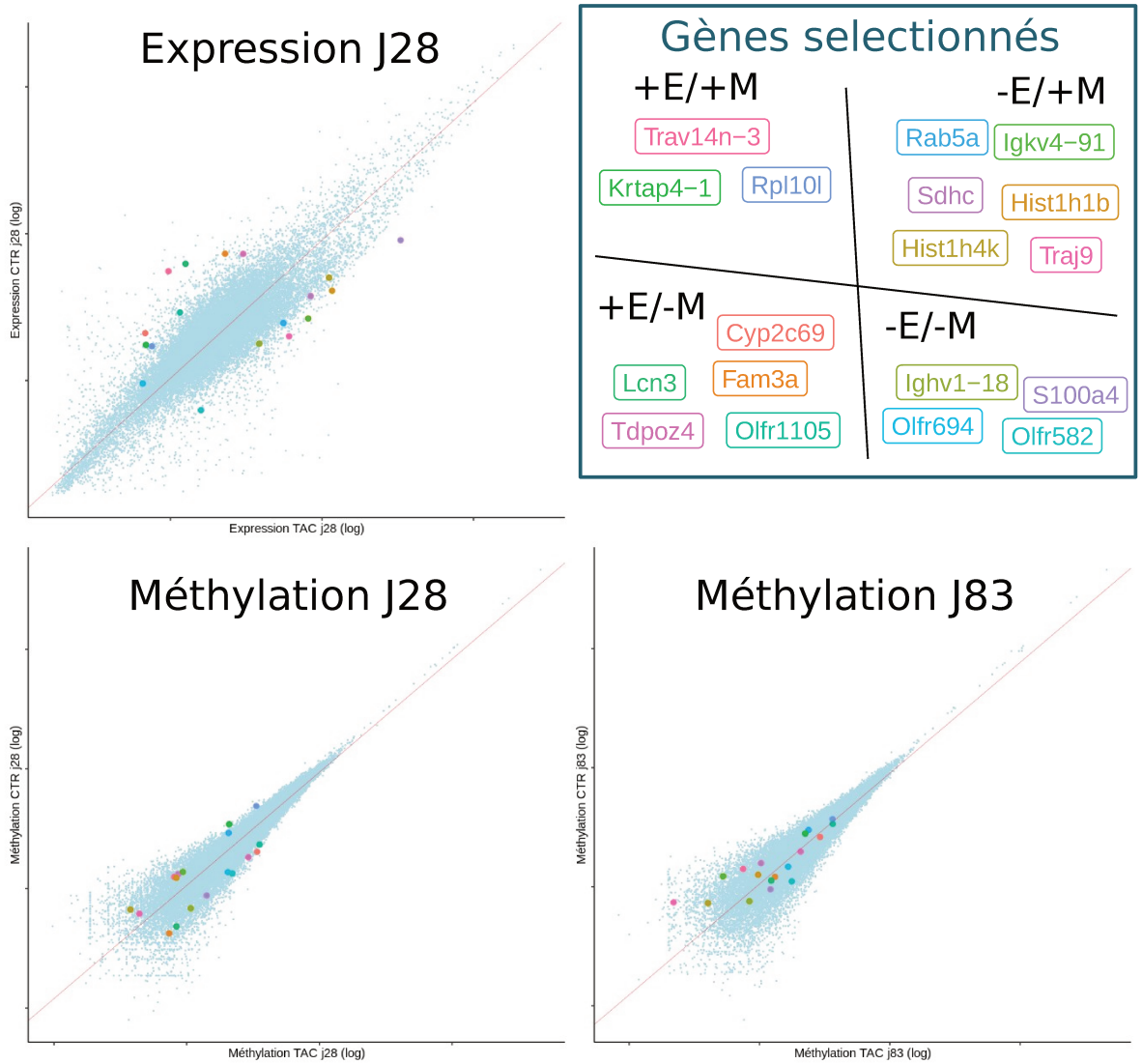

Figure 18: Valeur des profils de gènes, souris du groupe témoins par rapport aux souris traitées par tacrolimus (en log) : expression des gènes à J28, méthylation à J28, et méthylation à J83. Gènes sélectionnés en fonction des delta aberrants d'expression et de méthylation à J28, avec la différence de méthylation à J83 dans le même sens que celle à J28.

II.4. Discussion

Le petit nombre de souris sur lesquelles ce projet a été réalisée illustre une problématique récurrente des études impliquant l'expérimentation animale. En effet, si les techniques de séquençage ont connu une drastique diminution des coûts et des temps d'exécution dans la dernière décennie, les protocoles d'expérimentation animale eux-mêmes restent contraignants.

Les protocoles de normalisation que nous avons utilisé pour créer les profils d'expression et de méthylation de chaque souris sont standards. Cependant, il est courant d'analyser les différences entre plusieurs groupes en appliquant un test statistique (typiquement ANOVA) sur chaque profil simultanément. Bien qu'efficace lorsque le nombre d'échantillon par groupe est grand, un tel test aurait été non applicable avec nos groupes de 3 (ou parfois 2 souris) car ne respectait pas les conditions d'utilisation (normalité des distributions). Nous avons donc utilisé une approche alternative permettant une application aux petits échantillons.

Nous avons créé un seul profil pour chaque groupe, combinant les 2 ou 3 profils du groupe afin d'avoir des résultats plus robustes.

Nous avons utilisé des ratios pour comparer des profils de deux conditions, ce qui a permis de faire ressortir les différences dans les profils. Cela a aussi permis de travailler sur une unique variable de grande taille (1 point de donné par gène), plutôt que d'effectuer des analyse de comparaison sur de très petits échantillons.

La recherche de valeurs aberrantes se justifie par l'hypothèse que la majorité du génome ne serait pas affecté par la prise d'ICN, que ce soit en termes d'expression ou de méthylation. La normalité des distributions des delta confirme en partie cette hypothèse.

Les outils statistiques et mathématiques utilisés pour la recherche de *delta* aberrants sont standards dans ces disciplines, mais peu utilisés en PGx ou dans le domaine médical en général. Leur généralisation pour des jeux de données utilisant plus de variables ont cependant donné lieu à des méthodes novatrices utilisées en PGx, mentionnées dans la revue sur les Big Data en Pharmacogénomique[106]. En plus d'un plus grand nombre de variables, ces nouveaux outils nécessitent tout de même des cohortes plus larges que celles du projet EPIPHITe.

Les scores d'aberrance fournissent une mesure de l'intérêt des gènes sélectionnés, mais il en va de même pour les sens de méthylation et d'expression. De plus, l'interprétation au vu de l'annotation des gènes permet de formuler certaines hypothèses sur les mécanismes impliqués.

II.5. Conclusion

L'étude sans *a priori* de profils d'expressions et de méthylation sur le génome entier de souris a détecté 14 gènes d'intérêts relatifs au traitement par ciclosporine, et 18 gènes d'intérêts relatifs au traitement par tacrolimus. Ces gènes ont été sélectionnés car ils montraient à la fois des différences d'expression et de méthylation significatives après 28 jours de traitement. De plus, si les différences de méthylation après 83 jours de traitement n'étaient pas toujours significatives, elles étaient au moins cohérentes : augmentation continue de J28 à J83, ou bien baisse continue de J28 à J83.

Parmi les gènes détectés, on retrouve 6 régulateurs olfactifs. Étant une des familles de gènes les plus grande du génome, leur implication pourrait être aléatoire.

On retrouve également 3 histones H1, qui agissent en tant que régulateur de la transcription individuelle de gènes par le remodelage de la chromatine, l'espacement des nucléosomes et la méthylation de l'ADN. L'effet d'une sur ou sous-expression de ces histones pourrait influencer la méthylation d'autres gènes, mais cet effet sort du cadre de cette étude. De même pour le gène Rybp-ps et son effet sur la régulation à la fois de l'expression et de la méthylation[98-102].

