Approches génétiques et génomiques pour l'identification de gènes prédisposant à une maladie multifactorielle : le diabète de type 1

Morgane Bourmaud

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE

ÉCOLE PRATIQUE DES HAUTES ÉTUDES
Sciences de la Vie et de la Terre

MÉMOIRE

Présenté par

Morgane Bourmaud

pour l'obtention du Diplôme de l'École Pratique des Hautes Études

Approches génétiques et génomiques pour l'identification de gènes prédisposant à une maladie multifactorielle : le diabète de type 1

soutenu le 09 décembre 2016 devant le jury suivant :

Sophie Gad-Lapiteau, Maître de conférences, EPHE – Président
Claire Vandiedonck, Maître de conférences, Paris Diderot – Tuteur scientifique
Flore Renaud, Maître de conférences, EPHE – Tuteur pédagogique
Matthieu Giraud, CR1 INSERM – Rapporteur
Fabien Fauchereau, Maître de conférences, Paris Diderot – Examinateur

Mémoire préparé sous la direction de :
Docteur Claire Vandiedonck
INSERM UMRS958 – Génétique des diabètes, Université Paris Diderot, Paris
Directrice : Cécile Julier

Et de

Docteur Flore Renaud
Laboratoire Génétique et biologie cellulaire, EA4589 UVSQ/EPHE, Montigny-le-Bretonneux
Directeur : Bernard Mignotte

Remerciements

Je tiens tout d'abord à remercier tous les membres de mon jury.
Je remercie sincèrement Sophie Gad-Lapiteau de me faire l'honneur de présider ce jury.
Un très grand merci à Matthieu Giraud et à Fabien Fauchereau pour avoir accepté de faire partie de mon jury et de prendre du temps pour évaluer mon travail.

Merci beaucoup à ma tutrice pédagogique, Flore Renaud, pour l'aide et les conseils que vous m'avez apportés tout au long de ces 3 années.

Un grand merci à Claire Vandiedonck, ma tutrice scientifique, de m'avoir accueillie au sein de son équipe et de m'avoir fait partagé ses connaissances. J'ai beaucoup appris à tes côtés et je ne te remercierai jamais assez de m'avoir laissé une chance en tant qu'assistant ingénieure et surtout de m'avoir encouragée à m'inscrire à ce diplôme. Merci pour ta gentillesse, ta patience et je te souhaite vraiment le meilleur tant pour ton avenir professionnel que personnel…

Merci à Azadeh et Tharshana, mes collègues de bureau, pour leur soutien et les bons moments que l'on a passés.

Merci à Valérie Senée et Sophie Romero pour leurs conseils et surtout leur bonne humeur !

Merci à Cécile Julier de m'avoir accueillie au sein de son laboratoire et à tous les autres membres de l'unité 958.

Merci à Pierre-Antoine Defossez de m'avoir laissé du temps pour réaliser ce diplôme et à toute son équipe pour leur soutien.

Merci à Hervé Petite et aux membres de son unité B2OA - CNRS UMR 7052 pour m'avoir permis de réaliser mes cultures cellulaires dans leur pièce de culture.

Merci aux volontaires de la cohorte CIG de l'Institut Pasteur et à Marie-Noelle Ungeheuer pour nous avoir aidés à avancer sur le projet.

Merci à l'unité Inserm 1148 et notamment à Marc Clément pour m'avoir aidée à réaliser et à analyser les expériences de cytométrie et de microscopie confocale.

Table des matières

Index des figures

Index des tableaux

Abréviations

ADN : Acide DésoxyriboNucléique
ADNc : Acide DésoxyriboNucléique complémentaire
ARN : Acide RiboNucléique
BCA : *BiCinchoninic Acid*
BDA : *Bristish Diabetic Association* (Association britannique de diabète)
BSA : *Bovine Serum Albumin*
CGH : *Comparative Genetic Hybridization*
CIG : CosImmGen (Nom de la cohorte de volontaires de l'Institut Pasteur)
CMH : Complexe Majeur d'Histocompatibilité
CNG : Centre National de Génotypage
DL : Déséquilibre de Liaison
DNMT : *DNA Methyl Transferase* (ADN méthyltransférase)
DT1 : Diabète de Type 1
DT2 : Diabète de Type 2
EB : *Elution Buffer* (Tampon d'élution)
ECL : *Enhanced ChemiLuminescence*
ELISA : *Enzyme Linked Immuno Sorbent Assay* (Dosage d'immunoabsorption par enzyme liée)
eQTL : *expression Quantitative Trait Locus* (Locus d'un trait quantitatif d'expression)
eSNP : *expression regulatory Single Nucleotide Polymorphism* (Polymorphisme d'un seul nucléotide régulant l'expression)
EUR : Europe
EWAS : *Epigenome-Wide Association Study* (Etude d'association épigénétique pangénomique)
FID : Fédération Internationale du Diabète
FRET : *Fluorescence Resonance Energy Transfer* (Transfert d'énergie entre molécules fluorescentes)
GAD : *Glutamic Acid Decarboxylase* (Glutamate Acide Décarboxylase)
GWAS : *Genome Wide Association Study* (Etude d'association génétique pangénomique)
HbA1C : Hémoglobine glyquée A1C
HGPO : HyperGlycémie Provoquée par voie Orale
HLA : *Human Leukocyte Antigen* (Antigènes des leucocytes humains)
HRP : *HorseRadish Peroxidase*
IA2 : *tyrosine phosphatase-related islet antigen-2* (Antigènes des îlots 2)
ICAReB : Investigation Clinique et Accès aux Ressources Biologiques
ICA : *Islet Cell Antibodies* (Anticorps anti-Ilots)
ICR : *Imprinting Control Region* (Région de contrôle de l'empreinte)

IPEX : *Immune dysregulation, Polyendocrinopathy,Enteropathy, X-linked* (Dérèglement Immunitaire, polyendocrinopathie et entéropathie, liés à l'X)
JDRF : *Juvenile Diabetes Research Foundation*
JOS : *Joslin Diabetes Center* (Centre de Diabète Joslin)
Kb : Kilobase
KRAB : *Krüppel-Associated Box*
LCL : Lignée Cellulaire Lymphoblastoïde (Lymphocytes B immortalisés avec le virus Epstein-Barr)
LDS : *Lithium Dodecyl Sulfate*
MAF : *Minor Allele Frequency* (Fréquence de l'allèle mineur)
MEB : *Membrane Extraction Buffer* (Tampon d'extraction membranaire)
MIQE : *Minimal Information for qPCR Experiments* (Informations minimales pour des expériences de qPCR)
MODY : *Maturity Onset Diabetes of the Young*
NAm : *North American* (Nord Américain)
NEB : *Nuclear Extraction Buffer* (Tampon d'extraction nucléaire)
NIH : *National Institutes of Health* (Institut national de la santé américain)
NOD : *Non-Obese Diabetic*
OMS : Organisation Mondiale de la Santé
OR : *Odds Ratio*
PBS : *Phosphate Buffered Saline* (Tampon phosphate salin)
PCR : *Polymerase Chain Reaction* (Réaction d'amplification en chaine)
PBMC : *Peripheral Blood Mononuclear Cell* (Cellules mononucléées du sang périphérique)
PNDM : *Permanent Neonatal Diabetes Mellitus* (Diabète sucré néonatal permanent)
PVDF : *PolyVinyliDene Fluoride* (Polydifluorure de vinylidène)
qPCR : *quantitative Polymerase Chain Reaction* (Réaction d'amplification en chaine quantitative)
SDS : *Sodium Dodecyl Sulfate*
SNP : *Single Nucleotide Polymorphism* (Polymorphisme d'un seul nucléotide)
T1DGC : *Type 1 Diabetes Genetics Consortium* (Consortium génétique du diabète de type 1)
Taq : Thermus Aquaticus
TBS : *Tris Buffered Saline* (Tampon Tris salin)
TDT : Transmission Disequilibrium Test (Test du Déséquilibre de Transmission)
TMB : 3,3',5,5'-TetramethylBenzidine
TNDM : *Transient Neonatal Mellitus Diabetes* (Diabète sucré néonatal transitoire)
UK : *United Kingdom* (Royaume-Uni)
WB : Western Blot
WTCCC : *Wellcome Trust Case Control Consortium*
ZFP57 : *Zinc Finger Protein 57*
ZnT8 : *Zinc Transporter 8*

Glossaire

Cohorte de « découverte » et cohorte de « réplication » : En génétique humaine, une cohorte de « découverte » est une cohorte dans laquelle on réalise une première étude d'association génétique. Il est cependant nécessaire de « répliquer » les associations identifiées dans une seconde cohorte indépendante, dite cohorte de « réplication ». La réplication consiste à identifier le même variant avec le même allèle associé à la maladie. La réplication est nécessaire afin de prouver que les premières associations identifiées sont des vrais positifs et non des faux positifs.

Déséquilibre de liaison : Association préférentielle (non aléatoire) entre deux allèles de deux locus génétiquement liés. Une telle association allélique est favorisée en particulier par la proximité physique des locus et modulée par l'activité de recombinaison méiotique. D'autres facteurs peuvent intervenir (démographie, sélection naturelle, interactions épistasiques). Il se mesure à l'aide de deux coefficients normalisés, D' ou r^2.

Diabète : Passage anormal du glucose sanguin dans les urines. Il peut être dû à un trouble de l'assimilation, de l'utilisation ou du stockage des sucres apportés par l'alimentation. Il se traduit aussi par un taux de glucose dans le sang élevé, on parle d'hyperglycémie.

Epigénétique : Modifications de l'expression des gènes qui sont héritables lors de la mitose ou de la méiose, et qui ne résultent pas de modifications de la séquence de l'ADN (définition de Riggs et al., 1996). Ces modifications sont en principe réversibles.

Epistasie : Interaction entre deux ou plusieurs gènes.

Etude d'association allélique : Etude visant à établir les bases génétiques d'un phénotype et consistant à comparer la fréquence des allèles des variants génétiques entre témoins et patients. Un allèle est associé à un phénotype si sa fréquence diffère significativement entre les cas et les témoins. Cet allèle est soit directement le variant causal, soit le plus souvent en déséquilibre de liaison avec lui.

Etude de liaison génétique : Etude visant à établir les bases génétiques d'un phénotype et consistant à tester la co-transmission d'une génération à l'autre du phénotype et des allèles de marqueurs génétiques au sein de familles. Cette co-transmission traduit une liaison génétique, c'est-à-dire un taux de recombinaison génétique inférieur à 50%, entre le locus responsable du phénotype et le marqueur génétique.

eQTL : expression Quantitative Trait Locus ; locus contrôlant le niveau d'expression d'un gène, considéré comme un phénotype quantitatif.

Famille multiplex : Famille comportant au moins deux personnes atteintes d'une maladie donnée.

Famille nucléaire : Famille constituée du noyau parents et de leurs enfants.

Germains : Frères et sœurs d'une même fratrie.

Glycémie : Concentration de glucose dans le sang. La glycémie normale à jeun est comprise entre 0,7 et 1,1 g/L.

Haplotypes : Séquence d'allèles en phase (sur un même chromosome) à des locus voisins liés.

Hémoglobine glyquée (HbA1C) : Proportion d'hémoglobine (principalement l'hémoglobine A) ayant fixé du sucre. Cette fixation est plus importante chez des patients diabétiques du fait de leur hyperglycémie. L'intérêt du dosage de l'HbA1c réside dans le fait que cette valeur reflète la moyenne de la glycémie sur les 3 derniers mois de vie (durée de vie moyenne d'un globule rouge).

Héritabilité : Variabilité phénotypique expliquée par la variabilité génétique.

Hyperglycémie : Glycémie anormalement élevée, supérieure à 1,26 g/L à jeun et 1,4 g/L après une charge en glucose.

Incidence : Nombre de nouveaux cas d'une maladie apparus durant une période de temps donnée.

Insulino-résistance : Baisse de réponse des cellules à l'insuline se traduisant notamment par une diminution de la pénétration du glucose et des acides aminés dans les cellules.

MAF : Fréquence de l'allèle mineur (Minor Allele frequency) dans la population. Pour un SNP, sa fréquence p est comprise entre 0 et 50%, l'autre allèle majeur ayant une fréquence (1-p) comprise entre 50 et 100%. La MAF rapportée est celle dans la population de sujets sains.

Mimétisme moléculaire : Théorie stipulant que certains antigènes exogènes, par exemple d'un agent infectieux viral ou bactérien, peuvent partager des épitopes communs avec des antigènes du soi.

Odds ratio : C'est le rapport des côtes des allèles d'un polymorphisme génétique dans la population de patients atteints et dans celle des contrôles. S'il est significativement différent de 1 alors l'allèle du polymorphisme est dit associé à la maladie. S'il est supérieur à 1, l'allèle associé prédispose ; s'il est inférieur à 1, l'allèle est protecteur. C'est une mesure de l'effet de l'allèle sur le risque de développer la maladie au niveau de la population.

Ome/Omique : L'omique est l'étude globale d'un ensemble moléculaire appelé « ome ». L'étymologie peut être rapportée à la lettre grecque, omega, qui signifie « totalité ». En biologie, le terme omique a d'abord été employé pour la génomique, c'est-à-dire l'étude du génome, donc de l'ensemble de tous les gènes. Par extension, le néologisme « -omique » a été utilisé pour l'étude de l'ensemble de tous les transcrits, ou transcriptome, puis pour l'étude de l'ensemble des protéines, ou protéome. Aujourd'hui, il existe plus d'une dizaine de « omes » dont l'étude est le est le plus souvent réalisée par des technologies à haut débit (https://lhncbc.nlm.nih.gov/files/archive/pub2001047.pdf).

Pénétrance : Proportion d'individus possédant un génotype donné et qui expriment le phénotype correspondant.

Prévalence d'une maladie : Fréquence de la maladie dans la population à un moment donné.

Introduction

1. Le diabète – définition, présentations cliniques, importance en santé publique

1.1. Définition

Le diabète[1] sucré se caractérise par le passage anormal de glucose dans les urines (glycosurie). Etymologiquement, le terme diabète vient du grec διαβαινω (diabainein) qui signifie « traverser », « franchir ». En pratique, c'est une maladie métabolique chronique qui survient principalement lorsque l'organisme est incapable, soit de produire suffisamment d'insuline, soit d'utiliser l'insuline de manière efficace (**Figure 1**), aboutissant à une glycémie, c'est-à-dire la concentration de glucose dans le sang, anormalement régulée. L'insuline est en effet une hormone hypoglycémiante fabriquée par les cellules bêta des îlots de Langerhans du pancréas. Elle permet au glucose de pénétrer dans les cellules de l'organisme grâce à des récepteurs spécifiques, où il est transformé en énergie nécessaire au bon fonctionnement des muscles et des tissus.

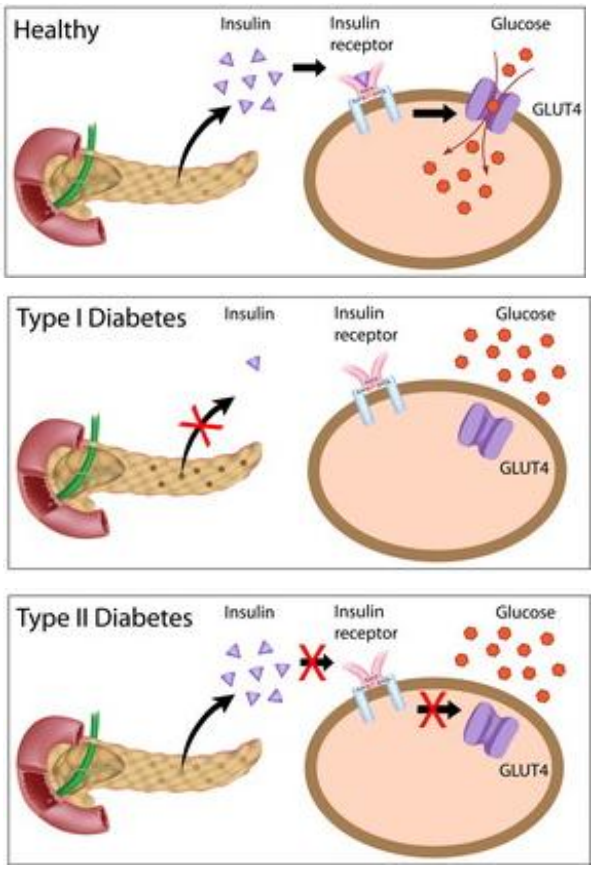


Figure 1. Métabolisme glucidique chez une personne saine, atteinte de diabète de type I ou de diabète de type II. Représentation schématique du déficit insulinique, par défaut de production (diabète de type 1) ou par mauvaise utilisation fonctionnelle (diabète de type 2) chez des patients diabétiques, aboutissant à une hyperglycémie. Illustration produite par Alila Medical Media (http://www.alilamedicalmedia.com/).

[1] La définition des mots en bleu se trouve dans le glossaire

Chez un diabétique, le glucose n'est pas absorbé correctement et s'accumule dans le sang, on parle d'hyperglycémie, source de complications chroniques très invalidantes, oculaires, cardio-vasculaires, rénales, ou encore infectieuses et pouvant mettre en jeu le pronostic vital. Selon l'organisation mondiale de la santé (OMS), un patient est diagnostiqué diabétique s'il présente :

- deux glycémies à jeun supérieures à 1,26 g/L (glycémie à jeun normale entre 0,7 et 1,1 g/L)
- une glycémie deux heures après ingestion de 75g de glucose (test d'hyperglycémie provoquée par voie orale : HGPO) supérieure ou égale à 2,00 g/L
- une mesure de l'hémoglobine glyquée (HbA1c) supérieure à 6,5%

L'hémoglobine est le transporteur sanguin de l'oxygène. Elle peut fixer une partie du sucre, on parle d'hémoglobine glyquée. Elle permet de juger l'équilibre de la glycémie au cours des 2 à 3 mois (durée de vie moyenne d'un globule rouge) qui précèdent un dosage sanguin.

1.2. Epidémiologie

En 2014, selon la Fédération Internationale du Diabète (FID) (http://www.idf.org), 387 millions de personnes étaient diabétiques dans le monde avec un décès causé par le diabète toutes les 7 secondes (http://www.idf.org/sites/default/files/Atlas-poster-2014_FR.pdf) (**Figure 2**).

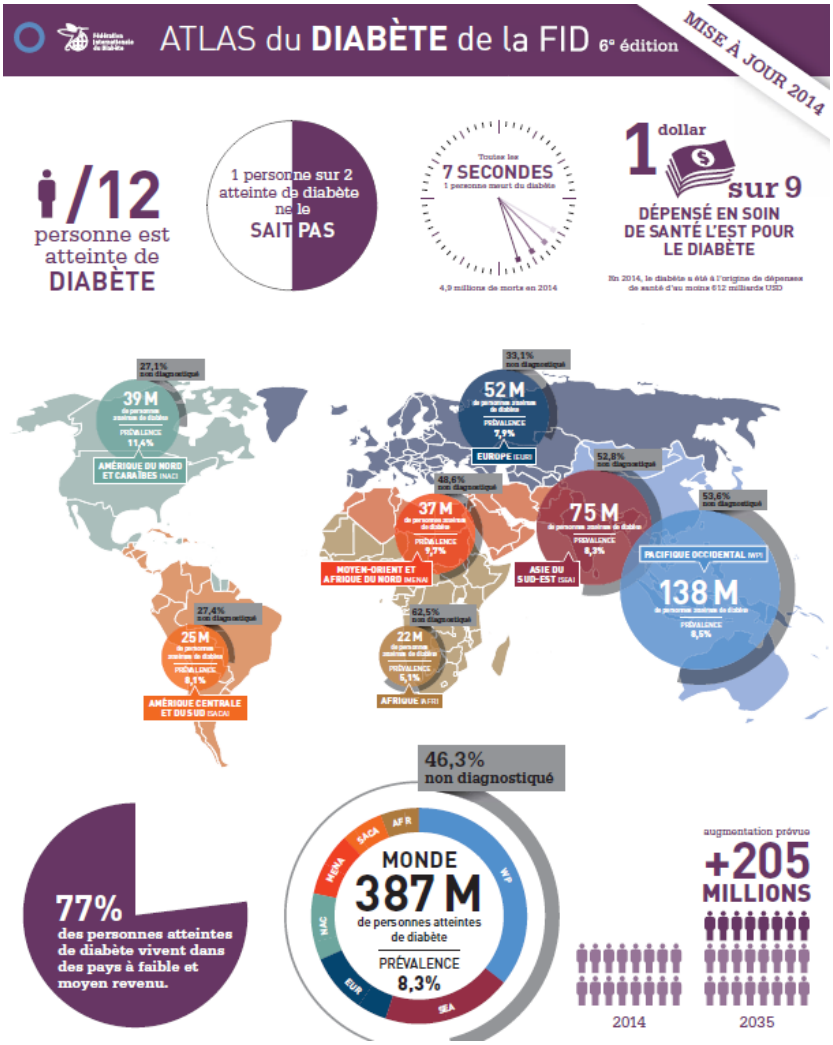


Figure 2. La prévalence mondiale du diabète en 2014 et son impact en santé publique. Poster édité par la Fédération Internationale du Diabète (http://www.idf.org).