La prise de ciclosporine s'accompagne d'une baisse d'expression de la cytokine Cd40lg, tandis que la prise de tacrolimus s'accompagne d'une hausse d'expression de la cytokine-like Fam3a. Cela pourrait signifier que l'ICN agit sur la cytokine réelle, tandis que la protéine similaire est sur-activée pour compenser.

Plusieurs gènes liés aux mécanismes du système immunitaire connaissent une hausse d'expression suite à la prise d'ICN. L'expression de Defb26 a augmenté suite à la prise de ciclosporine : il s'agit d'une bêta défensine donc la réponse immunologique aux micro-organismes étrangers pourrait être affectée en réaction à la baisse de lymphocytes. L'expression de Ccl7, lié au traitement des macrophages lors d'inflammation[103], augmente lors de la prise de ciclosporine. Au contraire, le rôle possible de Bcl2l11 dans l'apoptose lymphocytaire[104] suggère que sa hausse d'expression suite à la prise de ciclosporine serait accompagnée d'une diminution accrue des lymphocytes. Puisque la prise d'ICN provoque déjà leur épuisement, les raisons de cette hausse d'expression sont peu claires.

CYP2C69 n'ait pas de liens directs avec le système immunitaire ni avec les mécanismes épigénétiques. Il s'agit cependant d'un enzyme de métabolisation mais qui n'a pas été associée à la biotransformation du tacrolimus. Il est difficile de déterminer l'effet associé à sa hausse d'expression lors de la prise de tacrolimus.

2 gènes du segment V du domaine variable de la chaîne légère des immunoglobulines[105] ont

été détectés suite à leur baisse d'expression lors de la prise de tacrolimus. Celle de Igkv4-91 est accompagnée d'une hausse de méthylation qui s'accroît beaucoup à J83 (6 fois plus chez les TAC que chez les CTR), tandis que celle de Ighv-18 est liée à une baisse d'expression. Dans les deux cas, la baisse d'expression rendrait plus difficile la reconnaissance des anti-gènes.

2 gènes récepteurs de cellules T ont été détectés, mais pour des raisons inverses. La baisse d'expression du récepteur de cellule T Traj9[107] pourrait être un effet secondaire de la diminution du nombre de lymphocytes. Leur épuisement encore plus grand à J83 expliquerait le fait que l'augmentation de la méthylation de Traj9 y est encore plus grande (14 fois plus chez les CSA que chez les CTR). Au contraire, si la hausse d'expression du récepteur de cellule T Trav14n-3 après la prise de tacrolimus sous-entend un lien avec la diminution des lymphocytes[107], le mécanisme est moins clair.

Ces résultats sont bien sûrs à interpréter en considérant le petit nombre de souris étudiées. De plus, l'expression n'a été étudiée qu'à J28, il est donc impossible de conclure sur la pérennité des changements d'expression. Il conviendrait donc de vérifier les résultats avec des analyses complémentaires.

Chapitre III. Interprétation clinique

Les études des pharmacogènes ont permis d'établir des recommandations thérapeutiques concernant la prévention d'effets indésirables médicamenteux ou les doses de médicaments à utiliser. La plupart de ces recommandations repose sur l'utilisation de nomenclature allélique dites « étoile ». Ce mode de nomenclature a été proposé initialement pour les cytochromes P450 par un groupe de travail international. Elle est aujourd'hui largement utilisée pour définir la nomenclature de variations connues et documentées d'autres familles de pharmacogènes.

Cette nomenclature défini l'allèle *1 comme le gène avec une séquence de référence, c'est-à-dire la protéine entièrement fonctionnelle, les autres allèles (* 2, * 3, etc) correspondent à des versions uniques du gène, caractérisées par des combinaisons de variations génétiques pouvant être associée à des modifications de la structure, de l'expression ou de l'activité de la protéine[108].

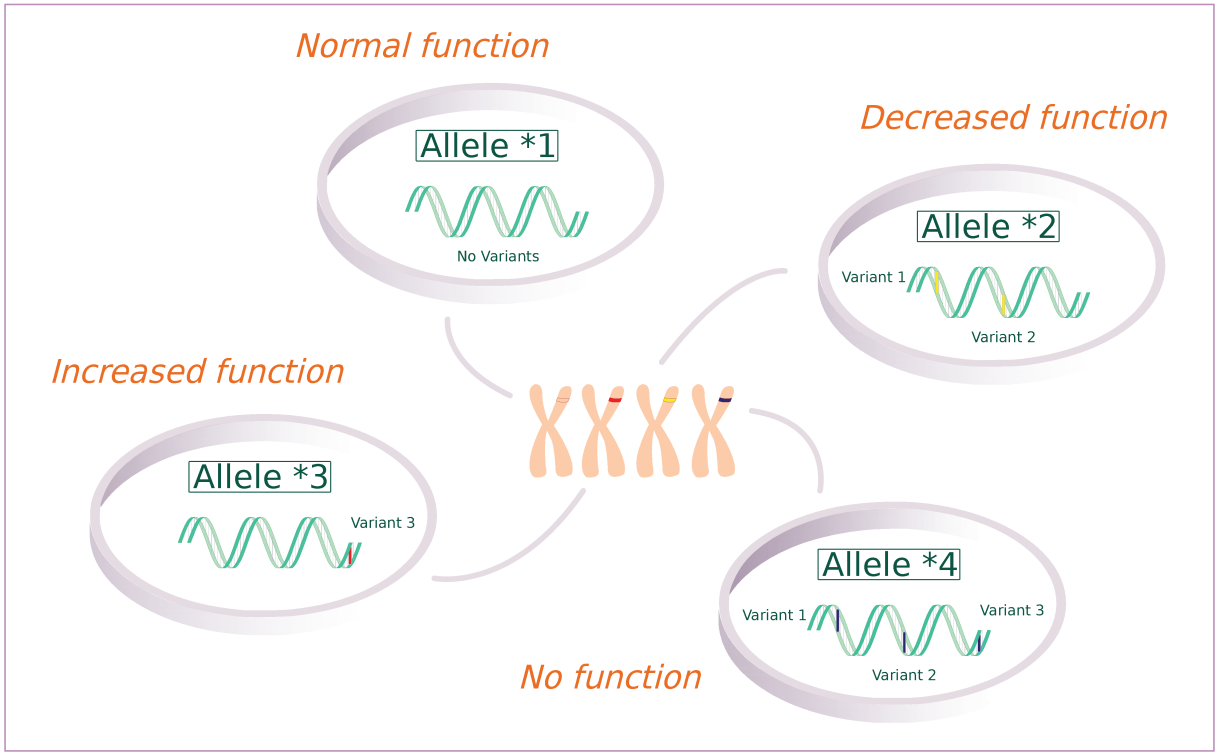

Figure 19: Nomenclature étoile. Une allèle * d'un gène correspond à une version spécifique de ce gène, caractérisée par une combinaison de variants. * 1 est la version de référence du gène, avec une fonction normale. Dans cet exemple, * 2 est une version du gène caractérisée par une combinaison de deux variations, se traduisant par une fonction diminuée, * 3 est caractérisé par un troisième variant, avec pour conséquence une fonction accrue, *4 se caractérise par l'ensemble des variants des allèles *2 et de *3, provoquant une absence de fonction.

Les allèles *1 des gènes sont souvent peu conservées : par exemple, le gène CYP2D6

impliquée dans le métabolisme de nombreux médicaments ne se retrouve sous sa forme de référence *1 chez seulement 33,1 % de la population européenne et 9,3 % de la population africaine[109]. La grande fréquence d'allèles alternatives associés à des changements dans l'effet ou la métabolisation de médicaments a conduit à l'élaboration de recommandations pour l'ajustement des dose de médicaments chez le patient, en fonction des allèles des pharmacogènes concernés.