Si cette tendance se poursuit, les estimations prévoient 592 millions de diabétiques en 2035, ce qui constitue un problème de santé publique majeur. De plus, le nombre de personnes atteintes de diabète, mais non diagnostiquées, était estimé à 179 millions, soit 46,3 % des diabétiques. Environ 77% des diabétiques vivent dans des pays à faible et moyen revenus et la prévalence de la maladie, c'est-à-dire sa fréquence, est supérieure dans les zones urbaines comparées aux zones rurales, avec une atteinte plus grande des groupes sociaux défavorisés. Les principaux éléments responsables de l'augmentation du diabète sont les changements de style de vie et le vieillissement des populations. En 2014, la tranche d'âge la plus touchée par le diabète était représentée par les 40-59 ans. En France, les données épidémiologiques actuelles indiquent une prévalence des diabètes de 7,2%, soit un peu plus de 3 millions de personnes, entraînant plus de 26 000 décès par an et un coût annuel moyen par patient de 5 000 euros par an.

1.3. Les différents types de diabète

Il existe trois principaux types de diabète : le diabète de type 1, celui de type 2 et le diabète gestationnel.

1.3.1. Le diabète de type 1

Le diabète de type 1 (DT1), aussi appelé diabète insulino-dépendant, est provoqué par une réaction auto-immune au cours de laquelle le système immunitaire attaque les cellules bêta du pancréas qui produisent l'insuline. Les lymphocytes T reconnaissent des molécules du « soi » présentes dans les cellules bêta et développent une réaction délétère contre ces cellules. Cette réaction se manifeste également par la production d'auto-anticorps dirigés contre les cellules des îlots de Langerhans (ICA= islet cell antibodies). Certaines des cibles de ces anticorps sont connues au niveau moléculaire : insuline, GAD (Glutamic Acid Decarboxylase), ZnT8 (Transporteur de Zinc) et tyrosine phosphatase (IA2). Les symptômes apparaissent plusieurs mois voire plusieurs années après le début de ces événements, quand plus de 80 % des cellules bêta ont été détruites. Cette destruction est irréversible et l'organisme devient alors incapable de fabriquer l'insuline dont il a besoin (**Figure 1**).

En général, le DT1 apparaît de manière rapide, en quelques jours à quelques semaines, et se traduit par un syndrome cardinal associant polyurie-polydipsie (l'augmentation du volume d'eau bue par jour et l'augmentation du volume d'urines émises par jour), polyphagie (sensation excessive de faim), amaigrissement et asthénie. Le diagnostic de la maladie passe par la détection des auto-anticorps circulants dans le sang, chez les personnes dont la glycémie est élevée. Cet outil diagnostique permet notamment de distinguer cette forme de diabète des autres. Le traitement substitutif par l'insuline est vital pour les personnes atteintes de DT1. Il se fait par voie parentérale, les doses devant être ajustées sur la glycémie. Les personnes atteintes de DT1 peuvent mener une vie normale grâce à la combinaison d'une insulinothérapie quotidienne, d'une éducation, d'une surveillance étroite, d'une alimentation saine et de la pratique régulière d'une activité physique.

Le DT1 est la troisième maladie chronique de l'enfant, avec une prévalence de 0,4% dans les populations d'origine européenne, soit un peu moins de 10% de l'ensemble des diabètes, ce qui représente environ 200 000 cas en France. La maladie touche également les hommes et les femmes. Le DT1 apparait dans la moitié des cas chez l'enfant et chez l'adolescent. Depuis les

années 1950, l'incidence, c'est-à-dire le nombre de nouveaux cas par an, est en augmentation dans de nombreux pays. Surtout, on observe un début de plus en plus précoce, avec un doublement attendu du nombre de cas inférieurs à 5 ans entre 2005 et 2020 en Europe (Ehehalt et al. 2012). En particulier, les formes du nourrisson ne sont plus rares. Les raisons de cette évolution ne sont pas encore connues avec certitude mais des modifications des facteurs de risque environnementaux, des événements survenant aux premiers stades de la grossesse, l'alimentation au début de la vie et des infections virales pourraient jouer un rôle.

1.3.2. Le diabète de type 2

Le diabète de type 2 (DT2) est la forme la plus courante de la maladie (environ 90% des diabétiques). Il touche généralement les adultes, mais là aussi, probablement sous l'effet de l'évolution des habitudes alimentaires notamment, les formes de l'enfant sont de plus en plus fréquentes. Chez les personnes atteintes de DT2, l'organisme est capable de produire de l'insuline, mais les cellules sont résistantes à l'action de cette hormone, ce qui entraîne une accumulation de glucose dans le sang car il ne pénètre pas ou mal les cellules (**Figure 1**). La production d'insuline peut devenir insuffisante voire déficiente, nécessitant une insulino-thérapie. Cette maladie est longtemps latente cliniquement alors même que l'hyperglycémie prolongée donne lieu aux complications chroniques du diabète déjà évoquées. Bien que l'étiologie du DT2 soit encore inconnue, il existe plusieurs facteurs de risque importants et bien établis : l'obésité, une alimentation mal équilibrée, l'inactivité physique, un âge avancé, des antécédents familiaux de diabète, l'ethnie, une glycémie élevée pendant la grossesse qui affecte l'enfant à naître.

1.3.3. Le diabète gestationnel

Il n'est pas rare qu'un diabète apparaisse au cours d'une grossesse, faisant suite à l'inhibition de l'action de l'insuline (également appelée insulino-résistance) probablement par les hormones placentaires. Ce diabète disparaît généralement après la grossesse. Cependant les femmes ayant eu un diabète gestationnel et leurs enfants ont un risque plus important de développer un DT2.

1.3.4. Autres formes de diabète sucré

Parmi les autres formes de diabète sucré (1 à 5%), il faut mentionner les diabètes iatrogènes, au cours d'un traitement corticoïde prolongé. Leur risque impose des précautions et une surveillance particulières lors de ces traitements. Il existe également des formes beaucoup plus rares de diabète sucré monogéniques, notamment néonatales (par exemple l'IPEX (Immune dysregulation, Polyendocrinopathy, Enteropathy, X-linked) (**Tableau 1**), ou touchant l'adulte (MODY= Maturity Onset Diabetes of the Young) (**Tableau 2**), ou syndromiques (associés à d'autres manifestations cliniques comme les diabètes mitochondriaux ou le syndrome de Wolfram) (**Tableau 3**). Des mutations de 7 de ces gènes (*ABCC8*, *GCK*, *HNFB1*, *INS*, *KCNJ11*, *NEUROD1* et *PDX1*) peuvent conduire à un diabète néonatal ou à un MODY (Yang et Chan 2016) (**Tableau 1 et 2**).
Ces formes rares sont particulièrement importantes du point de vue cognitif car l'identification de leur gène a apporté des informations essentielles sur la physiologie de la cellule bêta et du système immunitaire. De façon remarquable, des variants de ces gènes ont été associés aux formes de diabète les plus fréquents : deux de ces gènes (*INS* et *GLIS3*) sont actuellement connus comme étant associés au DT1 et au DT2, douze autres sont associés au DT2 uniquement (**Figure 3**).

Tableau 1. Gènes mutés dans les diabètes néonataux. TNDM, Transient Neonatal Diabetes Mellitus ; PNDM, Permanent Neonatal Diabetes Mellitus ; *, Gène également muté dans une forme de diabète syndromique

Diabète néonatal	Symbole du Gène	Nom du Gène	Chromosome	Protéine	Fonction
PNDM/ TNDM	*ABCC8*	ATP-binding Cassette, sub-family C (CFTR/MRP), member 8	11p15	Sous-unité de Kir6.2	Sous-unité d'un canal ATP-K qui contrôle la sécrétion d'insuline et le taux de sucre dans le sang
PNDM	*EIF2AK3**	Eukaryotic translation Initiation Factor 2-alpha Kinase 3	2p12	Kinase	Répression de la synthèse globale des protéines
PNDM/ IPEX	*FOXP3**	Forkhead box P3	Xp11	Facteur de transcription	Contrôle l'activité de gènes impliqués dans la régulation du système immunitaire
TNDM/ PNDM	*GATA4*	Globin Transcription Factor binding protein 4	8p23	Facteur de transcription	Régule l'expression de gènes impliqués dans l'embryogenèse, la différentiation et fonction myocardiqaque et le développement testiculaire
PNDM	*GATA6*	Globin Transcription Factor binding protein 6	18q11	Facteur de transcription	Important dans la régulation de la différenciation cellulaire et l'organogenèse durant le développement
PNDM	*GCK*	Glucokinase (hexokinase 4)	7p13	Enzyme glycolytique	Métabolisme du glucose
PNDM	*GLIS3*	Gli-similar family zinc finger 3	9p24	Facteur de transcription	Activateur ou répresseur de la transcription- Impliqué dans le développement des cellules bêta du pancréas, de la thyroïde, des yeux, du foie et du rein
PNDM/ TNDM	*HNF1B*	Hepatocyte Nuclear Factor 1 homeobox B	17q12	Facteur de transcription	Fonction dans le développement du rein et régulation du développement du pancréas embryonnaire
TNDM	*HYMA1*	Hydatidiform mole–associated and imprinted transcript	6q24	ARNm non codant	Gène soumis à empreinte
PNDM	*IER3IP1*	Immediate Early Response 3 Interacting Protein 1	18q12	Protéine du réticulum endoplasmique	Régulation de l'apoptose
PNDM	*INS*	Insulin	11p15	Hormone	Contrôle de la glycémie
PNDM/ TNDM	*KCNJ11*	Potassium Channel, inwardly rectifying subfamily J, member 11	11p15	Sous-unité du canal ATP-K Kir6.2	Sous-unité d'un canal ATP-K qui contrôle la sécrétion d'insuline et le taux de sucre dans le sang
PNDM	*MNX1*	Motor Neuron And Pancreas Homeobox 1	7q36	Facteur de transcription	Protéine nucléaire avec un domaine homéotique
PNDM	*NEUROD1*	Neuronal Differentiation 1	2q32	Facteur de transcription	Régulation de l'expression du gène de l'insuline
PNDM	*NEUROG3*	Neurogenin 3	10q21	Facteur de transcription	Impliqué dans la neurogénèse et requis pour le développement des cellules endocrines du pancréas et de l'intestin
PNDM	*NKX2-2*	NK2 Homeobox 2	20p11	Activateur transcriptionnel	Morphogenèse du système nerveux central
TNDM	*PAX6*	Paired Box 6	11p13	Facteur de transcription	Développement des yeux et du cerveau
	PCBD1	Pterin-4 Alpha-Carbinolamine Dehydratase/Dimerization Cofactor Of Hepatocyte Nuclear Factor 1 Alpha	10q22		
PNDM	*PDX1*	Pancreatic and Duodenal homeobox 1	13q12	Facteur de transcription	Activateur transcriptionnel de plusieurs gènes dont l'insuline
TNDM	*PLAGL1*	Pleiomorphic Adenoma Gene-Like 1	6q24	Régulateur transcriptionnel	Propriété anti-prolifération, agit comme un suppresseur de tumeur
PNDM	*PTF1A*	Pancreas specific transcription factor, 1a	10p12	Facteur de transcription	Rôle dans le développement du pancréas
PNDM	*RFX6*	Regulatory Factor X, 6	6q22	Facteur de transcription	Différenciation des ilots pendant le développement du pancréas endocrine
PNDM	*SLC2A2*	Solute Carrier Family 2 (facilitated glucose transporter), member 2	3q26	Transporteur du glucose GLUT2	Glycoprotéine de la membrane plasmique du foie, des cellules bêta, de l'intestin et de l'épithélium du rein- Facilite le transport du glucose
	*SLC19A2**	Solute Carrier family 19 (thiamine transporter), member 2	1q23	Transporteur de la vitamine B1	Aide à la conversion des carbohydrates en énergie
	TRMT10A	TRNA Methyltransferase 10A	4q23	ARNt méthyltransférase	Catalyse la méthylation de la guanine 9 dans les ARNt
	*WFS1**	Wolfram syndrome 1 (wolframin)	4p16	Protéine du réticulum endoplasmique	Régule la quantité de calcium-Aide au repliement du précurseur de l'insuline en hormone mature
TNDM	*ZFP57*	Zinc Finger Protein 57	6p22	Facteur de répression de la transcription	Contrôle la méthylation de l'ADN durant le développement

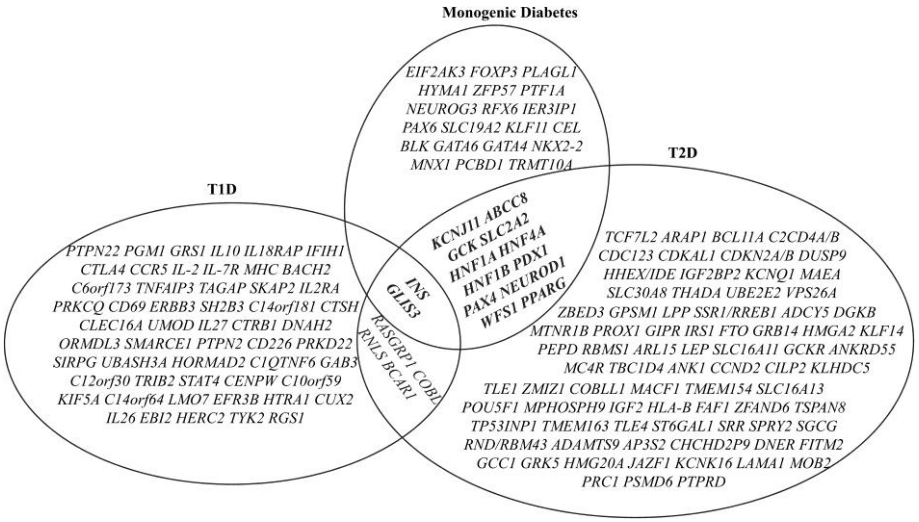


Figure 3. Diagramme de Venn des gènes/locus associés au DT1 ou au DT2 et connus dans les diabètes monogéniques. Plus de 50 locus sont impliqués dans le DT1 et en moyenne 100 dans le DT2. *INS*, *GLIS3*, *RASGRP*, *COBL*, *RNLS*, and *BCAR1* sont associés à la fois au DT1 et au DT2. Environ 1/3 des gènes impliqués dans des diabètes monogéniques sont également associés au DT2. *INS* et *GLIS3* sont deux gènes de diabètes monogéniques dont les variants sont associés au DT1 et au DT2. Figure tirée d'un article publié en 2016 dans Endocrine Reviews (Yang et Chan 2016).

Tableau 2. Différents sous-types de diabète MODY et leurs gènes mutés

Diabète MODY	Symbole du gène	Nom du gène	Bande cytogénétique	Protéine	Fonction
MODY 1	*HNF4A*	Hepatocyte Nuclear Factor 4, alpha	20q13	Facteur de transcription	Régulation de l'expression de gènes hépatiques
MODY 2	*GCK*	Glucokinase (hexokinase 4)	7p13	Enzyme glycolytique	Métabolisme du glucose
MODY 3	*HNF1A*	HNF1 homeobox A	12q24	Facteur de transcription	Régulation de l'expression de gènes hépatiques et des ilots pancréatiques
MODY 4	*PDX1*	pancreatic and duodenal homeobox 1	13q12	Facteur de transcription	Activateur transcriptionnel de plusieurs gènes dont l'insuline
MODY 5	*HNF1B*	Hepatocyte Nuclear Factor 1 homeobox B	17q12	Facteur de transcription	Fonction dans le développement du rein et régulation du développement du pancréas embryonnaire
MODY 6	*NEUROD1*	Neuronal Differentiation 1	2q31	Facteur de transcription	Régulation de l'expression de l'insuline
MODY 7	*KLF11*	Kruppel-Like Factor 11	2p25	Facteur de transcription	Liaison epsilon et gamma globine Inhibition de la croissance cellulaire, favorise l'apoptose
MODY 8	*CEL*	Carboxyl Ester Lipase	9q34	Lipase	Sécrétée par le pancréas et la glande mammaire. Hydrolyse et absorption de lipide
MODY 9	*PAX4*	Paired box 4	7q32	Facteur de transcription	Différenciation et développement des cellules Bêta
MODY 10	*INS*	Insulin	11p15	Hormone	Contrôle de la glycémie
MODY 11	*BLK*	B Lymphocyte Kinase proto-oncogene, Src family tyrosine kinase	8p23	Facteur de transcription	Développement, différenciation et signalisation des lymphocytes B
MODY 12	*ABCC8*	ATP-binding Cassette, sub-family C (CFTR/MRP), member 8	11p15	Sous-unité de Kir6.2	Sous-unité d'un canal ATP-K qui contrôle la sécrétion d'insuline et le taux de sucre dans le sang
MODY 13	*KCNJ11*	Potassium Channel, inwardly rectifying subfamily J, member 11	11p15	sous-unité du canal ATP-K Kir6.2	Sous-unité d'un canal ATP-K qui contrôle la sécrétion d'insuline et le taux de sucre dans le sang

Tableau 3. Gènes mutés dans des formes syndromiques de diabètes

Diabète syndromique	Symbole du gène	Nom du gène	Chromosome	Protéine	Fonction
Lipodystrophie congénitale	*AGPAT2*	1-acylglycerol-3-phosphate O-acyltransferase 2	9q34	Enzyme du réticulum endoplasmique	Croissance et développement des adipocytes
Syndrome APECED (polyencocrinopathie auto-immune de type 1)	*AIRE*	Autoimmune Regulator	21q22	Facteur de transcription	Rôle dans la tolérance centrale
Syndrome d'Alström	*ALMS1*	Alstrom Syndrome Protein 1	2p13	Protéine du centrosome	Rôle dans l'organisation des microtubules
Lipodystrophie congénitale	*BSCL2*	Berardinelli-Seip Congenital Lipodystrophy 2 (seipin)	11q13	Protéine transmembranaire du réticulum endoplasmique	Fonction inconnue-Potentiel rôle dans le développement précoce des adipocytes
Syndrome de Wolfram	*CISD2*	CDGSH iron sulfur domain 2	4q24	Protéine du réticulum endoplasmique	Fonction inconnue mais protéine retrouvée dans la membrane externe de la mitochondrie
Syndrome de Wolcott-Rallison	*EIF2AK3*	Eukaryotic Translation Initiation Factor 2-Alpha Kinase 3 = protein kinase R-like endoplasmic reticulum kinase (PERK)	2p12	Kinase	Répression de la synthèse globale des protéines, impliquées dans le contrôle de la morphologie et de la fonction mitochondriale
IPEX (Dérèglement Immunitaire, polyendocrinopathie et entéropathie, liés à l'X)	*FOXP3*	Forkhead box P3	Xp11	Facteur de transcription	Contrôle l'activité de gènes impliqués dans la régulation du système immunitaire
Ataxie de Friedreich	*FXN*	Frataxin	9q13	Protéine mitochondriale	Régulation mitochondriale
Lépréchaunisme et Syndrome de Rabson-Mendenhall	INSR	Insulin Receptor	19p13	Récepteur de l'insuline	Contrôle de la glycémie
Divers : hypertension, lipodystrophie partielle…	*PPARG*	Peroxisome Proliferator-Activated Receptor Gamma	3p25	Récepteur nucléaire	Régulateur de la différenciation des adipocytes et de l'homéostasie du glucose
Microcéphalie, petite taille et altération du métabolisme du glucose	*PPP1R15B*	protein phosphatase 1 regulatory subunit 15B	1q32.1	Phosphatase	Déphosphorylation du facteur de transcription EIF2α
Anémie mégaloblastique thiamine dépendante	*SLC19A2*	Solute Carrier family 19 (thiamine transporter), member 2	1q23	Transporteur de la vitamine B1	Aide à la conversion des carbohydrates en énergie
Syndrome de Wolfram	*WFS1*	Wolfram syndrome 1 (wolframin)	4p16	Protéine du réticulum endoplasmique	Régule la quantité de calcium-Aide au repliement du précurseur de l'insuline en hormone mature
Syndrome de Werner	*WRN*	Werner syndrome, RecQ helicase-like	8p12	Exonucléase	Rôle dans la réparation de l'ADN

2. Le diabète de type 1 : une étiologie multifactorielle

Le diabète de type 1, comme la majorité des maladies auto-immunes, est une maladie à hérédité complexe, polygénique, hétérogène et multifactorielle qui nécessite l'interaction de facteurs génétiques et environnementaux (Wandstrat et Wakeland 2001). Aucun gène ne peut expliquer à lui seul la maladie. De plus des données récentes mettent en lumière le rôle possible de mécanismes épigénétiques.

2.1. Facteurs génétiques

L'existence d'un terrain génétique de susceptibilité au DT1 est démontrée par deux types d'études : d'une part par des études d'agrégation familiale, c'est-à-dire par la mise en évidence d'un risque accru de récurrence chez les apparentés de patients en comparaison avec le risque de la population générale (exemple : pour des germains, le risque de récurrence = 6% versus une prévalence de 0,4% dans la population générale, soit un risque $\lambda_{sibling} = \lambda s = 6/0,4 = 15$) ; d'autre part par des études de taux de concordance chez des jumeaux, avec un taux de concordance chez les jumeaux monozygotes (~30-50%) supérieur à celui de jumeaux dizygotes (~6-7%).
Plus précisément, ces facteurs génétiques ont été identifiés à l'aide d'études de liaison génétique (dans des familles multiplexes) et surtout plus récemment grâce à des études d'association allélique (voir paragraphe 3 pour le principe de ces méthodes).
Ainsi, des études d'association pangénomique (Genome Wide Association Study - GWAS) effectuées notamment sur la cohorte du Type I Diabetes Genetics Consortium (T1DGC-www.t1dgc.org) ont permis de localiser à ce jour 52 régions génétiques associées significativement ($p=5.10^{-8}$ pour une association pangénomique) et 6 régions associées de manière suggestive (au seuil 10^{-5}), dans la susceptibilité au DT1 (Barrett et al. 2009 ; Onengut-Gumuscu et al. 2015) (www.t1dbase.org et www.immunobase.org) (**Tableau 4**), sans toutefois identifier tous les polymorphismes causaux. Le principe de ces études pangénomiques est de tester l'association entre des gènes, plus précisément leurs allèles, et le phénotype étudié (la maladie) chez un grand nombre d'individus.
Parmi toutes ces régions génétiques, l'une domine largement. Il s'agit du Complexe Majeur d'Histocompatibilité (CMH) situé sur le bras court du chromosome 6.
Pour cette région, l'Odds Ratio (OR), qui mesure l'amplitude de l'effet, peut atteindre 11,37 pour les allèles les plus fortement associés alors que pour toutes les autres régions, il ne dépasse pas 2,5 (**Tableau 4 et Figure 4**). Ainsi, par exemple l'OR est de 2,38 pour l'insuline (*INS*) en 11p15, pour *PTPN22* en 1p13 il est de 1,89 et pour *CTLA4* en 2q33 il est de 1,19.

Tableau 4. Régions génétiques associées au DT1 listées par position génomique. Tableau incluant les associations significatives issues du dernier GWAS sur le DT1 réalisé à l'aide de l'Immunochip (Onengut-Gumuscu et al. 2015) ainsi que les autres régions précédemment associées au DT1 recensées sur le site T1DBase. Sont indiqués les SNPs les plus fortement associés (seuil p $<5.10^{-8}$), avec l'effet de l'allèle mineur comparé à celui de l'allèle majeur. Pour 6 régions indiquées en gras (régions n°5, 17, 30, 34,37 et 48) les associations sont suggestives ($p<10^{-5}$ pour le DT1 mais $p<5x10^{-8}$ pour au moins une autre maladie auto-immune ou inflammatoire). Les régions surlignées en orange sont les trois régions prédominantes ayant les effets les plus élevés (**Figure 4**). Les maladies auto-immunes et inflammatoires associées au locus identifié selon immunobase sont indiquées dans la dernière colonne (AA, Alopécie Areata ; ATD, Thyroïdite auto-immune ; CEL, Maladie cœliaque ; CRO, Maladie de Crohn ; IBD, Maladie Inflammatoire de l'instestin ; JIA, Arthrite Idiopathique Juvénile ; MS, Sclérose en plaques ; NAR, Narcolepsie ; PBC, Cirrhose Biliaire Primitive ; PSO, Psoriasis ; RA , Polyarthrite rhumatoïde ; SLE, Lupus érythémateux disséminé ; UC, Rectocolite hémorragique ; VIT, Vitiligo ; MAF, Minor Allele Frequency).

N° région	Bande cytogénétique	Top SNP	Position du SNP (hg19)	Allèle Majeur>Allèle Mineur	MAF	OR	p-value	Gène(s) candidat(s)	Maladies associées dans immunobase
1	1p13.2	rs2476601	114,377,568	G>A	0,09	1,89	1×10^{-122}	*PTPN22*	RA, CRO, SLE, VIT, ATD, AA, JIA
2	1q32.1	rs6691977	200,814,959	T>C	0,19	1,13	$4,3 \times 10^{-8}$		
3	1q32.1	rs3024505	206,939,904	G>A	0,17	0,84	$1,9 \times 10^{-9}$	*IL10*	SLE, CRO, UC, IBD
4	2p23.3	rs478222	25,301,755	A>T	0,40	0,87	$3,5 \times 10^{-9}$		
5	**2q11.2**	**rs13415583**	**100,764,087**	**T>G**	**0,35**	**0,9**	**$1,1 \times 10^{-7}$**	**AFF3**	**RA**
6	2q13	rs4849135	111,615,079	G>T	0,29	0,89	$4,4 \times 10^{-8}$		
7	2q24.2	rs2111485	163,110,536	G>A	0,39	0,85	$3,8 \times 10^{-18}$	*IFIH1*	VIT, IBD, UC, PSO, SLE
8	2q33.2	rs3087243	204,738,919	G>A	0,45	0,84	$7,4 \times 10^{-21}$	*CTLA4*	CEL, RA, ATD
9	3p21.31	rs113010081	46,457,412	T>C	0,11	0,85	$4,6 \times 10^{-8}$	*CCR5*	CEL
10	4p15.2	rs10517086	26,085,511	G>A	0,30	1,09	$4,6 \times 10^{-10}$		RA
11	4q27	rs75793288	123,243,596	C>G	0,36	1,15	$5,6 \times 10^{-13}$	*IL2*, *IL21*	CEL, UC
12	4q32.3	rs2611215	166,574,267	G>A	0,15	1,18	$1,8 \times 10^{-11}$		
13	5p13.2	rs11954020	35,883,251	C>G	0,39	1,11	$4,4 \times 10^{-8}$	*IL7R*	
14	6p21.3	rs6916742	32,453,191	C>T	0,49	0,24	4×10^{-307}	Région du CMH	CEL,MS,RA,T1D
15	6q15	rs72928038	90,976,768	G>A	0,17	1,20	$6,4 \times 10^{-14}$	*BACH2*	CEL, ATD, RA, MS
16	6q22.32	rs1538171	126,752,884	C>G	0,45	1,12	$7,4 \times 10^{-10}$		
17	**6q23.3**	**rs6920220**	**138,006,504**	**G>A**	**0,22**	**1,12**	**$7,26 \times 10^{-6}$**		**CEL, IBD, PBC, RA, SLE, UC**
18	6q25.3	rs1738074	159,465,977	C>T	0,41	0,92	$7,59 \times 10^{-9}$	*TAGAP*	CEL, MS
19	6q27	rs924043	170,379,025	C>T	0,01	0,84	$8,6 \times 10^{-9}$		
20	7p15.2	rs7804356	26,891,665	T>C	0,24	0,88	$5,3 \times 10^{-9}$		
21	7p12.2	rs62447205	50,465,830	A>G	0,28	0,89	$2,5 \times 10^{-8}$	*IKZF1*	
22	7p12.1	rs4948088	51,027,194	C>A	0,05	0,77	$4,4 \times 10^{-8}$		
23	9p24.2	rs6476839	4,290,823	A>T	0,4	1,12	$1,0 \times 10^{-9}$	*GLIS3*	
24	10p15.1	rs61839660	6,094,697	C>T	0,1	0,62	$2,8 \times 10^{-39}$	*IL2RA, RBM17*	MS, RA, VIT
25	10p15.1	rs11258748	6,472,891	C>T	0,13	0,69	$9,84 \times 10^{-9}$	*PRKCQ*	
26	10p11.22	rs722988	33,426 ,147	A>G	0,37	1,11	$4,88 \times 10^{-8}$	*NRP1*	
27	10q23.31	rs12416116	90,035,654	C>A	0,28	0,85	$3,9 \times 10^{-15}$		
28	11p15.5	rs798968	2,050,299	A>T	0,22	1,25	$2,8 \times 10^{-14}$	*INS*	
29	11p15.5	rs689	2,182,224	T>A	0,3	0,42	$<10^{-100}$	*INS*	
30	**11q13.1**	**rs694739**	**64,097,233**	**T>C**	**0,37**	**0,95**	**$2,37 \times 10^{-7}$**	**BAD**	**AA, CRO, MS, PBC**
31	12p13.31	rs4763879	9,910,164	G>A	0,37	1,09	$1,9 \times 10^{-11}$	*CD69*	MS, T1D
32	12q13.13	rs11170466	53,585,859	C>T	0,05	1,19	$7,86 \times 10^{-9}$	*ITGB7*	
33	12q13.2	rs705705	56,435,504	G>C	0,34	1,25	$4,4 \times 10^{-32}$	*IKZF4*	AA
34	**12q14.1**	**rs10877012**	**58,162,085**	**G>T**	**0,33**	**0,82**	**$3,8 \times 10^{-6}$**		**MS**
35	12q24.12	rs653178	112,007,756	T>C	0,48	1,30	$1,6 \times 10^{-44}$	*SH2B3*	CEL, PBC, AA, JIA, PSC, RA, VIT
36	13q32.3	rs9585056	100,081,766	T>C	0,24	1,12	$3,3 \times 10^{-8}$	*GPR183*	
37	**14q24.1**	**rs911263**	**68,753,593**	**T>C**	**0,29**	**0,89**	**$4,93 \times 10^{-6}$**		**PBC, T1D**
38	14q24.1	rs1465788	69,263,599	G>A	0,29	0,86	$1,8 \times 10^{-12}$		
39	14q32.2	rs1456988	98,488,007	T>G	0,27	1,12	$2,9 \times 10^{-8}$		
40	14q32.2	rs56994090	101,306,447	T>C	0,41	0,88	$1,1 \times 10^{-11}$		
41	15q14	rs72727394	38,847,022	C>T	0,19	1,15	$3,6 \times 10^{-10}$	*RASGRP1*	
42	15q25.1	rs34593439	79,234,957	G>A	0,1	0,78	$9,0 \times 10^{-14}$	*CTSH*	CEL, NAR
43	16p13.13	rs12927355	11,194,771	C>T	0,32	0,82	$3,0 \times 10^{-22}$	*DEXI*	PBC, MS
44	16p11.2	rs151234	28,505,660	G>C	0,12	1,19	$4,8 \times 10^{-11}$	*IL27*	CRO
45	16q23.1	rs8056814	75,252,327	G>A	0,07	1,32	$3,0 \times 10^{-19}$	*BCAR1*	
46	17q12	rs12453507	38,053,207	G>C	0,49	0,90	$1,0 \times 10^{-8}$	*IKZF3*, *ORMDL3*, *GSDMB*	PBC, RA
47	17q21.2	rs7221109	38,770,286	C>T	0,36	0,95	$1,3 \times 10^{-9}$	*CCR7*	
48	**17q21.31**	**rs1052553**	**44,073,889**	**A>G**	**0,24**	**0,89**	**$8,2 \times 10^{-8}$**		**PBC**
49	18p11.21	rs1893217	12,809,340	A>G	0,16	1,21	$1,2 \times 10^{-15}$	*PTPN2*	CEL, CRO, IBD, UC
50	18q22.2	rs1615504	67,526,644	C>T	0,47	1,13	$1,8 \times 10^{-11}$	*CD226*	CEL, MS
51	19p13.2	rs34536443	10,463,118	G>C	0,04	0,67	$4,4 \times 10^{-15}$	*TYK2*	CRO, SLE, PBC, PSO, RA, JIA, MS
52	19q13.32	rs402072	47,219,122	T>C	0,16	0,87	$4,7 \times 10^{-8}$		
53	19q13.33	rs516246	49,206,172	T>C	0,49	0,87	$5,2 \times 10^{-14}$	*FUT2*	IBD, CRO
54	20p13	rs6043409	1,616,206	G>A	0,35	0,88	$3,0 \times 10^{-10}$		
55	21q22.3	rs11203202	43,825,357	C>G	0,33	1,16	$1,2 \times 10^{-15}$	*UBASH3A*	RA, VIT
56	22q12.2	rs4820830	30,531,091	T>C	0,38	1,14	$1,2 \times 10^{-12}$		
57	22q12.3	rs229533	37,587,111	A>C	0,43	1,11	$1,8 \times 10^{-8}$	*C1QTNF6*, *RAC2*	
58	Xq28	rs2664170	153,945,602	A>G	0,32	1,16	$7,8 \times 10^{-9}$		

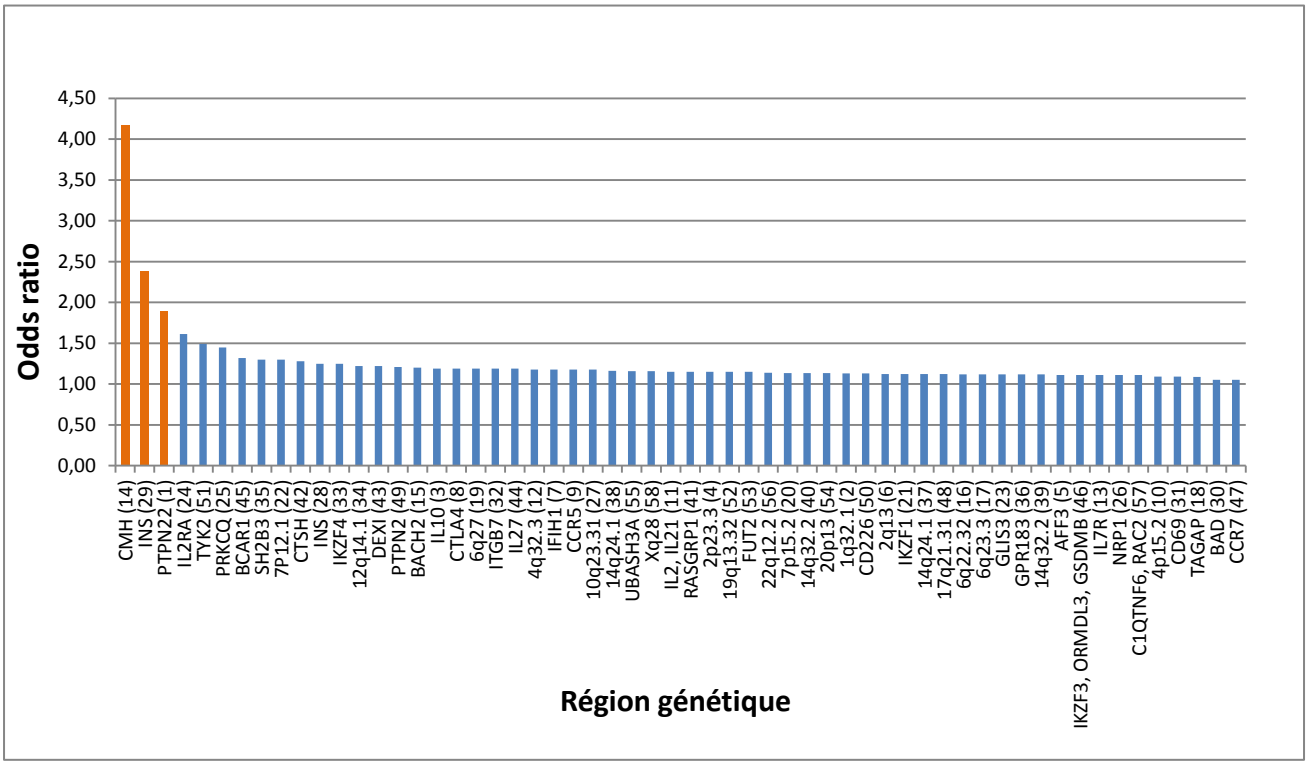


Figure 4. Odds ratio des différentes régions impliquées dans la susceptibilité au DT1. Figure représentant l'amplitude des effets des régions associées au DT1 classées par ordre décroissant de l'odds ratio exprimé par rapport à l'allèle de risque. Le numéro de la région selon sa position génomique (cf. **Tableau 4**) est donné entre parenthèse après le nom du gène candidat, s'il est connu, ou la bande cytogénétique. Les régions surlignées en orange sont les trois régions prédominantes avec les odds ratios les plus élevés.

2.2. Complexe Majeur d'Histocompatibilité (CMH)

Le Complexe Majeur d'Histocompatibilité (CMH) a été découvert il y a plus de 50 ans pour son rôle dans la compatibilité des greffes tissulaires entre individus. Il est localisé sur le bras court du chromosome 6 au niveau de la bande cytogénétique 6p21.3 (**Figure 5**). Il est subdivisé en 3 régions :

• La région de classe I, du côté télomérique, comprend les gènes HLA (Human Leukocyte Antigen) de classe I, tels que *HLA-A*, *HLA-B*, et *HLA-C*, dont les produits sont présents, combinés à la beta2-microglobuline, à la surface de toutes les cellules et sont impliqués dans la présentation des peptides antigéniques aux lymphocytes T $CD8^+$.

• La région dite de classe II, située du côté centromérique, comprend de nombreux gènes et pseudogènes, dont les trois paires de gènes HLA de classe II, codant les chaines alpha et beta des molécules HLA-DP, HLA-DQ et HLA-DR, exprimées à la surface des cellules spécialisées dans la présentation des peptides antigéniques (ex : cellules dendritiques). Certains de ces gènes comme *HLA-DRB1* peuvent être dupliqués et présents jusqu'à 9 copies.

• Au milieu, la région de classe III contient des gènes codants pour des protéines du système du complément (C2, C4, facteur B) et des cytokines pro-inflammatoires, telles que le TNF (Tumor Necrosis Factor) et les lymphotoxines alpha et bêta.

Ces trois régions comprennent de nombreux autres gènes, ayant ou non une fonction immunologique (Horton et al. 2004).

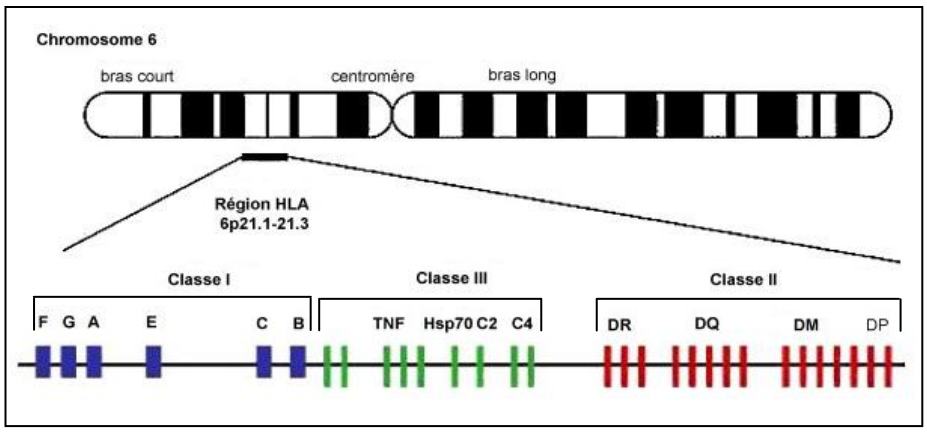


Figure 5. Localisation cytogénétique du CMH sur le bras court du chromosome 6 et structure schématique en trois classes. Le CMH est classiquement divisé en 3 régions incluant les gènes HLA de classe I (en bleu) du coté télomérique, les gènes HLA de classe II (en rouge) du côté centromérique, et entre les deux, la région dite de classe III incluant des gènes codant des cytokines pro-inflammatoires ou des protéines de la cascade du complément (en vert). Les coordonnées physiques (selon hg19) sur le chromosome 6 des trois classes du CMH sont : 29,640,260-31,485,656 pour la classe I, 31,485,656-32,228,466 pour la classe III et 32,228,466-33,123,113 pour la classe II.