Le séquençage de nouvelle génération (NGS) permet la détection et l'identification des variations courtes de séquences nucléotidiques (Short Nucleotide variations, SNV) dans l'ensemble d'un pharmacogène, mais ne permet pas de conclure pas directement quant à la nature de l'allèle selon la nomenclature étoile. Nous avons donc développé un outil en ligne permettant de récupérer les SNV appartenant à une liste de pharmacogènes proposée par le Réseau National de Pharmacogénétique (RNPGx). *Cypascan* traite les données NGS et rend presque instantanément le génotype du patient en nomenclatures étoile. Il utilise des fichiers au format vcf (Variant Calling Format) (GRCh37 / 38) et s'appuie sur les bases de données publiques concernant les gènes étudiés.

Ce travail fait l'objet d'un article soumis à la revue *British Journal of Clinical Pharmacology* à la date de rédaction de cette thèse.

Cypascan PGx Alleles Calling, An online tool using Next-Generation Sequencing (NGS) data for pharmacogenetics allele calling from NGS sequencing data, Claire-Cécile Barrot, Jean-Baptiste Woillard, Nicolas Picard

Conclusion

Avec le développement des méthodes de séquençage de plus en plus efficaces, les traitements bio-informatiques prennent de plus en plus en plus d'importance en PGx. Cela a conduit au développement d'outils novateurs dans le traitement de larges jeux de données, dont l'usage devient de plus en plus fréquent. Cependant, ces outils sont liés aux big data, et nécessitent donc de très grandes cohorte d'échantillons à analyser. Si la diminution constante des coûts liés aux divers types de séquençage atténue ce problème, les études PGx seront toujours limitées par l'accès aux patients appropriés ainsi que par la logistique complexe des protocoles d'expérimentation animale. La capacité de retirer des résultats fiables et cohérents à partir de petites cohortes reste donc primordiale.

Les méthodes de traitement bio-informatiques que j'ai développées durant cette thèse utilisent des outils statistiques et mathématiques bien connus dont la fiabilité intrinsèque n'est plus à démontrer. Ils n'ont cependant jusqu'ici jamais été utilisés dans le cadre des PGx, et le défi était de trouver comment les adapter aux problématiques rencontrées.

Entre autres, le choix de les appliquer à la recherche de gènes plutôt que de variants s'est imposé comme un compromis entre le nombre de variables (indépendamment de leur type) nécessaires et le nombre d'échantillons à analyser, chaque gène comportant un grand nombre de variants.

Toutes les méthodes développées ont été de type sans *a priori*, ce qui dans ce cas signifie que la connaissance préalable des voies pharmacodynamiques potentiellement impliquées n'a pas influencé les résultats.

Bien que sans *a priori*, ces méthodes ne sont malgré tout pas complètement libre d'hypothèses, puisque reposant sur la présomption d'un comportement 'standard' des variables (ici des gènes), qui permettrait de repérer celles qui dévieraient de cette norme (à ne pas confondre avec la loi normale). Ce fut particulièrement vrai dans lors de l'étude des profils d'expression et de méthylation, puisque j'y ai spécifiquement recherché les valeurs aberrantes, mais c'est aussi cette hypothèse qui a permis la classification des gènes selon leur type de variabilité lors des études exomiques.

La limite des méthodes développée est bien sûr leur besoin de confirmation. C'est une limite commune à tout traitement *in silico* de données biologiques, bien que largement atténué lorsque les cohortes sont assez grandes pour permettre à la fois analyse et validation. En tant que tel, il ne s'agit pas d'un obstacle à la validation des méthodes, et pourrait même être un atout.

En effet, l'étude spécifique des gènes sélectionnés plutôt que d'exomes (ou transcriptomes)

peut être faite sur de plus grands nombre d'échantillons, et peut aussi intéresser aux variations spécifiques des séquences de nucléotides plutôt qu'aux gènes dans leur ensemble.

Concernant l'analyse des données résultantes, les classifications et scores attribués lors de la sélection des gènes pourraient être très utiles lors de calculs d'inférences bayésiennes. Cela permettrait à la fois d'utiliser ces méthodes ayant d'ores et déjà fait leurs preuves en PGx, tout en s'affranchissant partiellement de leur principales faiblesse en utilisant des résultats observations pour calculer des probabilités *a priori*.

Références bibliographiques

1. Vogel, F. Moderne Probleme der Humangenetik. *Ergebn. inn. med. u. Kinderh* **12**, (1959).
2. Kalow, W. Pharmacogenetics and pharmacogenomics: Origin, status, and the hope for personalized medicine. *Pharmacogenomics J.* **6**, 162–165 (2006).
3. Pouché, L. *et al.* A candidate gene approach of the calcineurin pathway to identify variants associated with clinical outcomes in renal transplantation. *Pharmacogenomics* **17**, 375–391 (2016).
4. Wiseman, A. C. Immunosuppressive medications. *Clin. J. Am. Soc. Nephrol.* **11**, 332–343 (2016).
5. Ericson, J. E. *et al.* A systematic literature review approach to estimate the therapeutic index of selected immunosuppressant drugs after renal transplantation. *Ther. Drug Monit.* **39**, 13–20 (2017).
6. Lim, M. A., Kohli, J. & Bloom, R. D. Immunosuppression for kidney transplantation: Where are we now and where are we going? *Transplant. Rev.* **31**, 10–17 (2017).
7. Takaoka, A. & Yanai, H. Interferon signalling network in innate defence. *Cell. Microbiol.* **8**, 907–922 (2006).
8. Schroder, K., Hertzog, P. J., Ravasi, T. & Hume, D. A. Interferon-gamma: an overview of signals, mechanisms and functions. *J. Leukoc. Biol.* **75**, 163–89 (2004).
9. Platanias, L. C. Mechanisms of type-I- and type-II-interferon-mediated signalling. *Nat. Rev. Immunol.* **5**, 375–386 (2005).
10. Francisco-Cruz, A. *et al.* Granulocyte-macrophage colony-stimulating factor: Not just another haematopoietic growth factor. *Med. Oncol.* **31**, (2014).
11. Gasson, J. C. Molecular Physiology of Granulocyte-Macrophage Colony-Stimulating Factor. *Blood* **77**, 1131–1145 (1991).
12. Ke, H. & Huai, Q. Structures of calcineurin and its complexes with immunophilins-immunosuppressants. *Biochem. Biophys. Res. Commun.* **311**, 1095–1102 (2003).
13. Collège National de Pharmacologie Médicale (CNPM). Available at: https://pharmacomedicale.org.
14. Nankivell, B. J. *et al.* The Natural History of Chronic Allograft Nephropathy. *N. Engl. J. Med.* **349**, 2326–2333 (2003).
15. Dierickx, D. & Habermann, T. M. Post-transplantation lymphoproliferative disorders in adults. *N. Engl. J. Med.* **378**, 549–562 (2018).
16. Vincenti, F. *et al.* Results of an international, randomized trial comparing glucose metabolism disorders and outcome with cyclosporine versus tacrolimus. *Am. J. Transplant.* **7**, 1506–1514 (2007).

17. Morscio, J., Dierickx, D. & Tousseyn, T. Molecular pathogenesis of B-cell posttransplant lymphoproliferative disorder: What do we know so far? *Clin. Dev. Immunol.* **2013**, (2013).

18. Stojanova, J., Caillard, S., Rousseau, A. & Marquet, P. Post-transplant lymphoproliferative disease (PTLD): Pharmacological, virological and other determinants. *Pharmacol. Res.* **63**, 1–7 (2011).