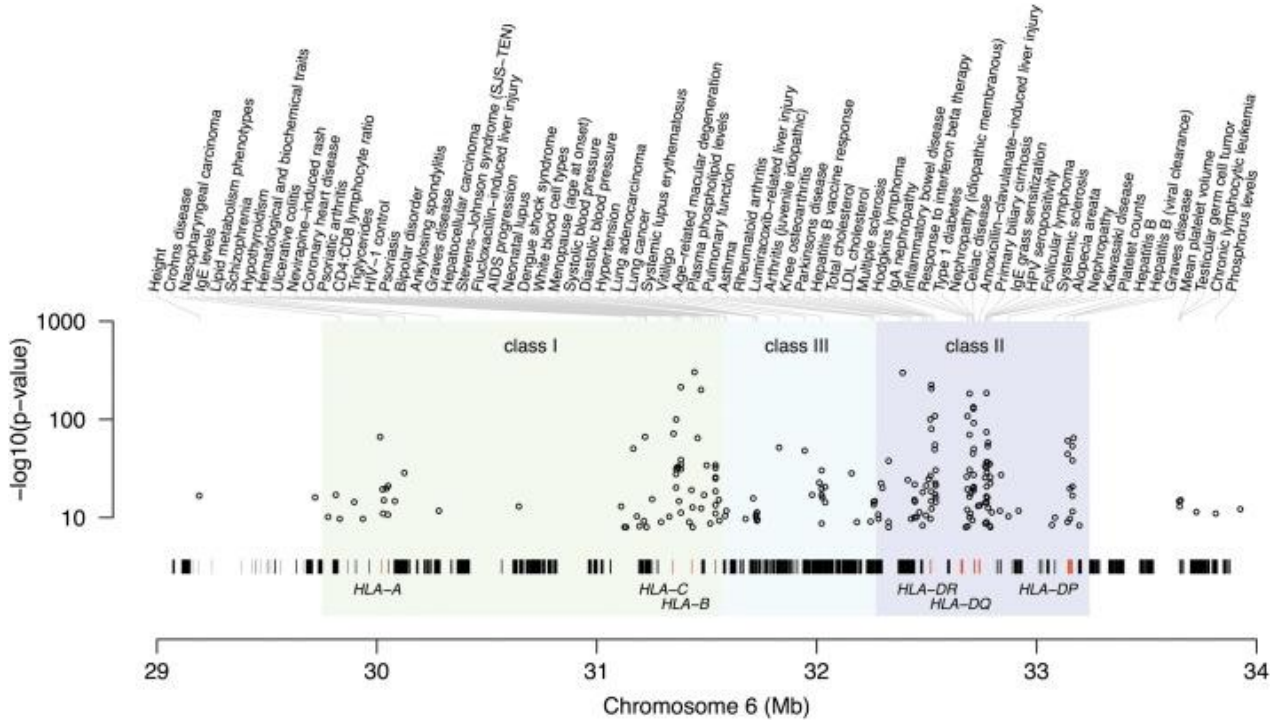


Figure 6. Associations significatives de nombreuses maladies avec des SNPs du CMH. Figure tirée d'un article publié en 2012 dans Human Molecular Genetics (de Bakker et Raychaudhuri 2012). Chaque point indique pour un SNP donné, sa position en abscisse et le niveau de signification de son association en ordonnée. La maladie correspondante est indiquée dans la partie supérieure du graphique.

Le CMH joue un rôle majeur dans le système immunitaire. C'est aussi la première région génétique de morbidité humaine de par le nombre et la force des associations aux maladies, et ce aujourd'hui encore à l'ère des études d'association génétique pangénomiques. Le CMH a ainsi été associé aux maladies auto-immunes et inflammatoires, aux maladies infectieuses, à certains cancers, à des maladies psychiatriques, voir aux effets secondaires des médicaments (**Figure 6**) (de Bakker et Raychaudhuri 2012). L'association entre les gènes du CMH et les maladies auto-immunes est connue depuis plus de trente ans. Ces associations ne sont pas dues à des mutations

mais à des variations alléliques particulières présentes avec une fréquence accrue chez les patients comparée à la population générale. Cependant, il est très difficile de distinguer clairement la contribution individuelle de chacun des gènes de cette région. Le CMH est en effet marqué par un fort déséquilibre de liaison (DL), c'est-à-dire par une association non aléatoire des allèles des locus le long de la région au sein d'haplotypes ancestraux étendus, et ce sur plus de 3 mégabases. La tâche est d'autant plus compliquée que le CMH est la région du génome humain la plus riche en gènes avec plus de 200 gènes qui sont en majorité exprimés dans le système immunitaire. C'est aussi la région la plus polymorphe, le gène *HLA-B* comprend par exemple 4242 allèles (http://hla.alleles.org/nomenclature/stats.html et https://www.ebi.ac.uk/ipd/imgt/hla/).

2.3. CMH et DT1

En ce qui concerne le DT1, le CMH contribue à 50% de la composante génétique, impliquant essentiellement les gènes de classe II *HLA-DRB1* et *HLA-DQB1* : 90 à 95% des patients caucasiens sont porteurs des haplotypes DR3 (HLA-DRB1*03-DQA1*0501-DQB1*0201) et/ou DR4-DQ8 (DRB1*04-DQA1*0301-DQB1*0302) au lieu d'environ 30% de la population générale. Il existerait un effet synergique de ces deux haplotypes présents chez 40% des patients à l'état hétérozygote contre 3% dans la population conduisant à un odds ratio de l'ordre de 30 (Hu et al. 2015). D'autres allèles comme DR15 (DRB1*15-DQB1*601) et HLA-DPB1*0402 semblent au contraire protecteurs vis à-vis du DT1.

Toutefois, les gènes de classe II n'expliquent pas la totalité de l'association au DT1 dépendante du CMH. En effet, il a été montré dans une cohorte prospective sur 20 ans que des germains de patients diabétiques porteurs des mêmes allèles DR3 et/ou DR4 avaient un risque de 55% de développer un DT1 lorsqu'ils partageaient avec le patient les deux mêmes haplotypes parentaux pour la totalité du CMH contre seulement 5% s'ils n'avaient pas hérité des deux mêmes haplotypes parentaux (Aly et al. 2006). Ainsi, d'autres facteurs de prédisposition génétiques que DR3 ou DR4 sont impliqués dans la pathogenèse du DT1 dépendante du CMH. Depuis, HLA-B*39 et HLA-A*24 ont été associés de manière indépendante des gènes de classe II (Nejentsev et al. 2007). Par une approche haplotypique, le gène *UBD* dans la région télomérique de classe I a aussi été impliqué (Baschal et al. 2011 ; Aly et al. 2008). La question de l'identification de gènes non HLA de classe II au sein du CMH demeure ouverte.

Tous les gènes impliqués dans le DT1 ne sont cependant pas identifiés. On parle d'héritabilité manquante. L'héritabilité représente la proportion de la variation phénotypique totale déterminée par les effets génétiques. Cette héritabilité manquante existe dans toutes les maladies multifactorielles, dont le DT1.

L'identification des variants prédisposant au DT1 par les études génétiques rencontre principalement un problème de puissance qui est limitée notamment par :

- la taille de l'échantillon
- la faible fréquence des variants causaux dans la population générale
- l'effet modeste de ces variants causaux
- l'hétérogénéité génétique

Outre ces problèmes de puissance insuffisante dans les études génétiques pour détecter de nouveaux variants de fréquence faible avec des effets modestes, l'héritabilité manquante pourrait être aussi expliquée par l'épistasie (interactions gènes-gènes), les interactions gènes-environnement ou encore l'épigénétique, qui correspond à une modification de l'expression des gènes (par méthylation, acétylation…) sans modification de la structure nucléotidique.

2.4. Facteurs environnementaux

L'intervention de facteurs extérieurs est en effet nécessaire pour déclencher la réaction auto-immune responsable du DT1. Ces facteurs sont cependant encore mal connus, d'autant plus qu'il est souvent difficile de mettre en cause l'environnement partagé familial. Un grand nombre de facteurs a été évoqué mais aucun n'est absolument prouvé et leur liste n'est pas exhaustive :

• Les agents infectieux peuvent prédisposer au DT1. Ainsi, l'activation polyclonale des lymphocytes en réponse à des infections chroniques ou les mécanismes inflammatoires associés aux oreillons ont été invoqués. Certaines infections entérovirales, en particulier au virus coxsackie, peuvent également prédisposer au DT1 de par leur tropisme pour les cellules bêta de Langerhans, et surtout car ils peuvent induire en erreur le système immunitaire du fait d'une ressemblance, appelée mimétisme moléculaire, entre les protéines virales et des antigènes des cellules β (ex : protéine virale 2C et GAD65 ou protéine virale VP1 et IA2). A l'opposé, les agents infectieux pourraient prévenir le DT1, selon la théorie d'un excès d'hygiène dans les pays industrialisés.

• Les facteurs diététiques peuvent également contribuer à l'apparition du DT1. L'introduction précoce du lait de vache dans l'alimentation du nouveau-né créerait une inflammation des cellules bêta du fait d'épitopes croisés entre l'insuline bovine et humaine. La consommation d'aliments riches en nitrosamines (composés toxiques pour les cellules β) ou en nitrites, ou de nitrates contenus dans l'eau de boisson a aussi été proposée.

• Des médicaments peuvent aussi avoir un effet toxique tel que la streptozotocine prescrite dans les insulinomes (cancer des cellules β).

2.5. Facteurs épigénétiques

Dans le DT1, une observation intrigante concerne le risque de récurrence de la maladie chez les enfants dont le père est diabétique, qui est 2 à 4 fois supérieur à celui des enfants dont la mère est diabétique (Warram et al. 1984 ; Rjasanowski et al. 1990). Les études génétiques cherchant un biais d'origine parentale n'ont pas mis en évidence d'association des gènes HLA classiques (Bronson et al. 2009), mais ont identifié un biais de transmission paternel au locus *DLK1-MEG3* soumis à empreinte (Wallace et al. 2010).

La première étude d'association épigénétique pangénomique (EWAS) dans les monocytes de jumeaux monozygotes discordants pour la maladie a identifié 132 régions de méthylation variable (Rakyan et al. 2011). Parmi les gènes sortants on retrouve *HLA-DQB1* ou encore *GAD2* qui code la protéine GAD65, un auto-antigène du DT1.

Une autre étude s'est intéressée au profil de méthylation dans les lymphocytes de jumeaux monozygotes discordants ou non pour la maladie (Stefan et al. 2014). Les résultats montrent une hyperméthylation globale chez les jumeaux concordants, qui est généralement associée à une répression des gènes. Cette hyperméthylation affecte des gènes impliqués dans les voies de défense et de réponses immunitaires. Des différences significatives de méthylation entre les jumeaux affectés et non affectés ont également été identifiées au niveau des gènes de susceptibilité au DT1 tel que *HLA* ou *INS*. Par ailleurs comme cela a été mentionné ci-dessus, l'incidence croissante du DT1 chez les jeunes enfants suggère un rôle de facteurs environnementaux entrainant des modifications épigénétiques. De fait, certains de ces facteurs comme le régime alimentaire, sont connus pour contribuer au DT1 et aux complications vasculaires en activant l'inflammation. Cette « mémoire métabolique » entraine des modifications épigénétiques à long terme. Dans les monocytes, l'activation de NFkB par le glucose augmente le

recrutement d'histones acétyltransférases au promoteur des gènes pro-inflammatoires comme le *TNF*, entrainant l'hyper-acétylation des histones H3 et H4 et l'activation transcriptionnelle de ces gènes (Patel et al. 2011) (Miao et al. 2004). Enfin, dans le modèle spontané auto-immun murin de diabète (souris NOD), le traitement par la trichostatine A, un inhibiteur d'histones déacétylases, réduit l'incidence de la maladie.

3. Approches actuelles dans les maladies multifactorielles

3.1. Limites des approches de génétique classique

Jusqu'à maintenant deux méthodes génétiques étaient utilisées afin de trouver de nouveaux gènes candidats (**Figure 7**). La première méthode est l'étude de liaison génétique. C'est une analyse génétique dans laquelle les gènes sont cartographiés par génotypage de marqueurs co-ségrégeant avec la maladie dans des familles. Elle permet de localiser les régions contenant les gènes responsables de la maladie. Cette méthode s'applique à des variants rares à pénétrance élevée, ce qui est plutôt adapté aux maladies monogéniques. La pénétrance représente la proportion d'individus possédant un génotype donné qui exprime le phénotype correspondant. Les analyses de liaison de type paramétrique, selon cette prévalence et le mode de transmission de la maladie, sont utilisées pour l'étude des maladies monogéniques. Dans les maladies polygéniques, des études de liaison sont aussi possibles, de type non paramétrique, basées sur les allèles identiques par descendance dans des paires de germains atteints. Par ces approches, les régions du CMH et de l'insuline avaient pu être liées au DT1 : ce sont les locus IDDM (Insulin-Dependant Diabetes Mellitus) 1 et 2. Une vingtaine d'autre locus IDDM ont été suggérés mais peu ont été confirmés (IDDM15, IDDM12, IDDM7, IDDM10 et la région 16p12-q24).
La deuxième méthode est l'association génétique, majoritairement de type études « cas-témoins ». Ces études ont pour objectif d'identifier des facteurs de susceptibilité génétique en comparant la fréquence des allèles des variants génétiques entre un groupe de cas, atteints de la maladie et un groupe de témoins, en utilisant le plus souvent des technologies de génotypage à haut débit. Elles s'appliquent plutôt aux variants avec des allèles fréquents à pénétrance faible, donc aux maladies multifactorielles. Par ces méthodes, plus de 50 régions ont été associées au DT1 comme décrit précédemment (**Tableau 4**).

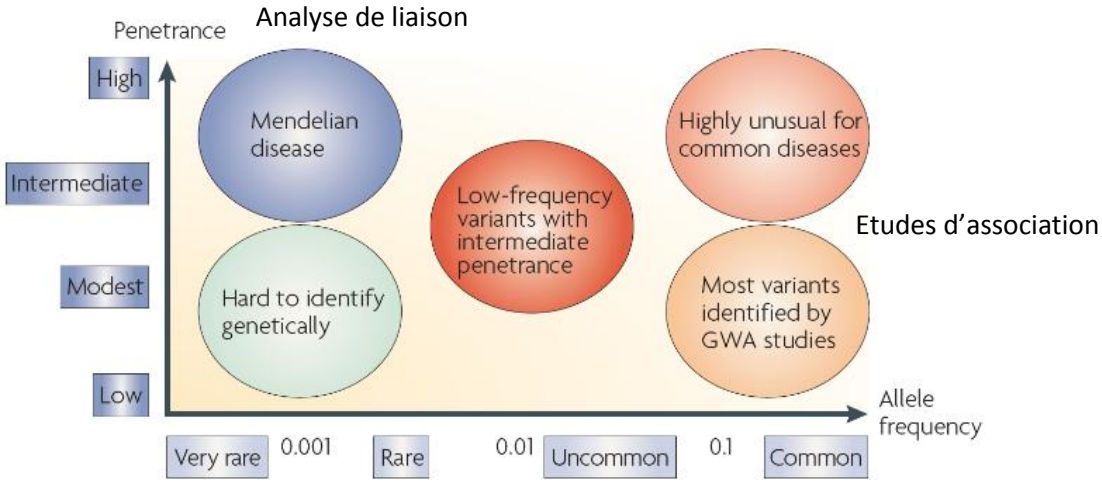


Figure 7. Importance pour les méthodes d'analyse génétique de la relation entre la fréquence dans la population générale des allèles morbides et l'effet de ces allèles sur le risque de développer les maladies. Figure adaptée d'un article de McCarthy publié en 2008 dans Nature Reviews Genetics (McCarthy et al. 2008).

Cependant deux écueils principaux demeurent :

1. Ces méthodes ne permettent pas de comprendre comment les gènes opèrent. La plupart des SNPs identifiés sont exceptionnellement les variants causaux mais sont en fait des marqueurs en déséquilibre de liaison avec le variant fonctionnel responsable de la maladie qu'il faut identifier.
2. Ces deux méthodes n'ont pas permis d'identifier tous les gènes responsables de la maladie : c'est l'héritabilité manquante.

Afin de résoudre ces deux problèmes, des approches différentes doivent être envisagées.

3.2. Vers la cartographie fine des variants causaux

Il est essentiel de réaliser une cartographie plus fine des régions candidates associées et répliquées dans plusieurs cohortes (même allèle de risque). La question est ici de déterminer, parmi tous les variants situés dans le locus associé, lequel est le mieux associé.

D'une part, il faut donc faire l'inventaire de tous les SNPs de la région associée : par re-séquençage des régions associées, par imputation (déduction) du génotype à d'autres SNPs que ceux génotypés grâce au déséquilibre de liaison entre ces SNPs connu dans une cohorte de référence de même origine ethnique, et par génotypage direct au moyen de puces à façon à plus haute densité réalisées par des consortiums. Dans cette perspective, un consortium travaillant sur plusieurs maladies auto-immunes et inflammatoires a notamment développé une puce à façon avec la technologie BeadChip d'Illumina, l'**immunochip**. Cette puce permet de génotyper environ 200 000 SNPs et insertions/délétions dans des régions génomiques associées à différentes maladies auto-immunes et inflammatoires et ciblant particulièrement la région du CMH (Trynka et al. 2011 ; Parkes et al. 2013). Appliquée au DT1, cette puce a permis de préciser quels sont les meilleurs SNPs associés dans les populations caucasiennes et d'identifier également des associations secondaires indépendantes dans ces régions (**Tableau 4**) (Onengut-Gumuscu et al. 2015) (www.t1dbase.org et immunobase.org).

D'autre part, une fois tous les variants testés par association génétique, il est fréquent que plusieurs variants donnent le même niveau d'association génétique et soient indistinguables en raison de leur déséquilibre de liaison (DL) parfait dans l'échantillon testé. En collectant un échantillon de plus grande taille, l'estimation du DL est plus précise et peut conduire à un DL moins parfait qu'il ne l'était avec un petit échantillon, permettant de distinguer le meilleur variant associé. Une alternative à une taille plus grande d'échantillon est de collecter des sujets issus d'une population différente, potentiellement avec des fréquences alléliques différentes, conduisant à une structure du DL différente. Parmi la liste finale de variants en parfait DL, il sera nécessaire de distinguer parmi eux le variant causal en étudiant l'impact fonctionnel de chacun.

Depuis quelques années, une nouvelle approche génomique est mise en avant pour aider à identifier ces variants causaux. Cette approche combine les données des études d'associations génétiques aux maladies, à des études d'association génétique de l'expression des gènes. En effet, les maladies génétiques complexes peuvent être considérées comme la manifestation extrême de la variation génétique, alors que la variation phénotypique au niveau transcriptomique représenterait une étape intermédiaire. Étant démontré que les niveaux d'expression des gènes sont héréditaires, il est possible de traiter le niveau d'expression génique comme un trait quantitatif, aussi appelé phénotype d'expression, qui est contrôlé par des « expression quantitative trait loci » (eQTLs) agissant en *cis* ou en *trans* (**Figure 8**) (Westra et al. 2013).

Ces eQTLs peuvent être cartographiés (« eQTL mapping ») par des études d'association génétique en combinant, pour les mêmes échantillons, les données de génotypage pangénomique et celles de leur transcriptome. La variation d'expression de chaque gène en fonction du génotype de chaque variant est testée par régression linéaire. La cartographie des eQTLs fournit un puissant outil pour identifier des variants régulateurs qui contrôlent les phénotypes d'expression (Dixon et al. 2007) : ce sont les eSNPs (expression regulatory SNPs) qui sont potentiellement des variants causaux dans les maladies. L'objectif est ainsi de réduire la liste des SNPs associés aux maladies à hérédité complexe issus des GWAS en les confrontant aux eSNPs : l'intersection de ces deux listes de SNPs à un même locus indique les variants candidats à étudier en priorité comme variants causaux des maladies.

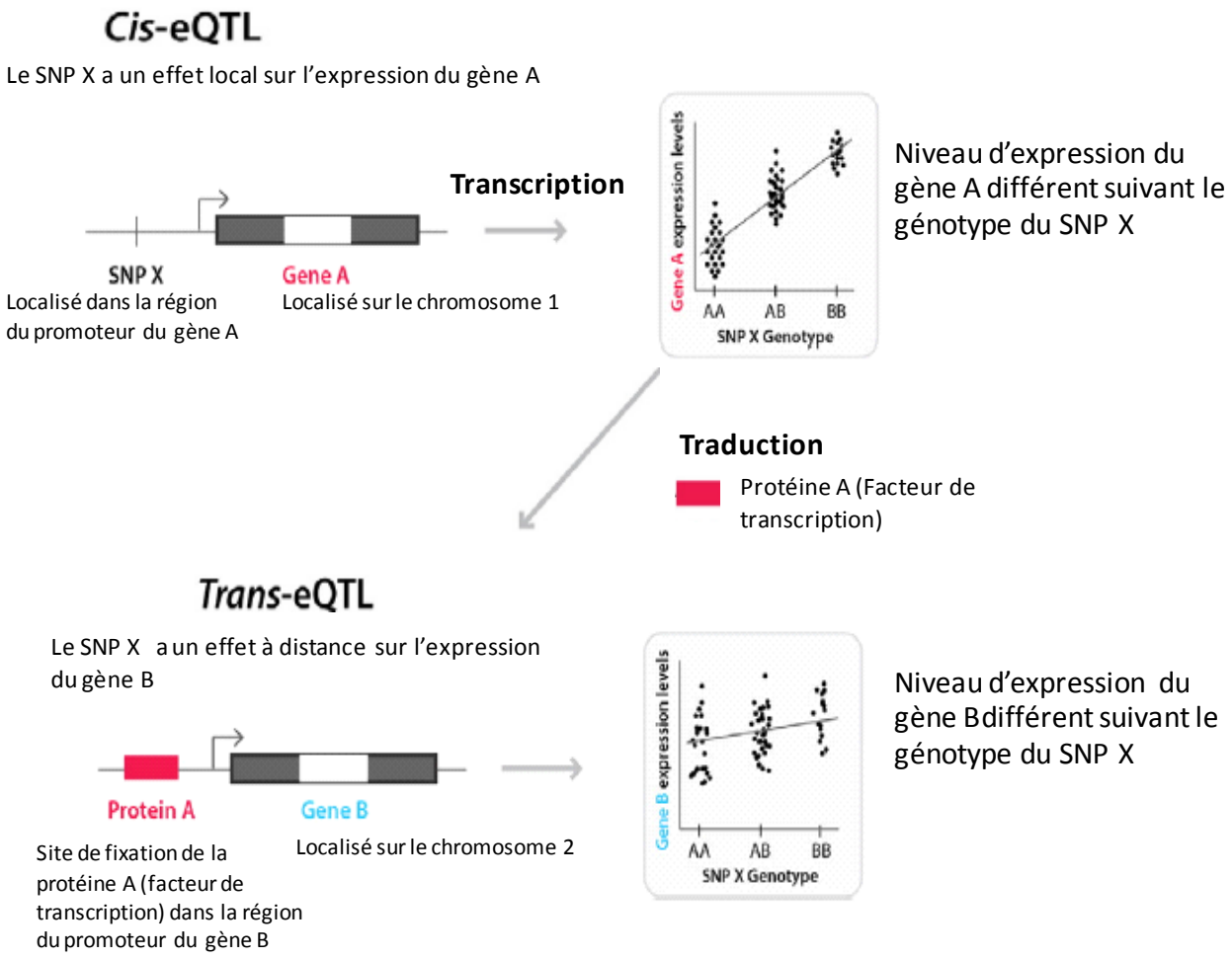


Figure 8. Effet des eQTLs en fonction de la distance du variant génétique. Lorsque le variant génétique à un effet sur l'expression d'un gène situé entre 1 à 5 mégabases, on parle de cis-eQTL. Si ce variant est situé à plus de 5 mégabases ou sur un autre chromosome, on parle de trans-eQTL. Figure adaptée de (Westra et Franke 2014). Un variant ayant un effet en cis sur l'expression d'un gène A peut avoir un effet en trans sur un gène B si le produit du gène A par exemple est un facteur de transcription du gène B.

3.3. A la recherche de l'héritabilité manquante

Les GWAS n'ont pas permis d'identifier toute la composante génétique des maladies à hérédité complexe. L'une des raisons possible est que les variants testés dans les GWAS sont des SNPs fréquents étudiés individuellement. Le choix de ces variants génétiques et les méthodes d'étude ont conditionné les résultats obtenus. D'une part, d'autres types de variants génétiques peuvent être impliqués dans les maladies à hérédité complexe. Les études d'association génétiques conduites sur des Copy Number Variants ont ainsi succédé aux GWAS sur les SNPs, à l'aide de puces à acide désoxyribonucléique (ADN) de type CGH (Comparative Genetic Hybridization) (Wellcome Trust Case Control Consortium et al. 2010). Dans le DT1, ces études ont confirmé l'association bien connue de la classe I du polymorphisme de type Variable Number of Tandem Reapeat (VNTR) localisé en 5' du gène de l'insuline, associé à une diminution de la transcription du gène *INS* dans le thymus. Cependant, aucun nouveau locus n'a pu être significativement impliqué dans l'héritabilité du DT1. Les variants rares, avec une fréquence de l'allèle mineur (MAF) inférieure à 1% chez les sujets contrôles, n'étaient pas initialement inclus sur les puces de génotypage haut-débit. Avec l'avènement des projets de séquençage haut-débit de génomes de plusieurs milliers de sujets contrôles (ex : projets 1000 genomes et UK10K) (Sudmant et al. 2015 ; UK10K Consortium et al. 2015), les puces commerciales et à façon peuvent désormais les inclure.

Le séquençage des exomes (Whole Exome Sequencing, WES) des échantillons de cas a aussi permis de tester les variants rares dans les régions codantes des gènes. Dans le DT1, quelques variants rares ont ainsi pu être associés au locus *IFIH1* indépendamment de l'effet du variant fréquent précédemment associé (Nejentsev et al. 2009). La contribution de ces variants rares dans les maladies à hérédité complexe, et en particulier auto-immunes, se révèle cependant à ce jour négligeable (Cirulli et Goldstein 2010 ; Hunt et al. 2013). Cependant, ces études manquent encore de puissance pour détecter des effets faibles. La baisse du coût du séquençage de l'ensemble du génome (Whole Genome Sequencing, WGS) et le développement de nouvelles méthodes statistiques, en particulier agrégeant les variants rares d'un même gène, pourraient réévaluer la part de ces variants rares dans l'héritabilité. D'autre part, les études d'associations pangénomiques ont le plus souvent été conduites en considérant les SNPs individuellement avec un effet additif des allèles de chaque variant sur le seul phénotype binaire du statut vis-à-vis de la maladie.

Il n'est pas exclu que des variants contribuent aux pathologies par des interactions épistatiques. Cependant, la multiplicité des tests statistiques ne permet pas de les tester systématiquement pour identifier de nouveaux variants qui n'auraient pas d'effet propre. La détection d'interactions entre les gènes et l'environnement sont également difficiles à mettre en évidence, nécessitant des cohortes prospectives.

L'étude de phénotypes intermédiaires, ou endophénotypes, constitue un domaine de recherche actif. En effet, l'amplitude des effets des variants peut être plus importante sur ces endophénotypes que sur le phénotype final, facilitant leur détection. A cet égard, les eQTLs, déjà décrits au paragraphe 3.2, sont très pertinents car au plus près de la variation génétique et pourraient constituer de très bons nouveaux candidats pour la pathologie. L'étude systémique intégrant d'autres phénotypes « omiques » est l'une des voies actuelles les plus prometteuses (**Figure 9**). Toutefois, il est à noter que, selon certaines hypothèses, la quasi-totalité des locus impliqués dans la susceptibilité aux maladies auraient en réalité déjà été détectés. Par exemple, il a été suggéré que les variants fréquents détectés dans les GWAS refléteraient des associations de

multiples variants rares aux effets plus forts que le variant fréquent : on parle de signal d'« association synthétique ». La difficulté d'estimer correctement l'héritabilité est aussi en cause et aurait conduit à sa surestimation.

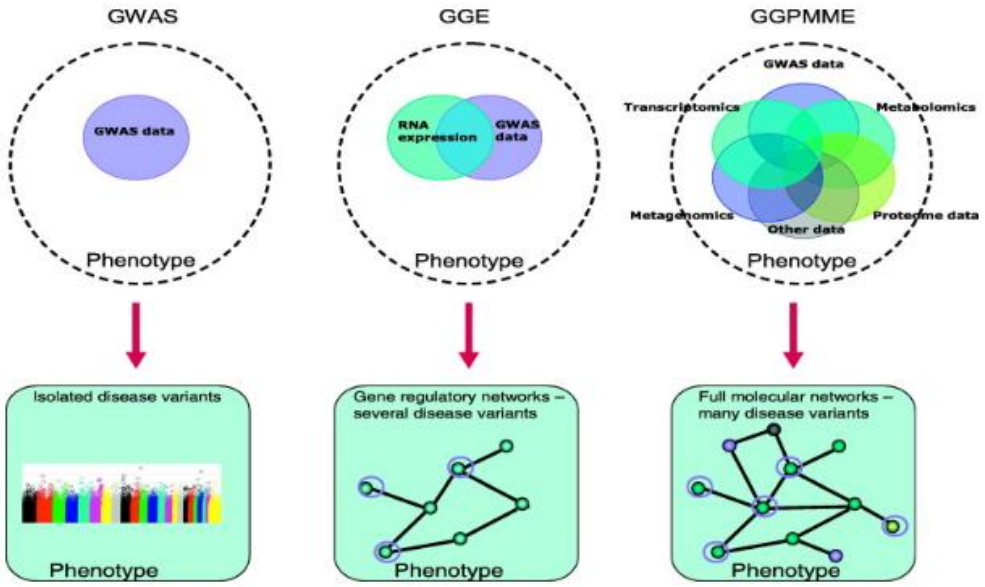


Figure 9. Comparaison de trois approches permettant d'identifier les SNPs causaux. Figure tirée de (Groop et Pociot 2014). GWAS : Genome Wide Association Study ; GGE : Genetics of Gene Expression ; GGPMME : Genetics of gene, Protein, Metabolite, and Metagenomics Expression. Les études de type GWAS testent l'association de variants génétiques à un phénotype donné. Les résultats sont représentés sous forme d'un graphique de type « manhattan plot », représentant en abscisse les variants testés le long du génome avec une alternance de couleur par chromosome, et le niveau d'association en ordonnée. Plusieurs variants sont généralement associés. La méthode « GGE » teste en plus l'association des variants pour l'expression des gènes. Les variants à la fois associés au phénotype testé dans le GWAS et à l'expression des gènes sont de bons candidats fonctionnels et permettent de définir un réseau de régulation transcriptionnelle. La méthode « GGGPME » combine aussi l'association génétique à d'autres « omes » comme le métagénome et le protéome. Leur étude intégrée apporte une information moléculaire et fonctionnelle aux variants associés, permettant de distinguer les meilleurs candidats.

4. *ZFP57* : un gène du CMH, candidat au DT1

Plusieurs raisons convergentes ont conduit notre équipe à s'intéresser au gène *Zinc Finger Protein 57* (*ZFP57*), localisé dans le CMH (en 6p21), comme candidat à la susceptibilité au DT1.

4.1. *ZFP57* : une cause génétique de diabète néonatal

Tout d'abord en 2008, l'équipe de Deborah Mackay a démontré que certaines formes de diabète néonatal transitoire (TNDM) avec une mosaïque d'hypométhylation de l'ADN dans les régions différentiellement méthylées (DMR) sont associées à des mutations récessives de *ZFP57*. Cette forme rare de diabète prédispose les patients au DT1 plus tard dans leur vie (Mackay et al. 2008).

4.2. *ZFP57* : un candidat dans le DT1

Puis en 2011, Claire Vandiedonck a généré la première carte de transcription du CMH (Vandiedonck et al. 2011). Cette carte d'expression des gènes de la région a été réalisée en tenant compte du fond haplotypique, puisque l'étude a été réalisée à l'aide de trois lignées cellulaires lymphoblastoïdes (LCL) entièrement homozygotes pour cette région. Chaque LCL porte un haplotype à risque pour des maladies auto-immunes :

- lignée COX, haplotype HLA-A1-B8-Cw7-DR3 associé au DT1, au lupus érythémateux systémique et à la myasthénie acquise auto-immune.
- lignée PGF, haplotype HLA-A3-B7-Cw7-DR15 de susceptibilité à la sclérose en plaques, au lupus érythémateux systémique ; protecteur pour le DT1.
- lignée QBL, haplotype HLA-A26-B18-Cw5-DR3-DQ2 associé au DT1 et à la maladie de Basedow.

Or, la comparaison de leur transcriptome au moyen d'une puce à façon a permis d'identifier le gène *ZFP57* comme étant le plus significativement différentiellement exprimé au sein du CMH, avec une valeur de P ajustée de $1{,}22.10^{-14}$. Un eQTL de l'expression de ce gène a été identifié au sein de cellules primaires, à savoir des cellules mononuclées du sang périphérique (PBMC), provenant d'une cohorte de 92 volontaires génotypés avec une puce Illumina (HumanCVDv1 BeadChip). La quantification en parallèle du niveau d'expression de *ZFP57* par réaction d'amplification en chaine quantitative (qPCR) a permis de montrer l'association entre le niveau d'expression de *ZFP57* et le SNP rs29228, situé à 16,8 kb de *ZFP57* ($p=1{,}2.10^{-14}$). Il est important de noter que l'allèle associé à un niveau d'expression élevé de *ZFP57* est porté par l'haplotype diabétogène DR3 (**Figure 10**). Ensuite, une cartographie plus fine a été réalisée par la même équipe dans les PBMCs dans un groupe de 288 volontaires génotypés pour 646 631 marqueurs génétiques avec une puce Illumina (Human OmniEpress-12v1.0 BeadChip). Cette analyse révèle que l'association de SNPs influençant fortement le niveau des transcrits de *ZFP57* ($p=4{,}6.10^{-52}$) sont situés dans le premier intron de ce gène (Plant et al. 2014). Ainsi, ce travail montre que *ZFP57*, muté dans des formes de diabète néonatal transitoire, est aussi le gène le plus fortement différentiellement exprimé chez les porteurs de l'haplotype DR3 du CMH, haplotype associé au DT1. Il s'agit donc d'un excellent candidat pour expliquer la prédisposition au DT1 liée au CMH mais indépendante des gènes *HLA* de classe II.

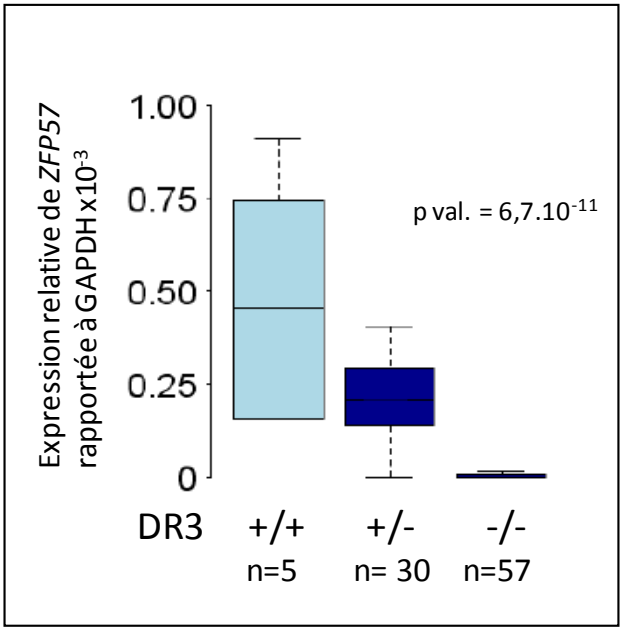


Figure 10. Expression relative du transcrit de *ZFP57* dans des PBMCs de 92 volontaires en fonction de leur génotype pour DR3 (Vandiedonck et al. 2011)

4.3. Un répresseur de la transcription impliqué dans les mécanismes épigénétiques

ZFP57 est connu comme étant un gène du développement précoce exprimé dans les cellules souches embryonnaires. Il a été caractérisé chez la souris par le Dr. Shinji Okazaki en 1994 dans des lignées cellulaires non différenciées. Après différenciation, son expression transcriptomique diminue fortement (Okazaki et al. 1994). Chez l'homme, ce gène est situé dans la région télomérique de classe I du CMH et comporte 6 exons s'étendant sur 8,8 kilobases (kb). Il code une protéine à doigts de zinc. Trois isoformes protéiques sont connues. La forme dominante, composée de 452 acides aminés, est caractérisée par un domaine de type KRAB (Krüppel associated box) dans la partie N-terminale, impliquée dans les interactions protéine-protéine, et 7 motifs en doigt de zinc permettent la liaison à l'ADN (**Figure 11**).

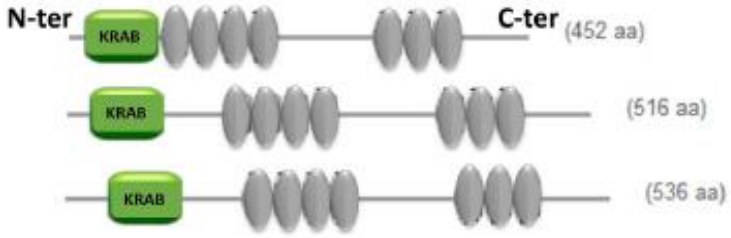


Figure 11. Structure des trois isoformes protéiques de ZFP57. Les trois isoformes contiennent un domaine KRAB (en vert) liant le co-facteur KAP1 et 7 doigts de zinc (en gris).

Chez la souris, il a été montré que ZFP57 est un facteur de répression de la transcription qui se fixe à l'ADN. Il est nécessaire au maintien de la méthylation de certains gènes, dont ceux soumis à empreinte parentale, dans les zygotes lors de la phase de déméthylation complète de l'ADN dans les premières étapes du développement embryonnaire (**Figure 12**). Il interagit *via* son domaine KRAB avec son co-facteur KAP1 (codé par *TRIM28*) qui va interagir avec une ADN méthyltransférase (DNMT) pour contrôler la méthylation allèle-spécifique (**Figure 13**).
Chez la souris, il a été montré que le complexe ZFP57-KAP1 se fixe sur les motifs hexanucléotides TGCCmetGC au niveau des régions de contrôle de l'empreinte (ICR) méthylées ainsi que sur d'autres régions soumises à méthylation différentielle sur l'ensemble du génome. Le complexe recrute différentes protéines maintenant la méthylation des dinucléotides CpG : HP1, SETDB1, les DNA méthyltransférases (DNMT) 1, 3A et 1B. La méthylation de l'ADN, en général, a pour effet de réprimer la transcription des gènes. Chez la souris, l'inactivation de *ZFP57* dans les ovocytes et les embryons est létale avant le sevrage, tandis que la perte d'expression uniquement dans les embryons est viable et n'affecte pas le développement de l'embryon (Li et al. 2008). On parle d'effet "maternel-zygotique". L'expression de *ZFP57* chez la mère et l'embryon sont donc essentiels durant le développement embryonnaire précoce. Comme chez les patients avec un diabète néonatal transitoire, la perte de fonction de ZFP57 chez la souris induit une perte de méthylation de l'ADN dans plusieurs régions soumises à empreintes parentales. Toutes ces données montrent l'importance de l'effet épigénétique de ce gène durant le développement embryonnaire.

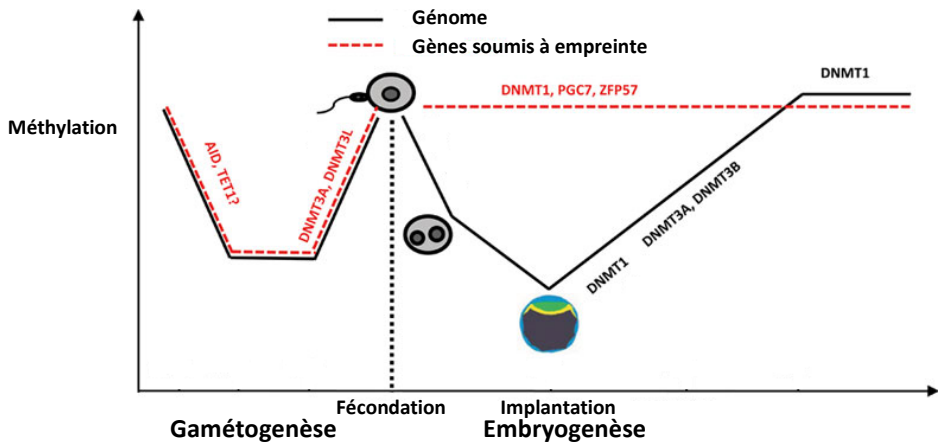


Figure 12. Méthylation et déméthylation programmées du génome. Pendant la gamétogenèse, la méthylation des gènes non soumis à empreinte suit le même profil de méthylation/déméthylation que les gènes soumis à empreinte. Au cours de l'embryogenèse, ils subissent un effacement de la méthylation puis la mise en place d'une nouvelle méthylation. En revanche, pour les gènes soumis à empreinte la méthylation est maintenue dans l'embryon pré-implantatoire.