19. Nuckols, J. D. *et al.* The pathology of liver-localized post-transplant lymphoproliferative disease: A report of three cases and a review of the literature. *Am. J. Surg. Pathol.* **24**, 733–741 (2000).

20. Tzellos, S. & Farrell, P. J. Epstein-barr virus sequence variation-biology and disease. *Pathogens* **1**, 156–175 (2012).

21. Santpere, G. *et al.* Genome-wide analysis of wild-type epstein-barr virus genomes derived from healthy individuals of the 1000 genomes project. *Genome Biol. Evol.* **6**, 846–860 (2014).

22. Epstein-Barr Virus and Infectious Mononucleosis. (2016). Available at: https://www.cdc.gov/epstein-barr/about-mono.html.

23. Young, L. S. & Rickinson, A. B. Epstein-Barr virus: 40 Years on. *Nat. Rev. Cancer* **4**, 757–768 (2004).

24. Rickinson, A. B. & Moss, D. J. Human Cytotoxic T Lymphocyte Responses To Epstein-Barr Virus Infection. *Annu. Rev. Immunol.* **15**, 405–431 (1997).

25. Dolcetti, R. Cross-talk between Epstein-Barr virus and microenvironment in the pathogenesis of lymphomas. *Semin. Cancer Biol.* **34**, 58–69 (2015).

26. Al-Mansour, Z., Nelson, B. & Evens, A. Post-transplant lymphoproliferative disease (PTLD): risk factors, diagnosis, and current treatment strategies. *Curr. Hematol. Malig. Rep.* **8**, 173–83 (2013).

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

28. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

29. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, 1–7 (2010).

30. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761 (2018).

31. Mi, H., Muruganujan, A., Ebert, D., Huang, X. & Thomas, P. D. PANTHER version 14: More genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res.* **47**, D419–D426 (2019).

32. Ogata, H. *et al.* KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **27**, 29–34 (1999).

33. Yu, G. & He, Q.-Y. ReactomePA: an R/Bioconductor package for reactome pathway

analysis and visualization. *Mol. Biosyst.* **12**, 477–479 (2016).

34. Team, R. C. R: A language and environment for statistical computing. (2019).

35. Syakur, M. A., Khotimah, B. K., Rochman, E. M. S. & Satoto, B. D. Integration K-Means Clustering Method and Elbow Method for Identification of the Best Customer Profile Cluster. *IOP Conf. Ser. Mater. Sci. Eng.* **336**, (2018).

36. Kassambara, A. & Mundt, F. factoextra: Extract and Visualize the Results of Multivariate Data Analyses. (2017).

37. Lê Cao, K. A., Boitard, S. & Besse, P. Sparse PLS discriminant analysis: Biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics* **12**, (2011).

38. Rohart, F., Gautier, B., Singh, A. & Lê Cao, K.-A. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Comput. Biol.* **13**, e1005752 (2017).

39. Tang, Z. *et al.* GEPIA: A web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res.* **45**, W98–W102 (2017).

40. Seo, S. *et al.* BBS6, BBS10, and BBS12 form a complex with CCT/TRiC family chaperonins and mediate BBSome assembly. *Proc. Natl. Acad. Sci.* **107**, 1488–1493 (2010).

41. Janovská, P. & Bryja, V. Wnt signalling pathways in chronic lymphocytic leukaemia and B-cell lymphomas. *Br. J. Pharmacol.* **174**, 4701–4715 (2017).

42. Bell, P. D. *et al.* Loss of Primary Cilia Upregulates Renal Hypertrophic Signaling and Promotes Cystogenesis. *J. Am. Soc. Nephrol.* **22**, 839–848 (2011).

43. Habbig, S. *et al.* NPHP4, a cilia-associated protein, negatively regulates the Hippo pathway. *J. Cell Biol.* **193**, 633–642 (2011).

44. Habbig, S. *et al.* The ciliopathy disease protein NPHP9 promotes nuclear delivery and activation of the oncogenic transcriptional regulator TAZ. *Hum. Mol. Genet.* **21**, 5528–5538 (2012).

45. Zemirli, N. *et al.* The primary cilium protein folliculin is part of the autophagy signaling pathway to regulate epithelial cell size in response to fluid flow. *Cell Stress* **3**, 100–109 (2019).

46. C. Boonstra, M. *et al.* Clinical Applications of the Urokinase Receptor (uPAR) for Cancer Patients. *Curr. Pharm. Des.* **17**, 1890–1910 (2011).

47. Rubio-Jurado, B. *et al.* Circulating Levels of Urokinase-Type Plasminogen Activator Receptor and D-Dimer in Patients with Hematological Malignancies. *Clin. Lymphoma, Myeloma Leuk.* **15**, 621–626 (2015).

48. Kouhpayeh, S. *et al.* Evaluation of Urokinase Plasminogen Activator Receptor, Soluble Urokinase Plasminogen Activator Receptor, and β1 Integrin in Patients with Hodgkin′s Lymphoma. *Adv. Biomed. Res.* **6**, 108 (2017).

49. Koppen, A. *et al.* Dickkopf-1 is down-regulated by MYCN and inhibits neuroblastoma cell proliferation. *Cancer Lett.* **256**, 218–228 (2007).

50. Menezes, M. E., Devine, D. J., Shevde, L. A. & Samant, R. S. Dickkopf1: a tumor suppressor or metastasis promoter? *Int J Cancer.* **130**, 1477–1483 (2012).

51. Rusu, V. *et al.* Type 2 Diabetes Variants Disrupt Function of SLC16A11 through Two Distinct Mechanisms. *Cell* **170**, 199–212 (2017).

52. Jung, S. Y., Song, H. S., Park, S. Y., Chung, S. H. & Kim, Y. J. Pyruvate promotes tumor angiogenesis through HIF-1-dependent PAI-1 expression. *Int. J. Oncol.* **38**, 571–576 (2011).

53. Lee, M. S. *et al.* Angiogenic activity of pyruvic acid in in vivo and in vitro angiogenesis models. *Cancer Res.* **61**, 3290–3293 (2001).

54. Schwarzfischer, P. *et al.* Comprehensive Metaboproteomics of Burkitt's and Diffuse Large B-Cell Lymphoma Cell Lines and Primary Tumor Tissues Reveals Distinct Differences in Pyruvate Content and Metabolism. *J. Proteome Res.* **16**, 1105–1120 (2017).

55. Liu, A., Claesson, H. E., Mahshid, Y., Klein, G. & Klein, E. Leukotriene B4 activates T cells that inhibit B-cell proliferation in EBV-infected cord blood derived mononuclear cell cultures. *Blood* **111**, 2693–2703 (2008).

56. Rabajdova, M. *et al.* The crucial role of emilin 1 gene expression during progression of tumor growth. *J. Cancer Res. Clin. Oncol.* **142**, 2397–2402 (2016).

57. Mongiat, M. *et al.* The Extracellular Matrix Glycoprotein Elastin Microfibril Interface Located Protein 2: A Dual Role in the Tumor Microenvironment. *Neoplasia* **12**, 294-IN1 (2010).

58. Mohren, M. *et al.* High coagulation factor VIII and von Willebrand factor in patients with lymphoma and leukemia. *Int. J. Hematol.* **103**, 189–195 (2016).

59. Magni, M. *et al.* TSPYL2 is a novel regulator of SIRT1 and p300 activity in response to DNA damage. *Cell Death Differ.* **26**, 918–931 (2019).

60. Davidson, J. *et al.* New-Onset Diabetes After Transplantation: 2003 International Consensus Guidelines1. *Transplantation* **75**, SS3–SS24 (2003).

61. Hjelmesæth, J. *et al.* The impact of early-diagnosed new-onset post-transplantation diabetes mellitus on survival and major cardiac events. *Kidney Int.* **69**, 588–595 (2006).

62. Kasiske, B. L., Snyder, J. J., Gilbertson, D. & Matas, A. J. Diabetes mellitus after kidney transplantation in the United States. *Am. J. Transplant.* **3**, 178–185 (2003).

63. Burroughs, T. E. *et al.* Diabetic complications associated with new-onset diabetes mellitus in renal transplant recipients. *Transplantation* **83**, 1027–1034 (2007).

64. Pham, P.-T., Pham, P.-M. & Pham, P.-C. New onset diabetes after transplantation (NODAT): an overview. *Diabetes, Metab. Syndr. Obes. Targets Ther.* 175 (2011). doi:10.2147/dmso.s19027

65. Suarez, O. *et al.* Diabetes mellitus and renal transplantation in adults: Is there enough evidence for diagnosis, treatment, and prevention of new-onset diabetes after renal transplantation? *Transplant. Proc.* **46**, 3015–3020 (2014).

66. Woodward, R. S. *et al.* Incidence and cost of new onset diabetes mellitus among U.S. wait-listed and transplanted renal allograft recipients. *Am. J. Transplant.* **3**, 590–598 (2003).

67. Villeneuve, C. *et al.* Adherence profiles in kidney transplant patients: Causes and consequences. *Patient Educ. Couns.* (2019). doi:10.1016/j.pec.2019.08.002

68. Villeneuve, C. *et al.* Evolution and determinants of health-related quality-of-life in kidney transplant patients over the first 3 years after transplantation. *Transplantation* **100**, 640–647 (2016).

69. Anchoori, R. K. *et al.* A bis-Benzylidine Piperidone Targeting Proteasome Ubiquitin Receptor RPN13/ADRM1 as a therapy for cancer. *Cancer Cell* **24**, (2013).

70. Soong, R. S. *et al.* RPN13/ADRM1 inhibitor reverses immunosuppression by myeloid-derived suppressor cells. *Oncotarget* **7**, 68489–68502 (2016).

71. Marcoux, A. A. *et al.* Molecular features and physiological roles of K+-Cl− cotransporter 4 (KCC4). *Biochim. Biophys. Acta - Gen. Subj.* **1861**, 3154–3166 (2017).

72. Melo, Z. *et al.* Molecular evidence for a role for K+-Cl- cotransporters in the kidney. *Am. J. Physiol. - Ren. Physiol.* **305**, 1402–1411 (2013).

73. Blessia, T. F., Singh, S. & Vennila, J. J. Unwinding the Novel Genes Involved in the Differentiation of Embryonic Stem Cells into Insulin-Producing Cells: A Network-Based Approach. *Interdiscip. Sci. Comput. Life Sci.* **9**, 88–95 (2017).

74. Bergeron, V., Ghislain, J. & Poitout, V. The P21-activated kinase PAK4 is implicated in fatty-acid potentiation of insulin secretion downstream of free fatty acid receptor 1. *Islets* **8**, 157–164 (2016).

75. Fadista, J. *et al.* Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 13924–13929 (2014).

76. Zhou, T. C. *et al.* Novel genetic findings in a Chinese family with early-onset femalerelated type 2 diabetes. *Acta Endocrinol. (Copenh).* **13**, 364–369 (2017).

77. Waeber, G. *et al.* The gene MAPK8IP1, encoding islet-brain-1, is a candidate for type 2 diabetes. *Nat. Genet.* **24**, 291–295 (2000).

78. Fan, J., Tatum, R., Hoggard, J. & Chen, Y.-H. Claudin-7 Modulates Cl− and Na+ Homeostasis and WNK4 Expression in Renal Collecting Duct Cells. *Int. J. Mol. Sci.* **20**, 3798 (2019).

79. Ghodsian, N. *et al.* Novel Association of WNK4 Gene, Ala589Ser Polymorphism in Essential Hypertension, and Type 2 Diabetes Mellitus in Malaysia. *J. Diabetes Res.* **2016**, (2016).

80. Galimov, A. *et al.* Growth hormone replacement therapy regulates microRNA-29a and

targets involved in insulin resistance. *J. Mol. Med.* **93**, 1369–1379 (2015).

81. Schierding, W. & O'Sullivan, J. M. Connecting snps in diabetes: A spatial analysis of meta-GWAS loci. *Front. Endocrinol. (Lausanne).* **6**, 1–6 (2015).

82. Wada, J., Sun, L. & Kanwar, Y. S. Discovery of genes related to diabetic nephropathy in various animal models by current techniques. *Contrib. Nephrol.* **169**, 161–174 (2011).

83. Shihab, H. A. *et al.* Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models. *Hum. Mutat.* **34**, 57–65 (2013).

84. Shihab, H. A. *et al.* Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models. *Hum. Mutat.* **34**, 57–65 (2013).

85. Karczewski, K. J. *et al.* Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv* 531210 (2019). doi:10.1101/531210

86. Vinson, C. & Chatterjee, R. CG methylation. *Epigenomics* **4**, 655–63 (2012).

87. Lötsch, J. *et al.* Common non-epigenetic drugs as epigenetic modulators. *Trends Mol. Med.* **19**, 742–753 (2013).

88. Mai, A. & Altucci, L. Epi-drugs to fight cancer: From chemistry to cancer treatment, the road ahead. *Int. J. Biochem. Cell Biol.* **41**, 199–213 (2009).

89. Thomson, J. P., Moggs, J. G., Wolf, C. R. & Meehan, R. R. Epigenetic profiles as defined signatures of xenobiotic exposure. *Mutat. Res. - Genet. Toxicol. Environ. Mutagen.* **764**–**765**, 3–9 (2014).

90. Corley, M. J., Zhang, W., Zheng, X., Lum-Jones, A. & Maunakea, A. K. Semiconductor-based sequencing of genomewide DNA methylation states. *Epigenetics* **10**, 153–166 (2015).

91. J, S. Développement de méthodes de séquençage de seconde génération pour l'analyse des profils de méthylation de l'ADN. (Université Pierre et Marie Curie - Paris VI).

92. Carvalho, B. S. & Irizarry, R. A. A framework for oligonucleotide microarray preprocessing. *Bioinformatics* **26**, 2363–2367 (2010).

93. Irizarry, R. A. *et al.* Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. *Biostatistics* 249–264 (2003).

94. LeFranc, M. P. Nomenclature of the Human Immunoglobulin Lambda (IGL) genes. *Exp. Clin. Immunogenet.* **18**, 242–254 (2001).

95. Graf, D., Korthäuer, U., Mages, H. W., Senger, G. & Kroczek, R. A. Cloning of TRAP, a ligand for CD40 on human T cells. *Eur. J. Immunol.* **22**, 3191–3194 (1992).

96. Blotta, M. H., Marshall, J. D., DeKruyff, R. H. & Umetsu, D. T. Cross-linking of the

CD40 ligand on human CD4+ T lymphocytes generates a costimulatory signal that up-regulates IL-4 synthesis. *J. Immunol.* **156**, 3133–40 (1996).

97. Mikolajczak, S. A. *et al.* The Modulation of CD40 Ligand Signaling by Transmembrane CD28 Splice Variant in Human T Cells. *J. Exp. Med.* **199**, 1025–1031 (2004).

98. Gao, Z. *et al.* AUTS2 confers gene activation to Polycomb group proteins in the CNS. *Nature* **516**, 349–354 (2014).

99. Danen-van Oorschot, A. A. A. M. *et al.* Human death effector domain-associated factor interacts with the viral apoptosis agonist Apoptin and exerts tumor-preferential cell killing. *Cell Death Differ.* **11**, 564–573 (2004).

100. Ma, W. *et al.* Proapoptotic RYBP interacts with FANK1 and induces tumor cell apoptosis through the AP-1 signaling pathway. *Cell. Signal.* **28**, 779–787 (2016).