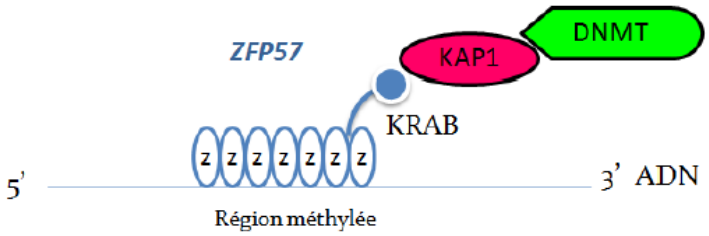


Figure 13. Interaction entre ZFP57 (en bleu), KAP1 et une ADN Méthyl Transférase au niveau d'une région de l'ADN méthylée.

4.4. Rôle potentiel de ZFP57 dans des pathologies de l'adulte

Récemment deux articles ont montré le rôle potentiel de ZFP57 dans le cancer, avec notamment sa surexpression dans les cancers pancréatiques (Tada et al. 2015). Une étude récente indique également que les régions susceptibles d'être liées par ZFP57 chez l'homme sont enrichies en région de méthylation différentielle dans des maladies inflammatoires et métaboliques (Tang et al. 2015). Jusqu'en 2011, l'expression de *ZFP57* n'était connue que dans l'embryon ainsi que dans les testicules, les ovaires et le cerveau chez l'adulte. Mais les travaux de l'équipe montrent que chez l'adulte, *ZFP57* est aussi exprimé au niveau transcriptionnel dans les PBMCs, donc dans le système immunitaire (Vandiedonck et al. 2011), ce qui renforce la candidature de ce gène pour un rôle génétique et épigénétique dans la prédisposition au DT1.

Projet de recherche

Les méthodes d'analyse de liaison et études d'association ont permis de découvrir plus de 50 régions génétiques impliquées dans la susceptibilité au DT1. Ces études ont essentiellement été réalisées avec la cohorte internationale du T1DGC, laquelle ne comporte pas de patients français. De plus, ces études n'expliquent pas toute l'héritabilité du diabète de type 1.
Dans ce cadre, le premier objectif du laboratoire a été de conduire une étude d'association génétique pangénomique afin d'évaluer l'impact dans la population française des variants connus, voire de découvrir de nouveaux variants. Le second objectif est de développer une approche fonctionnelle basée sur l'expression génique, soit pangénomique, soit ciblée, afin de rechercher de nouveaux gènes impliqués dans la pathogenèse du DT1 et donc de réduire la part manquante de l'hérédité.
Sur cette base, mes objectifs spécifiques étaient de :

- répliquer dans une cohorte étendue de « réplication » des résultats d'association génétique qui avaient été obtenus dans le laboratoire à l'aide d'une puce de génotypage ciblant les gènes immunitaires, l'immunochip, dans une cohorte familiale française de « découverte »
- valider des gènes différentiellement exprimés entre patients et leurs contrôles familiaux et identifiés au laboratoire à l'aide d'une puce d'expression pangénomique dans la cohorte internationale du T1DGC
- caractériser fonctionnellement l'expression du gène candidat *ZFP57*

De façon plus détaillée, mes objectifs étaient donc les suivants :

1. Etude génétique : « French GWAS »

Pour la première étude, le laboratoire a recruté une cohorte de patients de DT1 et de leurs apparentés au 1er degré en collaboration avec le Pr. Marc Nicolino qui travaille à l'hôpital Femme Mère Enfant – Hospices Civils de **Lyon**. La cohorte inclut 3464 sujets.
Une analyse d'association familiale avait été réalisée au laboratoire sur un sous-ensemble de cette cohorte, incluant 685 sujets génotypes avec l'immunochip.
Mon but était d'étendre les résultats afin de confirmer les associations significatives et suggestives obtenues à différents locus sur le reste de la cohorte.

2. Etude transcriptomique : « T1DGC-Express »

Cette étude s'inscrit dans un projet collaboratif international financé par le National Institutes of Health (NIH) et repose sur les cohortes internationales du **T1DGC**. Avant mon arrivée au laboratoire, 264 puces d'expression pangénomiques avaient été hybridées à l'acide ribonucléique (ARN) de lignées lymphoblastoïdes de 45 sujets (patients et germains non atteints), cultivées dans trois conditions de stimulation mitogénique avec du PMA (Phorbol Myristate Acetate), en duplicat. Leur analyse avait identifié 199 gènes differentiellement exprimés entre les patients et leurs contrôles familiaux.

Dans un premier temps, ma mission a été de valider et de répliquer ces résultats sur les ARNs des 264 mêmes échantillons et sur 169 autres échantillons du T1DGC. Cinq gènes d'intérêt ont été choisis parmi les 199 : 4 appartenant à la voie de l'interféron-gamma (*IFNγ*, *IL-27*, *IRF1* et *STAT1*) et un cinquième gène, *UBD*, appartenant au CMH de classe I et ayant été déjà identifié dans une étude comme étant impliqué dans le DT1 (Baschal et al. 2011 ; Aly et al. 2008).
Dans un second temps le but était de comparer les résultats de transcriptomique à l'expression protéique par ELISA des protéines IFNγ et IL-27 sur les 311 surnageants de culture cellulaire à disposition.
Enfin, le projet s'étend aussi à des cellules mononucléées du sang périphérique, dont j'ai extrait l'ARN et dont le transcriptome est en cours par la technologie RNA-Seq.

3. Etude génétique et fonctionnelle d'un gène candidat, *ZFP57*

Le dernier objectif était d'étudier génétiquement et fonctionnellement le gène *ZFP57*.
L'hypothèse est que l'expression de ce gène doit être un élément épigénétique important dans la susceptibilité au DT1.
Le but est donc d'élucider les mécanismes à travers lesquels ce gène est impliqué au moyen de différentes ressources biologiques. Pour cela mes objectifs étaient de :

- cartographier les variants causaux de *ZFP57* grâce à la cohorte familiale de **Lyon**
- localiser la protéine dans la cellule au moyen de lignées lymphoblastoïdes **COX, PGF et QBL**
- mettre en évidence l'expression protéique de ZFP57 dans les cellules primaires du sang de sujets sains volontaires adultes de la plateforme **ICAREB**
- caractériser les cibles de ZFP57 par Immunoprécipitation puis séquençage (ChIP-seq)

Matériel et méthodes

1. Cohortes et lignées cellulaires

Dans l'ensemble de ce projet, trois cohortes de sujets et quatre lignées cellulaires lymphoblastoïdes (LCL) modèles ont été étudiées. Elles sont décrites ci-après avec les données disponibles au moment où j'ai rejoint le laboratoire. Quatre figures (**Figure 14, 16, 17 et 18**) récapitulent les échantillons et expériences réalisées, en indiquant explicitement quelles étapes j'ai prises en charge. Je les détaille dans les sections suivantes.

1.1. Cohorte française de Lyon

Cette cohorte a été établie aux Hospices Civils de Lyon par le Pr. Marc Nicolino et le Dr. Cécile Julier sous l'égide de la JDRF (Juvenile Diabetes Research Foundation) entre 2002 et 2009, avec pour objectif de recruter de jeunes patients et leurs apparentés. Une déclaration de collection a été faite par les Hospices Civils de Lyon pour les échantillons d'ADN (sous la référence DC-2008-72).

La cohorte est composée essentiellement de familles simplex, c'est-à-dire de familles avec un seul enfant atteint de diabète de type I. Seules 5% des familles présentent entre 2 et 4 patients. Les germains non atteints des patients ont aussi été recrutés dans la mesure du possible. Il s'agit de sujets tous caucasiens, originaires de France (2/3) ou du Maghreb (1/3). Au total, la cohorte comporte 1019 patients (53% d'hommes et 47% de femmes) ayant débuté leur diabète entre l'âge de 10 mois et 17 ans (âge moyen=7,38 ±4,12 ans). Elle est divisée en deux groupes : Lyon 1 (685 sujets, 237 familles, 245 patients) et Lyon 2 (2779 sujets, 867 familles, 774 patients). Pour quelques familles de la cohorte Lyon 2, l'ADN des patients n'a pu être collecté, expliquant une différence avec le nombre de familles et de patients.

Les ADNs et ARNs de 366 sujets sont disponibles pour la cohorte Lyon 1 (209 familles et 199 patients), tandis que l'ADN seul est disponible pour les sujets de la cohorte Lyon 2 (**Figure 14**). L'extraction de ce matériel nucléique, collecté à partir de sang total sur tubes PAXgene, a été réalisée par Anne Boland à la Biobanque du Centre National de Génotypage (CNG, Evry). Après contrôle qualité, 636 ADNs de la cohorte Lyon 1 et 2493 ADNs de la cohorte Lyon 2 ont été mis en plaque à une concentration de 5 ng/µL. Chaque plaque comportait un sujet en commun comme contrôle positif ainsi qu'un contrôle négatif (eau) disposés dans des puits différents sur chaque plaque afin de contrôler l'absence d'inversion de plaques grâce à la bonne correspondance entre les génotypes obtenus et les plans de plaque. Les 34 plaques de 96 puits correspondantes, ainsi que les solutions mères en microtubes, ont été conservées à -20°C au laboratoire.

La cohorte Lyon 1 a aussi été entièrement génotypée avec la puce « immunochip » (Illumina Infinium beadarray) après mise en plaques de 96 puits à 50 ng/µl (volume de 10 µl) incluant également chacune un sujet en commun comme contrôle positif et un contrôle négatif (eau).

La puce « immunochip » inclut 196 524 SNPs et insertions/délétions répartis sur l'ensemble du génome, dans des régions génomiques associées à différentes maladies auto-immunes et inflammatoires, et ciblant particulièrement la région du CMH (**Figure 15**). Le génotypage a été

effectué sur la plate-forme Illumina de la plate-forme PS3 de la Pitié-Salpêtrière (Paris) par Wassila Carpentier. Les génotypages ont été réalisés en aveugle des données cliniques. Les génotypes pour les allèles A et B d'Illumina ont été extraits avec le logiciel GenomeStudio d'Illumina au laboratoire U958. L'analyse d'association familiale a été réalisée avant mon arrivée par Windy Luscap, étudiante en M2 au laboratoire.

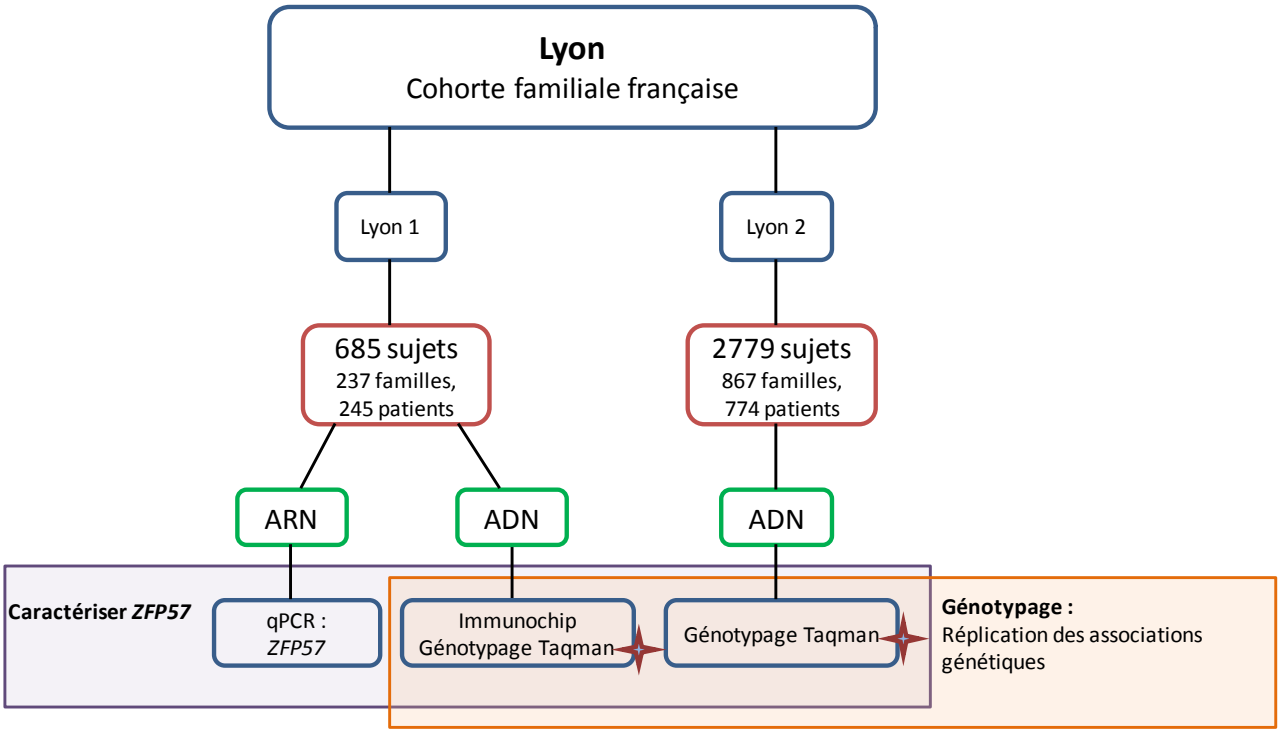


Figure 14. Bilan des échantillons disponibles et des techniques réalisées dans la cohorte de Lyon. Les étoiles indiquent les techniques que j'ai prises en charge au laboratoire.

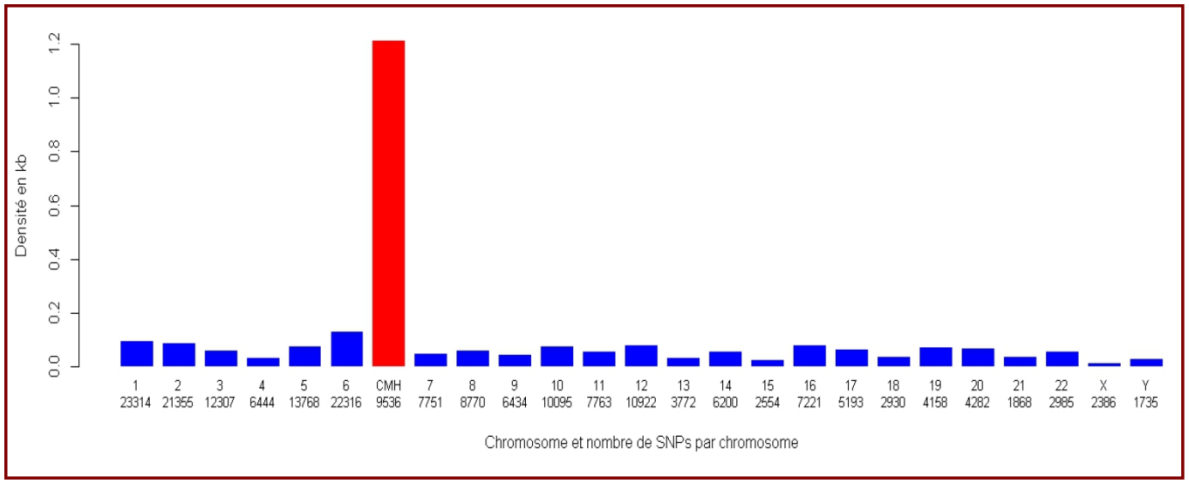


Figure 15. Densité de SNPs par kilobase pour chaque chromosome. Le nombre de SNPs par chromosome (en bleu) est indiqué sous le nom de chaque chromosome. La couverture pour le CMH (en rouge) est également indiquée après le chromosome 6.

1.2. Cohorte française CosImmGene (CIG)

Cette cohorte de volontaires sains a été recrutée par le Dr. Marie-Noëlle Ungeheuer à la plate-forme « Investigation Clinique et Accès aux Ressources Biologiques » (ICAReB) de l'Institut Pasteur. Elle présente l'avantage unique d'autoriser le rappel des volontaires pour de nouveaux prélèvements de sang, en fonction de leur génotype, phénotype et/ou pour des études longitudinales. Elle est constituée pour l'instant de 44 adultes sains, 31 femmes et 13 hommes, âgés de 21 à 67 ans (médiane=39,3 ans, quartile inférieur = 26,8 ans, quartile supérieur = 48,5 ans) sans différence significative d'âge entre les hommes et les femmes. Tous ces sujets, sauf trois, sont caucasiens. Leur ADN a été extrait à partir du sang total par la plate-forme ICAReB. Tous ces sujets ont été également génotypés sur l'immunochip à la plate-forme PS3 (Pitié-Salpêtrière, Paris, W. Carpentier). Pour 18 de ces volontaires, l'ARN a également été extrait à partir du sang total par Tharshana Stephen, étudiante de Master 2, avec le kit Paxgene (Qiagen). Quatre de ces sujets ont été prélevés à deux reprises à quelques mois d'écart et leurs ARNs extraits à chaque fois. Les leucocytes ont été également obtenus pour 10 sujets afin de conduire des expériences de cytométrie en flux que j'ai réalisées en binôme avec Tharshana (**Figure 16**).

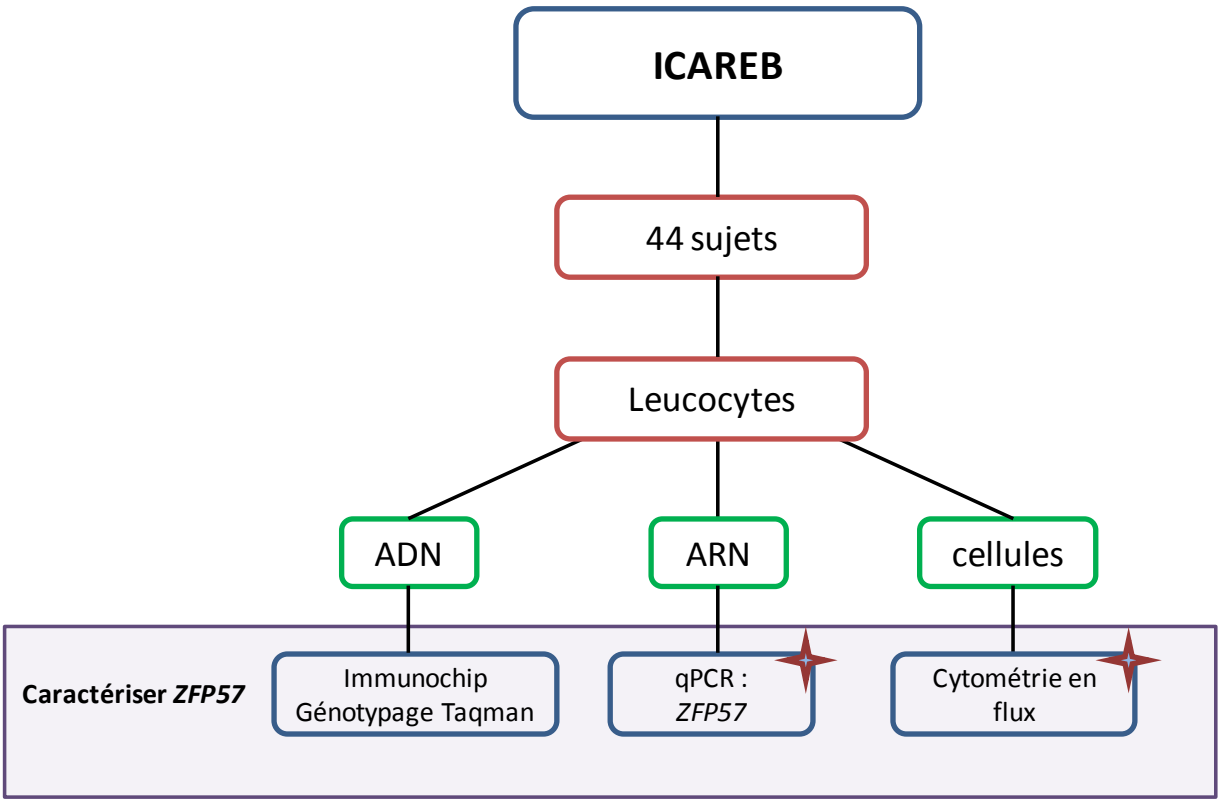


Figure 16. Bilan des échantillons disponibles et des techniques réalisées dans la cohorte ICAREB. Les étoiles indiquent les techniques que j'ai prises en charge au laboratoire.