101. Zhou, H. *et al.* RING1 and YY1 binding protein suppresses breast cancer growth and metastasis. *Int. J. Oncol.* **49**, 2442–2452 (2016).

102. Chen, D. *et al.* RYBP stabilizes p53 by modulating MDM2. *EMBO Rep.* **10**, 166–172 (2009).

103. Liu, Y., Cai, Y., Liu, L., Wu, Y. & Xiong, X. Crucial biological functions of CCL7 in cancer. *PeerJ* **2018**, 1–21 (2018).

104. W., L. M. *et al.* Critical roles of Bim in T cell activation and T cell–mediated autoimmune inflammation in mice. *J. Clin. Invest.* **119**, (2009).

105. Lefranc, M. P. Immunoglobulin and T cell receptor genes: IMGT® and the birth and rise of immunoinformatics. *Front. Immunol.* **5**, 1–22 (2014).

106. Barrot, C.-C., Woillard, J.-B. & Picard, N. Big data in pharmacogenomics: current applications, perspectives and pitfalls. *Pharmacogenomics* **20**, 609–620 (2019).

107. Stelzer, G. *et al.* The GeneCards suite: From gene data mining to disease genome sequence analyses. *Curr. Protoc. Bioinforma.* **2016**, 1.30.1-1.30.33 (2016).

108. Kalman, L. V, Black, J. L., Clinic, M., Sw, S. & Bell, G. C. *Pharmacogenetic Allele Nomenclature: International Workgroup Recommendations for Test Result Reporting*. **99**, (2017).

109. Zhou, Y., Ingelman-Sundberg, M. & Lauschke, V. Worldwide Distribution of Cytochrome P450 Alleles: A Meta-analysis of Population-scale Sequencing Projects. *Clin. Pharmacol. Ther.* **102**, (2017).

Annexes

Annexe 1, Protocole d'isolation des LTCD4 à partir de sang total et solution cellulaire de rate

Le sang total et la solution cellulaire de rate étaient incubés à l'obscurité avec un tampon de lyse des hématies pendant 8 minutes (tampon Amonium-Chloride-Potassium (ACK), 3ml de tampon pour 100µl de matrice biologique). Ce tampon a été préparé avec du NH4Cl (chlorure d'ammonium, 8,26g/l), du KHCO3 (bicarbonate de potassium, 1g/l) et de l'EDTA Na2 (0,03g/l). 20 ml de RPMI était ajouté pour arrêter la réaction de lyse. Après centrifugation (1200 rpm, 10 min à +4°C), les culots cellulaires de sang et rate étaient repris dans 1 ml de PBS.

L'homogénat était filtré sur colonne MS (Columns MS, Miltenyi Biotech) pour procéder à une sélection positive de cellules CD4 à l'aide des microbilles magnétiques couplées à des anticorps anti-CD4 (CD4 L3T4 MicroBeads mouse, Miltenyi Biotech, 10 µl pour 107 cellules).

La pureté des LT CD4 a été évaluée par Cytométrie de Flux (CMF) (FacsCalibur, Becton Dickinson), en utilisant des anticorps BD Pharmigen™ FITC rat anti-mouse CD4, BD Biosciences® (0,5 µl d'anticorps pour 106 cellules).

Annexe 2, Protocole MeDip-Seq

Protocole MeDIP-Seq

I. *Genomic DNA isolation*

A. QI Amp® DNA Blood Mini Kit (Qiagen)

= Extraction ADN génomique

- ✓ Déposer 20µl de protéase au fond de tubes de 1,5ml
- ✓ Ajouter 2.10^6 LT CD4 total dans un volume final de 200µl
- ✓ Ajouter 200µl de tampon AL
- ✓ Vortexer 15 secondes
- ✓ Incuber 10min à 56°C
- ✓ Centrifuger brièvement (pour éliminer les gouttes du bouchon des tubes)
- ✓ Ajouter 200µl d'éthanol (96-100%) dans chaque tube
- ✓ Vortexer 15 secondes et centrifuger brièvement
- ✓ Déposer le contenu des tubes sur les colonnes Qiagen, ne pas mouiller les bords
- ✓ Centrifuger à 8000rpm pendant 1min
- ✓ Placer la colonne dans un nouveau tube de 2ml Qiagen et jeter l'ancien
- ✓ Déposer 500µl de tampon AW1
- ✓ Centrifuger à 8000rpm pendant 1min
- ✓ Placer la colonne dans un nouveau tube de 2ml Qiagen et jeter l'ancien
- ✓ Déposer 500µl de tampon AW2
- ✓ Centrifuger à vitesse maximale pendant 1min
- ✓ Placer la colonne dans un nouveau tube de 2ml Qiagen et jeter l'ancien
- ✓ Centrifuger à nouveau à vitesse maximale 3min
- ✓ Placer la colonne sur un tube de 1,5ml
- ✓ Déposer au centre de la colonne 200µl de tampon AE
- ✓ Attendre 2min puis centrifuger à 8000rpm pendant 1min

Remarque : L'ADN peut être stocké à -20°C.

B. Qubit® DNA Assay Kit avec le Qubit® 2.0 Fluorometer (Life technologiesTM)

= Dosage ADN génomique

Remarque : L'ADN peut être stocké à -20°C.

II. *Ion Torrent MeDIP-Seq Library Construction*

C. Bioruptor® Pico (Diagenode)

= Sonication ADN génomique

- ✓ Allumer le sonicateur au moins 20min avant l'utilisation (le temps que l'eau descende à 4°C)

- ✓ Conserver sur glace

Remarque : L'ADN peut être stocké à -20°C.

D. End-repair DNA : Ion Plus Fragment Library

= Réparation extrémités de l'ADN fragmenté

NB : Avant utilisation, centrifuger 2 secondes les composants du **Ion plus Fragment Library Kit** afin de les faire redescendre au fond du tube

- ✓ Ajouter de la **Nuclease-free Water** à l'ADN fragmenté pour atteindre le volume total suivant :

100 ng ADN	1µg d'ADN
79µl	158µl

- ✓ Mélanger par pipetage dans un tube Eppendorf de 1 ,5ml :

Composant	Volume (µl)	
	100ng	1µg
ADN fragmenté	79	158
5X End Repair Buffer	20	40
End Repair Enzyme	1	2
Total	100	200

- ✓ Incuber 20min à température ambiante

E. Agencourt AMPure XP kit (Beckman Coulter)

= Purification de l'ADN fragmenté

ATTENTION : Toujours utiliser de l'éthanol à 70% fraîchement préparé (1ml par échantillon + excès) pour les étapes suivantes. Un pourcentage trop élevé d'éthanol pourrait causer un lavage inefficace des petits fragments d'ADN. Un pourcentage trop faible d'éthanol pourrait causer une perte d'échantillon

- ✓ Ajouter un volume d'**Agencourt AMPure XP Reagent** à l'échantillon (1,8X le volume d'échantillon)

100 ng ADN	1µg d'ADN
180µl	360µl

- ✓ Pipeter 5 fois pour mélanger complètement la suspension de billes avec l'ADN
- ✓ Centrifuger brièvement
- ✓ Laisser incuber 5min à température ambiante
- ✓ Centrifuger brièvement
- ✓ Placer les tubes sur le **DynaMag™-2**
- ✓ Attendre 3min ou jusqu'à ce que la solution soit limpide
- ✓ Enlever soigneusement et jeter le surnageant, sans perturber le culot de billes
- ✓ Ne pas enlever les tubes du **DynaMag™-2**
- ✓ Ajouter 500µl d'éthanol 70% fraichement préparé
- ✓ Incuber 30s
- ✓ Tourner les tubes accrochés sur le **DynaMag™-2** sur eux-mêmes deux fois pour faire bouger les billes autour de la paroi des tubes
- ✓ Attendre que la solution soit limpide
- ✓ Enlever soigneusement et jeter le surnageant, sans perturber le culot de billes
- ✓ Répéter le lavage à l'éthanol 70% une seconde fois
- ✓ Pour enlever l'éthanol résiduel, centrifuger brièvement
- ✓ Remettre les tubes sur le **DynaMag™-2**
- ✓ Enlever doucement le reste du surnageant avec une P20, sans perturber le culot de billes
- ✓ Ne pas enlever les tubes du **DynaMag™-2**
- ✓ Laisser sécher les billes à température ambiante ≤5min
- ✓ Enlever les tubes du **DynaMag™-2**
- ✓ Ajouter 25µl de **Low TE** directement sur le culot pour disperser les billes
- ✓ Mélanger complètement la suspension par 5 pipetages successifs
- ✓ Vortexer les échantillons 10s
- ✓ Centrifuger brièvement
- ✓ Placer les tubes sur le **DynaMag™-2**
- ✓ Attendre au moins 1min ou jusqu'à ce que la solution soit limpide
- ✓ Transférer soigneusement le surnageant contenant l'ADN élué dans un tube PCR 0,2ml

Remarque : L'ADN peut être stocké à -20°C.

F. Ion plus fragment Library Kit

= Liaison des adaptateurs aux fragments d'ADN double brin

ATTENTION : Lors de la manipulation des adaptateurs et des barcodes, faire plus particulièrement attention à la contamination croisée !!! Changer fréquemment de gants et ouvrir un tube à la fois

- ✓ Dans des tubes PCR de 0,2mL, mélanger :

Réactifs	Volume (µl)	
	50-100ng	1µg
DNA	≈25	≈25
10X Ligase Buffer	10	10
Ion P1 Adapter	2	10
Ion Xpress™ Barcode X	2	10
dNTP Mix	2	2
Nuclease-free Water	49	31
DNA Ligase	2	4
Nick Repair Polymerase	8	8
Total	100	100

- ✓ Mélanger par pipetage (**ATTENTION : ça mousse !!!**)
- ✓ Placer les tubes dans un thermocycleur et appliquer le programme suivant

1. Préparation des billes

ATTENTION : Ne jamais laisser les billes à l'air libre sans tampon, elles ne doivent pas sécher !!!

- ✓ Préparer le tampon de lavage (**bead wash Buffer)** en diluant au 1/5 le **MagBuffer A** (4°C) avec **Water** (4°C). Le volume nécessaire de tampon de lavage par IP est de 100µl → 100/5 = 20µl **MagBuffer A** + 80µl **Water** (200µl **MagBuffer A** + 800µl **Water** pour 10 IP)
- ✓ Resuspendre les billes (**Magbeads** 4°C) et tranférer 11µl de billes par IP dans un nouveau tube eppendorf de 1,5ml (RNase et DNase free) (110µl pour 10 IP)
- ✓ Mettre sur **Magnetic Rack**
- ✓ Enlever le surnageant et garder les billes
- ✓ Resuspendre les billes dans le tampon de lavage froid
- ✓ Mettre sur **Magnetic Rack**
- ✓ Enlever le surnageant et garder les billes
- ✓ Répéter ce lavage une fois de plus

Nombre d'IP	Volume de tampon (µl) par lavage	Volume de billes (µl)
1	27,5	11
10	275	110

- ✓ Resuspendre les billes dans 22µl de tampon de lavage par 1 IP (220µl pour 10 IP) et laisser dans la glace

2. Préparation du mix pour l'IP

- ✓ Préparation du mix dans des tubes PCR 0,2ml avec bouchons plats pour pas que ça fonde dans le thermocycleur
- ✓ Seulement dans un tube d'IP faire le mix suivant avec le meDNA et le unDNA

Réactifs	Volume pour 1 IP + Input (1) µl
Water	57-20 = 37
MagBuffer A	24
MagBuffer B (4°C)	6
Positive meDNA control (-20°C)	1,5
Negative unDNA control (-20°C)	1,5
DNA sample (-20°C)	20
Volume Total	90

- ✓ Dans les autres tubes faire le mix suivant sans le meDNA et le unDNA

Réactifs	Volume pour 1 IP + Input (1) µl
Water	3 (volume des deux contrôles) + 37 = 40
MagBuffer A	24
MagBuffer B	6
DNA sample	20
Volume Total	90

→ **1µg d'ADN est nécessaire par IP !!!**

Nombre d'IP	Mix (µl)	Echantillon ADN (µl) à une concentration de 20ng/µl	Volume total (µl)
1	70	20	90

Remarque : Volume total de 90µl = 1 IP (75µl), 10% d'input (7,5µl) et un excès (7,5µl)

- ✓ Incuber 3min à 95°C
- ✓ Refroidir rapidement l'échantillon en mettant sur glace
- ✓ Centrifuger brièvement
- ✓ Prendre 7,5µl (10% input) de chaque IP et transférer dans une nouvelle barrette. Conserver l'input à 4°C
- ✓ Transférer 75µl de chaque IP dans une nouvelle barrette. Conserver à 4°C
- ✓ Jeter l'ancienne barrette
- ✓ Dans un nouveau tube, préparer le **Dilutated Antibody mix** (voir tableau ci-dessous). Ajouter l'AC, le **MagBuffer A** et l'eau en premier. Ajouter le **MagBuffer C** par la suite

Réactifs	1 IP	10 IP
AC (µl)	0,15	1,5
MagBuffer A (µl)	0,60	6
Water (µl)	2,25	22,5
MagBuffer C (µl)	2	20
Volume final (µl)	5	50

- ✓ Ajouter 5µl de **Dilutated Antibody mix** par tube d'IP qui contient déjà l'**IP incubation mix** et l'échantillon d'ADN
- ✓ Mélanger et ajouter 20µl de billes lavées par tube d'IP (volume final pour 1 tube d'IP = 100µl)
- ✓ Placer sur une roue et laisser en rotation toute la nuit ou 4H à 4°C

ATTENTION : bien boucher les tubes pour ne pas avoir de fuite !!!

3. Lavages

- ✓ Placer les **MagWash buffers** et le **Magnetic Rack** sur la glace. Réaliser les lavages sur glace ou en chambre froide
- ✓ Centrifuger brièvement les tubes d'IP et les placer sur le **Magnetic Rack** dans la glace, attendre 1min et enlever le tampon
- ✓ Ajouter par tube d'IP 100µl de **MagWash Buffer-1** froid
- ✓ **Pipetter doucement** pour resuspendre les billes (**sinon ça mouse !!!**)
- ✓ Incuber 4min à 4°C sur une roue en rotation
- ✓ Centrifuger brièvement
- ✓ Placer sur le **Magnetic Rack**
- ✓ Attendre 1min et enlever le tampon
- ✓ Garder les billes accrochées à la paroi
- → Répéter le lavage 1 fois de plus
- ✓ Laver les billes une fois avec 100µl de **MagWash buffer-2** froid
- ✓ Après le dernier lavage, enlever le tampon, garder le culot de billes sur glace

4. Isolation de l'ADN

- ✓ Récupérer les tubes input qui sont à 4°C
- ✓ Préparer 50µl par IP et 100µl par Input de **complete buffer DIB** : ajouter 1µl de **Proteinase K** (-20°C) pour 100µl de **Buffer DIB** (50 µl sont nécessaires pour l'IP et 92,5µ sont nécessaires pour l'input → soit pout 10 IP + 10 inputs = 1500µl de **Buffer DIB** + 15µ de **Proteinase K**)
  - Ajouter 50µl de **complete buffer DIB** par IP et resuspendre les billes
  - Ajouter 92,5µl de **complete buffer DIB** au 7,5µl d'Input
  - ⇨ Volume final dans chaque tube = 50µl pour IP et 100µl pour Input
- ✓ Incuber 15min à 55°C + incuber 15min à 100°C (Thermocycleur → User « Lucie » « medip-dna-isol1 »)
- ✓ Centrifuger les tubes pendant 3min ou les placer sur le **Magnetic Rack** (attendre 1min)
- ✓ Transférer le surnageant dans un nouveau tube de 1,5ml

Remarque : L'ADN peut être stocké à -20°C.