1.3. Cohorte internationale du T1DGC

Le T1DGC (Type 1 Diabetes Genetics consortium) est un consortium international qui a été créé dans le but d'identifier les gènes et les allèles de risque dans le DT1 en recrutant plus de 10 000 sujets appartenant à 2500 familles multiplexes, c'est-à-dire avec au moins deux enfants atteints de DT1 (Hilner et al. 2010). Le recrutement a eu lieu en Asie-Pacifique, Europe (EUR), Amérique du Nord (NAm = North American et JOS = Joslin Diabetes Center), au Royaume-Uni (BDA = Bristish Diabetic Association, UK = United Kingdom), au Danemark et en Sardaigne.

Les deux principaux objectifs du T1DGC étaient : (1) d'identifier les régions génomiques et les gènes candidats dont les variants modifient le risque de développer un DT1, et (2) de créer une base de données et de ressources utilisables par la communauté scientifique (t1dbase.org). Dans ce but, ces cohortes ont été génotypées pour de très nombreux marqueurs pangénomiques, incluant récemment l'immunochip pré-citée (Onengut-Gumuscu et al. 2015). Dans la région du CMH, les sujets ont aussi été génotypés pour 66 marqueurs microsatellites.

Dans la présente étude, nous avons étudié l'expression génique chez des sujets de la cohorte sélectionnés selon des critères génétiques précis, afin d'optimiser nos chances d'identifier de nouveaux gènes de prédisposition au DT1. Ces sujets ont été choisis de façon à masquer l'effet des facteurs de risques majeurs HLA-DR3 et DR4-DQ8. La stratégie suivie a consisté à sélectionner dans une première étape les paires de germains discordants pour la maladie et de génotype HLA de classe II DR3/DR3, DR3/DR4 ou DR4/DR4. Puis, leurs haplotypes pour la région entière du CMH (pas seulement les gènes de classe II) ont été reconstruits. Les paires de germains présentant au moins un haplotype parental distinct au niveau du CMH ont été finalement retenues.

Cette stratégie vise à identifier les facteurs génétiques au sein du CMH distinct des gènes de classe II. Au total, 215 paires de germains satisfaisaient ces critères sur l'ensemble des cohortes du T1DGC. Seuls des échantillons cellulaires (LCL et/ou PBMC) des cohortes EUR et NAm étaient cependant disponibles pour étudier l'expression génique : 53 sujets pour la cohorte EUR incluant 24 diabétiques issus de 19 familles avec 25 paires de germains discordants, et 57 sujets pour la cohorte NAm (**Figure 17**) incluant 32 diabétiques issus de 24 familles avec 32 paires de germains discordants. Pour note, certaines familles comptaient plus de deux germains.

Pour les 53 échantillons EUR, les LCLs ont été mises en culture à Ulm (Allemagne) par notre collaborateur Bernard Boehm. Pour chaque lignée, deux réplicats ont été préparés, selon trois conditions: (1) non stimulées, (2) stimulées 6h avec du PMA (phorbol myristate acetate) et (3) stimulées 6h avec du PMA puis lavées et collectées 18h plus tard, soit 24h après la stimulation. Pour 45 sujets (incluant 20 diabétiques issus de 16 familles avec 23 paires discordantes), les ARNs ont été collectés en vue de caractériser leur transcriptome au moyen de puces d'expression en 3' de type Illumina (Human_HT12 V4 Expression BeadChip) auprès de notre collaborateur Grant Moraham à Perth (Australie). Au total, 264 puces ont été hybridées correspondant pour 39 sujets aux lignées cultivées en 3 conditions de stimulation avec 2 réplicats de culture, tandis que pour 6 autres sujets non atteints, il manque un réplicat de culture dans une des trois conditions.

Les données de puces ont été analysées avant mon arrivée au laboratoire. De plus, pour chaque réplicat et condition de culture des 53 sujets EUR, un aliquot d'ARN, un culot de cellules et les surnageants de culture ont été collectés et nous ont été envoyés. J'ai réalisé toutes les validations des résultats d'expression sur ces échantillons, au niveau des transcrits et des protéines.

Pour les LCLs des NAm, deux réplicats de culture ont été réalisés sans stimulation à l'université de Virginie par notre collaborateur Patrick Concannon. Un culot de cellules nous a été envoyé. J'ai réalisé les extractions d'ARN de ces échantillons et les expériences de réplication des données d'expression des transcrits sur ce matériel.

Par ailleurs, des ampoules congelées de PBMCs étaient également disponibles pour un sous-échantillon : 39 sujets de la cohorte EUR (21 diabétiques, 19 familles, 21 paires discordantes) et 49 sujets de la cohorte NAm (26 diabétiques, 24 familles, 25 paires discordantes). J'ai réalisé toutes les extractions d'ARN à partir de ces PBMCs en vue d'expériences de transcriptome par la technologie RNA-Seq brin spécifique réalisées par la compagnie Fasteris (Suisse).

Pour chaque échantillon, 10 millions de paires de reads de 125 bases ont été obtenues et les analyses sont en cours dans notre laboratoire.

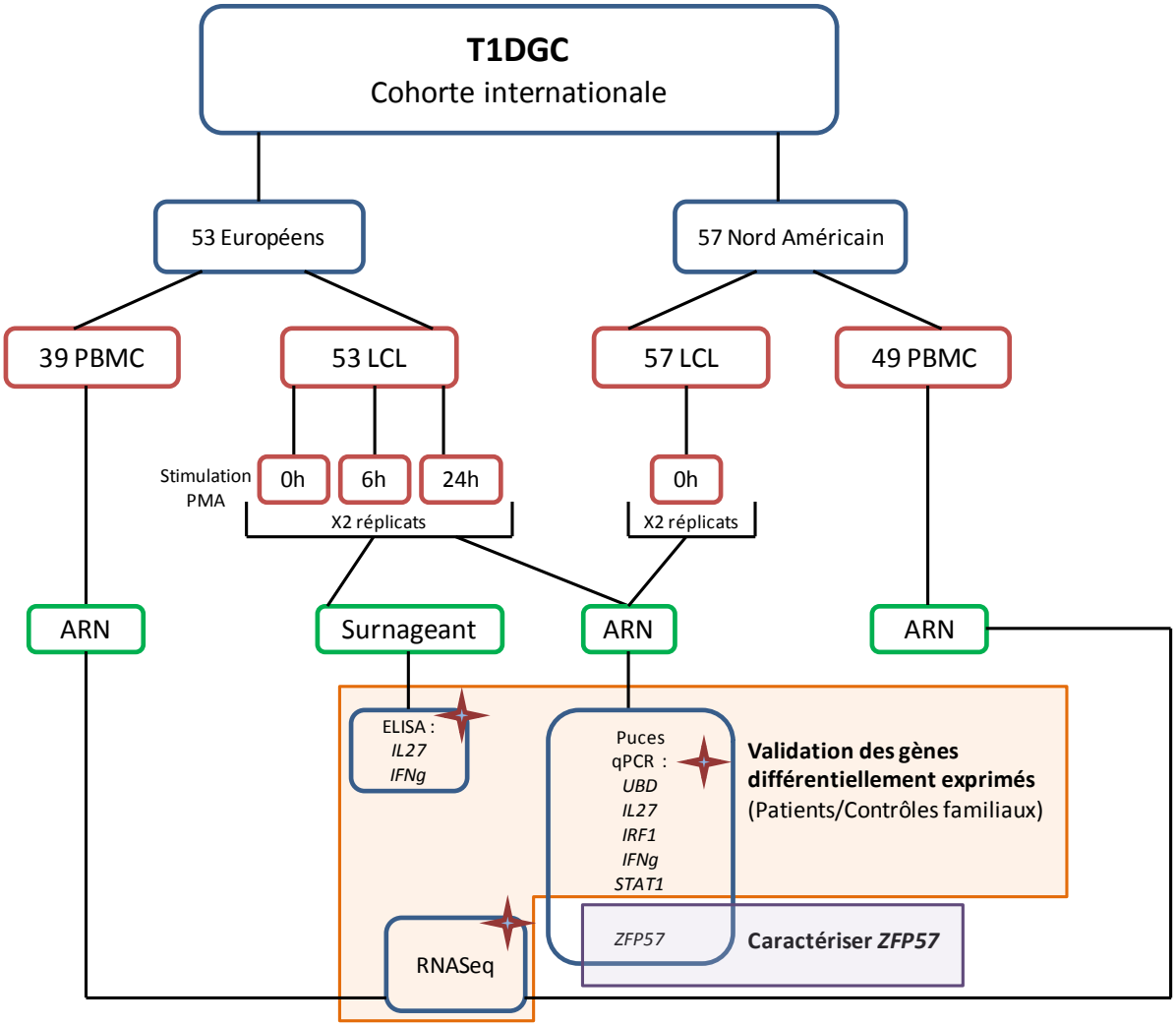


Figure 17. Bilan des échantillons disponibles et des techniques réalisées dans la cohorte du T1DGC. Les étoiles indiquent les techniques que j'ai prises en charge au laboratoire.

1.4. Lignées cellulaires

Quatre lignées cellulaires lymphoblastoïdes (LCLs), appelées COX, PGF, QBL et APD (ECACC, European Collection of Authenticated Cell Cultures), établies à partir d'individus homozygotes pour différents haplotypes du CMH ont été étudiées comme modèle (**Figure 18**). Elles ont été maintenues en culture dans du milieu RPMI 1640 (Sigma) en présence de 10% de sérum de veau fœtal (PAA), 2 mM de glutamine (Sigma) et 100 unités par ml de pénicilline et streptomycine (Sigma) à 37°C dans un environnement humidifié avec 5% de CO2. J'ai pris en charge tout le travail concernant ces lignées.

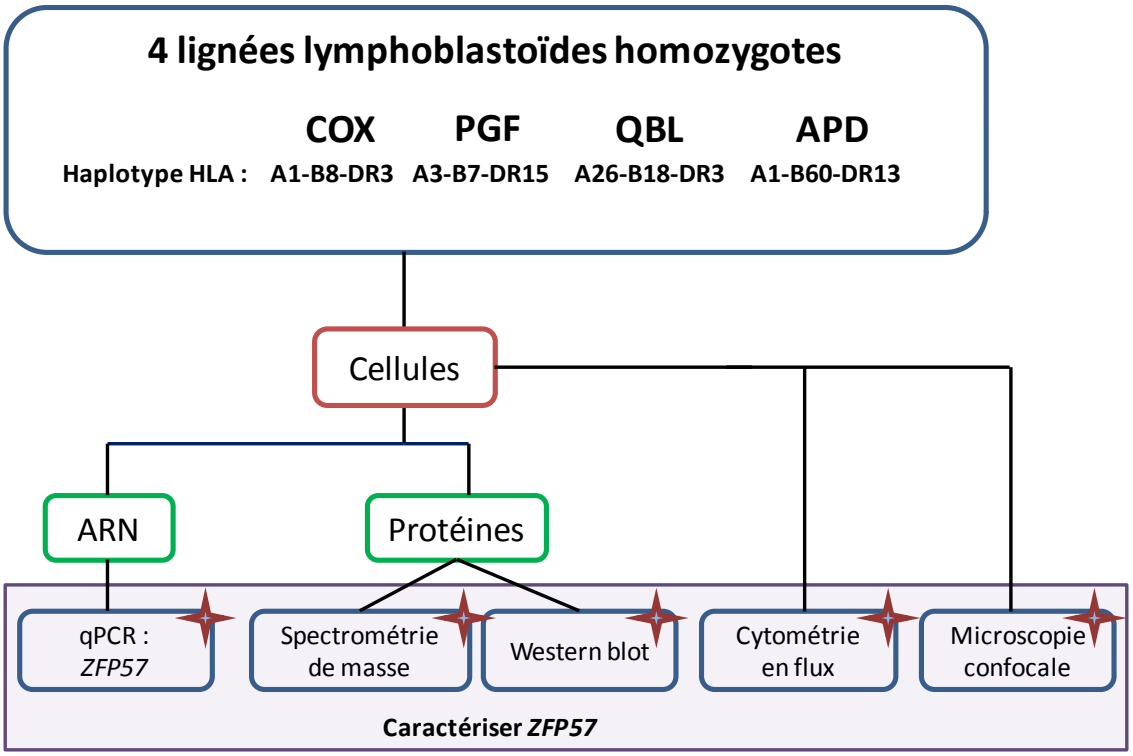


Figure 18. Bilan des échantillons disponibles et des techniques réalisées dans les 4 lignées lymphoblastoïdes. Les étoiles indiquent les techniques que j'ai prises en charge au laboratoire.

2. Génotypage

Tous les génotypages ont été conduits en aveugle des données cliniques et relus en double.

2.1. Génotypage par puce Illumina, l'immunochip

Les données brutes ont été extraites dans notre équipe avec le logiciel Genome Studio (Illumina). Après filtrage (unicité génomique, concordance entre les séquences des sondes et la séquence sur le génome et concordance des allèles du manifeste avec ceux du génome), 185 114 SNPs (82,7%) ont été conservés, dont 8056 SNPs dans le CMH.

L'analyse de la cohorte de « découverte » Lyon 1 a permis d'identifier des associations significatives au seuil 5.10^{-8} au niveau de la région du CMH, incluant rs9273363 présentant l'association la plus significative, mais aussi rs2187668 et rs7454108 taggant respectivement HLA-DR3 et HLA-DR4 (cf **Tableau 6**, section Résultats). En dehors du CMH, 3 SNPs (rs62408234, rs62408223 et rs17638639) présentaient une association suggestive au seuil 10^{-5} et 4 SNPs (rs75460852, rs1604643, rs2838514 et rs7739286) au seuil 10^{-4}. Un onzième SNP (rs11160607) a enfin été identifié comme présentant une association suggestive avec un biais de transmission paternel. Enfin deux SNPs, l'un dans le gène de l'insuline (rs689), l'autre dans le gène PTPN22 (rs2476601), avaient été précédemment identifiés dans d'autres cohortes caucasiennes comme ayant les effets les plus importants après le CMH. Ils n'étaient cependant pas significatifs dans notre cohorte de « découverte », mais nous les avons inclus dans la cohorte de « réplication ». Le travail de génotypage de la cohorte de « réplication » Lyon 2 a été réalisé par trois stagiaires de M2, Xia Zhao, Neda Rezaei et Tharshana Stephen, par une ingénieure d'études, Valérie Sénée, et par moi-même (**Tableau 5**).

Tableau 5. Récapitulatif des SNPs génotypés pour le projet de réplication French GWAS (gris) et celui de l'étude génétique de *ZFP57* (blanc). Pour chaque SNP sont précisés sa position génomique (GRCh38), le gène cible, sa position par rapport à celui-ci, le nucléotide reconnu par la sonde Taqman couplée au fluorophore VIC ou FAM, le brin sur lequel il est positionné et la fréquence de l'allèle mineur (MAF) dans la population de sujets sains.

Projet	SNP	GRCh38	Cible	Position par rapport à la cible	VIC	FAM	Brin	MAF
Réplication	rs2187668	chr6:32638107	*HLA-DQA1*	Intron 1	C	T	-	0,08 (T)
	rs7454108	chr6:32713706	*DR4_DQ8*	47 kb en 5' de *HLA-DQB1* et 27,6 kb en 5' de *HLA-DQA2*	C	T	+	0,08 (C)
	rs9273363	chr6:32658495	*HLA-DQB1*	969 pb en 3' de *HLA-DQB1*	A	C	+	0,24 (A)
	rs62408234	chr6:90268442	*BACH2*	Intron 1	G	A	-	0,08 (A)
	rs62408223	chr6:90239477	*BACH2*	Intron 1	T	C	-	0,08 (G)
	rs17638639	chr2:178161391	Proche de *OSBPL6*, *RBM45* et *PDE11A*	33 kb en 5' de *OSBPL6* et à 53 kb en 5' de *PDE11A* et 31 kb en 3' de *RBM45*	C	G	+	0,12 (G)
	rs75460852	chr1:101074135	Proche de *DHP5* (1p21.2)	84,5 kb en 5' de *DHP5*	T	A	+	0,02 (A)
	rs1604643	chr3:158935845	Entre *MFSD1* et *IQJC*	106 kb en 3' de *MFSD1* et 133 kb en 5' de *IQJC*	A	G	+	0,07 (G)
	rs2838514	chr21:44188454	*C21ORF33*	42,7 kb en 3' de *C21ORF33*	C	T	+	0,20 (C)
	rs7739286	chr6:105985700	*PRDM1*	100 kb en 5' de *PRDM1*	A	C	+	0,22 (C)
	rs689	chr11:2160994	*Insuline*	Intron 1	A	T	+	0,35 (A)
	rs2476601	chr1:113834946	*PTPN22*	Exon 13	A	G	+	0,03 (A)
	rs941576	chr14:100839708	*MEG3*	Intron 6	A	G	+	0,38 (G)
	rs11160607	chr14:100836890	*MEG3*	Intron 6	C	T	+	0,20 (T)
ZFP57	rs3129054	chr6:29681279	*ZFP57*	4 kb en 5' de *ZFP57*	C	T	+	0,34 (T)
	rs2235383	chr6:29725722	*ZFP57*	Intron 5 de HLA-F, 48,5 kb en 5' de *ZFP57*	A	G	+	0,23 (G)
	rs7741807	chr6:29712880	*ZFP57*	35,7 kb en 5' de *ZFP57*	C	G	+	0,23 (G)
	rs7759272	chr6:29717217	*ZFP57*	40 kb en 5' de *ZFP57*	A	G	-	0,23 (G)
	rs9257932	chr6:29658599	*ZFP57*	Intron 1 de *MOG*, 13,8 kb en 3' de *ZFP57*	C	T	+	0,04 (C)
	rs2076177	chr6:29725336	*ZFP57*	Intron 5 de *HLA-F*, 48 kb en 5' de *ZFP57*	C	T	+	0,23 (T)
	rs396660	chr6:29678388	*ZFP57*	1 kb en 5' de *ZFP57*	C	T	+	0,36 (T)
	rs445150	chr6:29679102	*ZFP57*	2 kb en 5' de *ZFP57*	A	G	-	0,36 (G)
	rs375984	chr6:29676725	*ZFP57*	Intron 1	C	T	+	0,27 (T)
	rs2747429	chr6:29680600	*ZFP57*	3,4 kb en 5' de *ZFP57*	C	T	+	0,27 (C)
	rs2535238	chr6:29677261	*ZFP57*	107 pb en 5' de *ZFP57*	A	C	+	0,27 (A)
	rs2747421	chr6:29677341	*ZFP57*	187 pb en 5' de *ZFP57*	C	G	+	0,27 (C)
	rs2747431	chr6:29680787	*ZFP57*	3,6 kb en 5' de *ZFP57*	C	T	+	0,34 (T)
	rs365052	chr6:29680621	*ZFP57*	3,5 kb en 5' de *ZFP57*	C	G	+	0,34 (C)
	rs416568	chr6:29679851	*ZFP57*	2,7 en 5' de *ZFP57*	T	A	+	0,34 (A)
	rs2269422	chr6:32183517	*ZFP57*	Intron 3 d'*AGER*, 2,5 Mb en 5' de *ZFP57*	C	T	+	0,04 (C)

2.2. Génotypage par sonde Taqman

Le génotypage par la méthode Taqman permet un criblage rapide des SNPs d'intérêt. Cette technique a été réalisée sur des plaques de 384 puits avec le thermocycleur Biorad CFX384 Touch Real-Time PCR Detection System (Bio-Rad).

La technique Taqman utilisée pour le génotypage a pour particularité d'utiliser deux sondes fluorescentes : une sonde complémentaire de l'allèle 1 marquée en 5' par le fluorophore FAM (excitation à 495 nm/émission à 515 nm) et une sonde complémentaire de l'allèle 2 marquée en 5' par le fluorophore VIC (excitation à 535 nm/émission à 555 nm). Au cours de la réaction en chaine d'amplification (PCR, Polymerase chain reaction), les sondes fluorescentes s'hybrident à la séquence qui leur est complémentaire. Lorsque la Taq (*Thermus aquaticus*) polymérase synthétise le brin d'ADN complémentaire, elle clive ces sondes grâce à son activité 5' exonucléasique libérant ainsi le fluorophore et permettant son émission de fluorescence qui peut être quantifiée (**Figure 19**). Les réactions ont été réalisées dans un volume final de 3 µl : 1 µl de Master Mix (Applied Biosystem), 0,0375 µl de sondes (Applied Biosystem), 1 µl d'ADN à une concentration de 5 ng/µl et complété avec de l'eau. Le programme de PCR était le suivant : 92°C-10 sec + (92°C-15 sec +60°C-60 sec + lecture plaque) x (40 à 60 cycles) + 72°C -30 sec (courbe de fusion). Les génotypes sont assignés à l'aide du logiciel Bio-Rad CFX Manager 3.1 en fonction de la fluorescence pour chaque allèle (**Figure 20**).