I. Platinum® PCR SuperMix High Fidelity

= Amplification de la library

- ✓ Combiner les réactifs suivants dans un tube de taille approprié et mélanger par pipetage

Composé	Volume par échantillon d'ADN (µl)	
	50-100ng	1µg
Platinum® PCR SuperMix High Fidelity	100	200
Library Amplification Primer Mix	5	10
Library non amplifiée	25	50
Total	130	260

- ✓ Partager le mélange réactionnel dans plusieurs tubes PCR de 0,2ml pour ajuster au volume réactionnel recommandé par le fabricant du thermocycleur
- ✓ Placer les tubes dans le thermocycleur et lancer le programme de PCR indiqué dans le premier tableau. Appliquer le nombre de cycles préconisé dans le deuxième tableau

Stade	Etape	Température (°C)	Temps
Holding	Dénaturation	95	5min
Cycling	Dénaturation	95	15sec
	Hybridation	58	15sec
	Elongation	70	1min
Holding		4	Hold

Nombre de cycles par library
1µg
18

- ✓ Regrouper les différents échantillons séparés précédemment dans un tube Eppendorf de 1,5ml

J. Agencourt AMPure XP kit (Beckman Coulter)

= Purification des complexes ADN/adaptateurs et sélection de taille.

ATTENTION : Toujours utiliser de l'éthanol à 80% fraîchement préparé (1ml par échantillon + excès) pour les étapes suivantes

- ✓ Ajuster le volume d'échantillon à 100µl. Si cela n'est pas possible adapter les volumes suivants
- ✓ Ajouter un volume d'**Agencourt AMPure XP Reagent** à l'échantillon

Paramètres de la banque d'ADN génomique	Taille approximative après sonication	150pb	200pb	250pb	300-400pb	400-500pb	500-700pb
	Taille après construction de la banque d'ADN génomique	270pb	320pb	400pb	400-500pb	500-600pb	600-800pb
Volume de billes (µl)	1ère sélection	65	55	45	40	35	30
	2ième sélection	25	25	25	20	15	15

- ✓ Pipeter 10 fois pour mélanger complétement la suspension de billes avec l'ADN
- ✓ Laisser incuber 5min à température ambiante
- ✓ Centrifuger brièvement
- ✓ Placer les tubes sur le **DynaMag™-2**
- ✓ Attendre 5min ou jusqu'à ce que la solution soit limpide
- ✓ Récupérer soigneusement et mettre dans un nouveau tube le surnageant qui contient l'ADN d'intérêt
- ✓ Jeter les billes qui contiennent les fragments larges non désirés
- ✓ Ajouter le volume approprié d'**Agencourt AMPure XP Reagent** au surnageant
- ✓ Pipeter 10 fois pour mélanger complétement la suspension de billes avec l'ADN
- ✓ Laisser incuber 5min à température ambiante
- ✓ Centrifuger brièvement
- ✓ Placer les tubes sur le **DynaMag™-2**
- ✓ Enlever soigneusement et jeter le surnageant, sans perturber le culot de billes
- ✓ Ne pas enlever les tubes du **DynaMag™-2**
- ✓ Ajouter 200µl d'éthanol **80%** fraichement préparé
- ✓ Incuber 30s
- ✓ Tourner les tubes accrochés sur le **DynaMag™-2** sur eux-mêmes deux fois pour faire bouger les billes autour de la paroi des tubes
- ✓ Attendre que la solution soit limpide
- ✓ Enlever soigneusement et jeter le surnageant, sans perturber le culot de billes
- ✓ Répéter le lavage à l'éthanol 80% deux fois supplémentaires
- ✓ Ne pas enlever les tubes du **DynaMag™-2**
- ✓ Laisser sécher les billes à température ambiante ≤10min
- ✓ Enlever les tubes du **DynaMag™-2**
- ✓ Ajouter 25µl de **Low TE** directement sur le culot pour disperser les billes
- ✓ Mélanger complètement la suspension par pipetages successifs
- ✓ Vortexer les échantillons 10s
- ✓ Centrifuger brièvement
- ✓ Placer les tubes sur le **DynaMag™-2**
- ✓ Attendre au moins 5min ou jusqu'à ce que la solution soit limpide
- ✓ Transférer soigneusement le surnageant contenant l'ADN élué dans un tube Eppendorf de 1,5ml

Remarque : L'ADN peut être stocké à -20°C.

K. Agilent 2100 Bioanalyzer (Agilent Technologies)

Index

ACP : analyse en composante principale
CSA : ciclosporine
CTR : contrôle
CYP : cytochromes P450
DLBC : lymphomes diffus à grandes cellules B
DNMT : méthyl-transférases de l'ADN
EBV : virus d'Epstein-Barr
EMX : métabolisation des xénobiotiques
GM-CSF : facteur stimulant les colonies de granulocytes et de macrophages
ICN : inhibiteurs de la calcineurine
IFNγ : interféron gamma
IL : interleukine
LT : lymphocytes T
LTCD4 : lymphocytes T auxiliaires
NFAT : facteur nucléaire des lymphocytes T activés
NGS : séquençage de nouvelle génération
NODAT : diabète d'apparition tardive après transplantation
PGx : pharmacogénomique
PTLD : syndrome lymphoprolifératif post-transplantation
RMA : robust multi-array analysis
RNPGx : Réseau National de Pharmacogénétique
SNV : variations courtes de séquences nucléotidiques
sPLS-DA : analyse discriminante de régression parcimonieuses des moindres carrés partiels
TAC : tacrolimus
TCR : récepteurs des cellules T
THYM : thymomes
TME : micro-environnement tumoral
VCF : variant calling format

Index des scripts

Table des matières

Recherche de Pharmacogènes associés aux effets indésirables des Inhibiteurs de la Calcineurine. Développement d'approches bio-informatiques adaptées aux petits échantillons.

L'efficacité des méthodes de séquençage de nouvelle génération et la réduction de leur coût ont conduit à l'utilisation de nouvelles méthodes, basées sur le concept de l'analyse de big data, appliqué aux données génomiques. Cependant, ces outils nécessitent de très grandes cohortes d'échantillons à analyser, limitées en pharmacogénomique par l'accès aux patients appropriés ainsi que par la logistique complexe des protocoles d'expérimentation animale. La capacité de retirer des résultats fiables et cohérents à partir de petites cohortes reste donc primordiale. Dans le cadre de cette thèse, de nouvelles approches bio-informatiques adaptées aux petites cohortes ont été explorée en recherchant des pharmacogènes associés aux effets indésirables des Inhibiteurs de la Calcineurine.

Mots-clés : bioinformatique, génétique, pharmacogénomique, inhibiteur de calcineurine

Investigation of pharmacogenes related to Adverse Effects of Calcineurin Inhibitors. Development of bioinformatics approaches adapted to small cohorts.

Efficiency of new generation sequencing methods and the reduction of their cost have led to the use of new methods based on big data to analyse genomic data. However, these tools require very large cohorts of samples to be used, which is limited in pharmacogenomics due to access to appropriate patients and complex logistics of animal testing protocols. The ability to obtain reliable and consistent results from small cohorts remains an important challenge. As part of this thesis, new bioinformatics approaches adapted to small cohorts were explored by investigation of pharmacogenes related to adverse effects of Calcineurin Inhibitors.

Keywords : bioinformatics, genetics, pharmacogenomics, calcineurin inhibitors