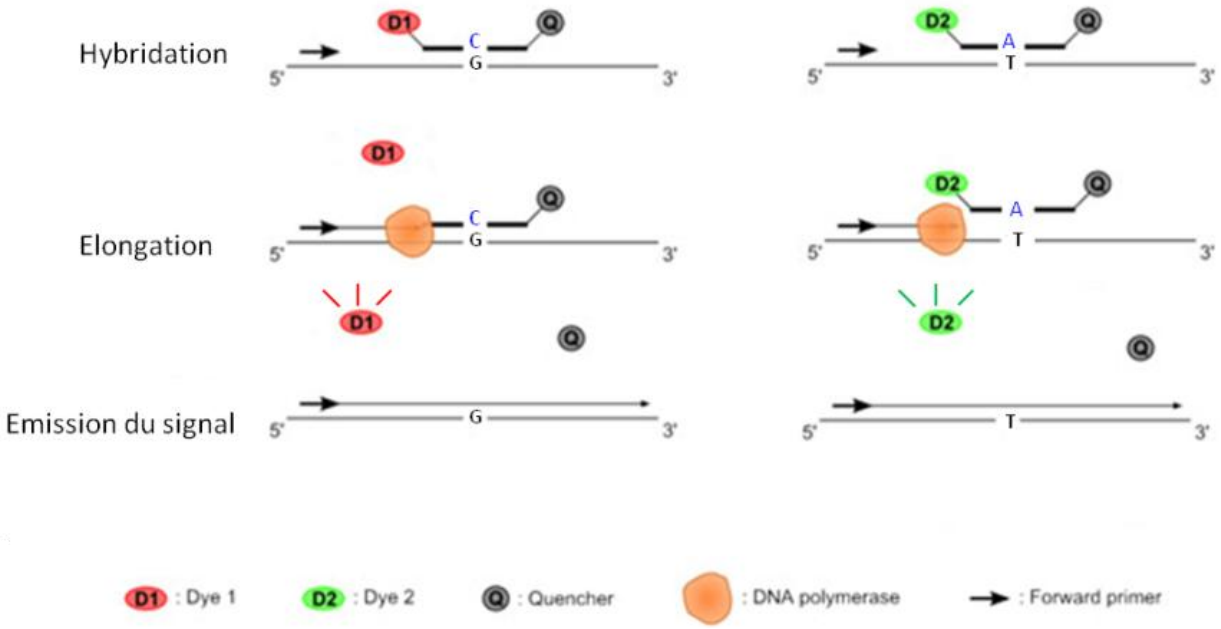


Figure 19. Principe de la méthode de génotypage par sondes Taqman (Applied Biosystems).

1. Après dénaturation, il y a hybridation des amorces (primers) et de la sonde allèle-spécifique qui contient un fluorophore FAM (Dye 1) ou VIC (Dye 2) ainsi qu'un quencher, qui du fait de sa proximité empêche le fluorophore d'émettre son signal (principe de la méthode FRET, Fluorescence Resonance Energy Transfer).
2. Elongation grâce à la DNA polymérase et libération du fluorophore indépendamment du quencher
3. Emission de fluorescence FAM ou VIC selon le génotype de l'individu et détection du signal

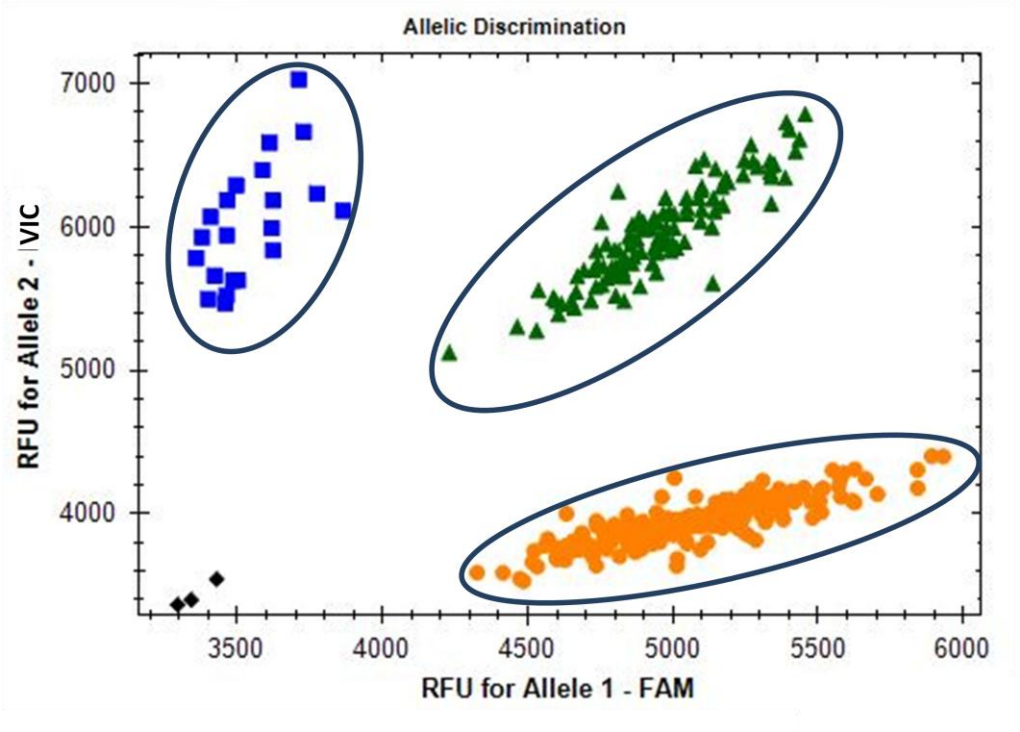


Figure 20. Capture d'écran de résultats obtenus sur le logiciel Bio-Rad CFX Manager 3.1 pour le génotypage par sondes Taqman d'un SNP. Chaque symbole de couleur représente un sujet génotypé : les cercles orange correspondent aux sujets homozygotes pour l'allèle 1 (émission du fluorophore FAM), les carrés bleus aux sujets homozygotes pour l'allèle 2 (émission du fluorophore VIC), les triangles verts aux sujets hétérozygotes pour les allèles 1 et 2 (émission des deux fluorophores). Les losanges noirs représentent les témoins négatifs (sans matrice) (émission de fluorescence sous le seuil).

3. Analyse des transcrits

3.1. Extraction d'ADN et d'ARN

Les ARNs et ADNs sont extraits selon le protocole du kit AllPrep DNA/RNA Mini (Qiagen, ref. 80204) incluant pour l'ARN l'étape de digestion à la DNAse I (**Annexe 1**). Des culots de 1 à 5.10^6 cellules sont collectés. Les cellules en culot sont lysées en ajoutant 350 µL (600 µL pour 5.10^6 à 1.10^7) de tampon RLT Plus (agent chaotropique dénaturant les protéines enrichi en détergents permettant la liaison efficace de molécules d'ADN) complété par 1% de beta-mercaptoethanol (agent réducteur des protéines par action anti-oxydante) et en vortexant. Pour bien homogénéiser, le lysat cellulaire est passé 5 à 10 fois à travers une seringue 20G (0,9 mm). Le lysat est ensuite déposé sur une colonne AllPrep DNA puis centrifugé 30 secondes à 10 000 rpm. L'ADN (15 à 30 kb) se fixe à la membrane de silice de la colonne (associée à un nouveau tube collecteur de 2 mL) et est temporairement conservé à 4°C avant purification.

Purification de l'ARN total

La partie non retenue par la colonne correspond à l'ARN total (>200 nucléotides) à purifier. Pour cela un volume d'éthanol à 70% (350 ou 600 µL suivant le volume de tampon RLT Plus utilisé) est ajouté, puis le lysat est homogénéisé par pipettage. L'échantillon est transféré sur une colonne RNeasy placée sur un tube collecteur de 2 mL puis centrifugé 30 secondes à 10 000 rpm. La partie non retenue par la colonne est jetée. Si l'échantillon dépasse 700 µL, le reste est placé sur la colonne puis centrifugé de la même façon. Une première étape de lavage consiste à ajouter à la colonne 350 µL de tampon RW1, tampon de lavage stringent contenant des sels de guanidine et de l'éthanol, permettant d'éliminer efficacement les biomolécules non spécifiquement liées à la

membrane, telles que les carbohydrates, protéines ou acides gras, tout en maintenant l'ARN spécifiquement lié à la membrane. La colonne est centrifugée 30 secondes à 10 000 rpm. La partie non retenue est jetée. Puis, un mélange de 10 µL de DNAse I et 70 µL de tampon RDD est ajouté à la colonne qui est incubée 15 min à température ambiante, afin de digérer l'ADN contaminant qui n'aurait pas été retenu par la colonne à l'étape initiale. Un second lavage avec 350 µL de tampon RW1 est effectué comme précédemment. Un volume de 500 µL du tampon de lavage doux RPE contenant de l'éthanol est ajouté puis la colonne est centrifugée 2 min à 10 000 rpm afin d'éliminer les traces de sels restant. La partie non retenue de la colonne est jetée puis un nouveau tube collecteur est placé en dessous. La colonne est centrifugée à vitesse maximale 1 min, puis placée sur un tube collecteur eppendorf de 1,5 mL. Enfin, l'ARN est élué de la colonne par le dépôt de 20 à 50 µL d'H_2O RNase-free (en fonction du nombre de cellules de départ ou de la taille du culot) et centrifugation 1 min à 10 000 rpm. L'éluat est récupéré et déposé de nouveau sur la colonne puis centrifugé 1 min à 10 000 rpm afin d'augmenter la concentration en ARN élué. L'ARN est dosé et stocké à -80°C.

Purification de l'ADN

Un volume de 500 µL de tampon AW1, contenant du chlorure de guanidinium afin de dénaturer les protéines, est ajouté sur la colonne AllPrep DNA, conservée précédemment à 4°C. Après centrifugation 30 secondes à 10 000 rpm, la partie non retenue est jetée. Afin d'éliminer les sels contaminants, 500 µL de tampon AW2 (tampon salin contenant essentiellement de l'étanol 70%) sont ajoutés. La colonne est centrifugée 2 min à 13000 rpm puis placée sur un tube collecteur eppendorf de 1,5 mL. Enfin, 100 µL de tampon EB (Elution Buffer) sont ajoutés. Une incubation finale d'1 min à température ambiante est nécessaire avant de centrifuger 1 min à 10 000 rpm pour éluer l'ADN. Cette dernière centrifugation est renouvelée avec le premier éluat afin d'augmenter la concentration finale d'ADN obtenu. L'ADN est dosé et stocké à -20°C.

Dans le cas où seul l'ARN est extrait, le kit Rneasy Mini Kit (Qiagen, ref. 74104) sera utilisé.

Les étapes sont identiques au précédent kit sauf qu'il n'y a pas de colonne à ADN.

3.2. Quantification et contrôle de qualité de l'ADN et de l'ARN

Le dosage de l'ADN et de l'ARN est effectué par spectrophotométrie avec le Nanodrop 2000 (Thermo scientific). La concentration de ces acides nucléiques est calculée à partir de leur propriété d'absorbance à une longueur d'ondes de 260 nm, sachant qu'une unité de densité optique à 260 nm équivaut à 50 µg/mL d'ADN double brin et 40 µg/mL d'ARN. En plus de la quantification des acides nucléiques, la spectrophotométrie permet d'évaluer leur pureté. Les protéines absorbant à 280 nm et les sels à 230 nm, la mesure du ratio DO260nm/DO280nm permet d'estimer la contamination protéique, tandis que le ratio DO260nm/DO230nm évalue la contamination en composants organiques et en sels chaotropiques. Pour un ADN correctement purifié, ces rapports doivent être proches de 1,8 et pour l'ARN proche de 2. Pour l'ARN total, l'intégrité est évaluée par électrophorèse sur un Bioanalyzer 2100 (Agilent) avec le kit RNA 6000 Nano qui permet de vérifier que l'ARN n'est pas dégradé. Deux paramètres sont considérés :

- le rapport des aires correspondant aux ARNs ribosomaux 28S (5kb) et 18S (1,9kb), dont la valeur optimale est de 2. Une mauvaise qualité de l'ARN se traduit par une diminution de ce ratio.
- le RNA Intergity Number (RIN), valeur déterminée par le bioanalyzer sur la base du profil de migration. Le RIN est compris entre 1 et 10, 10 reflétant une excellente qualité.

3.3 Reverse transcription

Les ARNs sont rétro-transcrits en ADNs complémentaires (ADNc) simple brin en utilisant le kit SuperScript III (Invitrogen) selon les recommandations du fabricant. Brièvement, 1 µg d'ARN est dénaturé pendant 5 min à 65°C en présence d'1 µl d'amorces hexamères aléatoires à 50 ng/µL et 1 µL de dNTP à 10 mM en complétant avec de l'eau jusqu'à 10µL, puis mis à 4°C. Après ajout de 2 µL de tampon 10X, 4 µL de MgCl2 à 25 mM, 2 µL de DTT à 0,1M, 1 µL de RNase OUT à 40 U/µL et 1 µL de SuperScript III, la reverse transcription est réalisée selon le programme suivant : 25°C-10 min, 50°C-50 min, 85°C-5 min, 4°C-fin. Les ARNs hybridés à l'ADNc sont éliminés en présence d'1 µL de RNAse H pendant 20min à 37°C. Pour chaque échantillon, un contrôle négatif sans reverse transcriptase est également préparé. Les ADNc ont été ajustés aux concentrations de 50 ng/µL équivalent ARN.

La synthèse de l'ADNc est vérifiée par amplification par PCR avec des amorces correspondant à la β-actine dans un volume final de 15 µL avec 1,5 µL de tampon 10 X, 0,6 µL de MgCl2 à 50 mM, 0,6 µL de dNTP à 10 mM, 0,15 µL de chaque amorce à 10 µM, 0,5 µL d'ADNc, 0,1 µL de Taq Platinum (Invitrogen) (5 U/µL), et complété avec de l'eau. Le programme de la PCR est le suivant : 94°C-1 min, (94°C-30 sec, 58°C-30 sec, 72°C-1 min) x 40 cycles, 72°C-7 min, 10°C-fin. Les produits de PCR sont ensuite analysés par électrophorèse sur gel d'agarose 2%.

3.4. Quantitative-Polymerase Chain Reaction (qPCR)

L'expression de six gènes cibles a été étudiée par qPCR : *ZFP57*, *IRF1*, *IL27*, *STAT1*, *IFNγ* et *UBD*. Aucun facteur de dilution n'a été appliqué pour *ZFP57*. Pour les quatre autres gènes, une dilution au demi de la matrice d'ADNc a été effectuée.

Trois gènes de ménage, *GAPDH*, *BACT* et *RPL30* sont utilisés pour la normalisation. Leur expression étant plus forte que celle des gènes cibles, la matrice d'ADNc a été diluée au 8ème.

L'amplification des produits est mesurée par l'émission de la fluorescence d'un intercalant, le SYBR Green (Biorad), lorsqu'il est incorporé dans le double brin d'ADNc. Les cycles de PCR et la mesure de la fluorescence sont réalisés dans un thermocycler de type Biorad CFX384 Real Time system, dans des plaques de 384 puits remplies à la main.

Les réactions de qPCR sont réalisées dans un volume final de 15 µL avec 7,5 µL de SYBR Green 2X, 0,15 µL de chaque amorce à 10 µM, 2 µL d'ADNc à 50 ng/µL et complété avec de l'eau. Le programme de la qPCR est le suivant : 50°C-10 min, 94°C-5 min, (94°C-10 sec, 57,5°C-30 sec-lecture) x 59 cycles, 95°C-10 sec, courbe de fusion, fin.

L'efficacité de chaque paire d'oligonucléotides a été évaluée à l'aide d'une dilution en série d'un pool d'ADNc. Toutes les amplifications ont été réalisées en triplicats techniques et incluent les contrôles négatifs sans reverse-transcriptase. Pour les réplicats techniques, un coefficient de variation (écart type/moyenne) supérieur à 1% a été considéré comme mauvais et les expériences ont été refaites pour ces échantillons.

L'analyse de la fluorescence a été réalisée au moyen du logiciel Bio-Rad CFX Manager version 3.1 pour chaque paire d'amorce indépendamment. Parmi les paramètres utilisés, la ligne basale est déterminée en choisissant, parmi tous les puits de la plaque, la valeur minimale observée comme valeur de début et la valeur maximale observée comme valeur de fin. Les valeurs Cq sont déterminées par la méthode de la ligne de seuil. L'ordonnée de cette ligne de seuil est calculée automatiquement par le logiciel avec les paramètres par défaut pour toutes les plaques, puis la valeur maximale observée est appliquée manuellement à toutes les plaques de manière à assurer

une calibration identique pour minimiser le biais éventuel conféré par le facteur « plaque » dans les analyses statistiques. Les niveaux relatifs de transcription des gènes cibles ont été déterminés par la méthode ΔΔCT en utilisant les efficacités réelles de chaque paire d'oligonucléotides (**Annexe 2**).
Afin de s'assurer de la comparabilité des résultats obtenus, nous nous conformons aux recommandations MIQE (**Annexe 3**) (Minimal Information for qPCR Experiments, http://www.rdml.org/miqe.php) (Bustin et al. 2009).

4. Analyse des protéines

4.1. Extraction protéique

Les cellules en suspension sont préalablement lavées dans du tampon phosphate salin (PBS) froid puis culotées par centrifugation à 6000 rpm pendant 5 minutes à 4°C.

4.1.1. Fraction totale

Pour préparer un lysat de protéines totales, selon le volume estimé du culot, on rajoute 5 volumes équivalents de tampon de lyse RIPA (RadioImmunoPrecipitation Assay) buffer (Sigma, ref. R0278) auquel on a ajouté des inhibiteurs de protéases (Roche, ref 04693159001). Le culot est homogénéisé par pipetage et vortex et laissé 5 min dans la glace, le temps que la lyse agisse. Le lysat protéique est conservé à -80°C.

4.1.2. Fraction cytoplasmique et nucléaire

Pour préparer un lysat de protéines cytoplasmiques et nucléaires, selon le volume estimé du culot, on rajoute 5 volumes de tampon d'extraction cytoplasmique (CE) froid : 10 mM d'HEPES à pH 7,9, 10 mM de KCl, 0,1mM d'EDTA, dans lequel est ajouté 0,3% de NP-40 ainsi que des inhibiteurs de protéases. Les cellules sont resuspendues et incubées 5 min dans la glace en vortexant régulièrement puis centrifugées 5 min à 3000 rpm à 4°C. Le surnageant qui correspond au lysat cytoplasmique est récupéré.
Le culot est repris avec 100 µl de tampon d'extraction cytoplasmique (CE) sans NP-40 puis lavé par centrifugation 5 min à 3000 rpm à 4°C. Selon le volume estimé du culot, on rajoute un volume équivalent de tampon d'extraction nucléaire (NE) : 20 mM d'HEPES à pH 7,9, 0.4 M de NaCl, 1 mM d'EDTA, 25% de glycérol, contenant des inhibiteurs de protéases. Le culot est resuspendu et incubé 10 min dans la glace en vortexant régulièrement. Après centrifugation 5 min à 14000 rpm à 4°C, le surnageant qui correspond au lysat nucléaire est homogénéisé. Les lysats cytoplasmiques et nucléaires sont conservés à -80°C.

4.1.3. Fractionnement infracellulaire des protéines cytoplasmiques, membranaires, nucléaires, chromatiniennes et du cytosquelette

Le fractionnement des protéines infracellulaires permet de localiser et d'enrichir les protéines à partir de compartiments cellulaires spécifiques. Cette extraction séquentielle des cinq fractions, cytoplasmique, membranaire, nucléaire soluble, liées à la chromatine et cytosquelettique est réalisée à l'aide du kit Subcellular Protein Fractionation Kit for Cultured Cells (Thermo Scientific, ref 78840) (**Figure 21**).

Le protocole part d'un nombre total de 10.10^6 cellules qui sont culotées dans un tube eppendorf de 1,5 mL par centrifugation 5 min à 500g. Le surnageant est retiré précautionneusement afin de garder le culot sec. Puis le tampon d'extraction du cytoplasme (CEB = Cytosol Extraction Buffer) contenant des inhibiteurs de protéases est ajouté au culot en fonction de sa taille. Cette étape perméabilise la membrane de manière sélective afin de libérer la fraction cytoplasmique soluble. Le culot est incubé 10 min à 4 °C sous agitation puis centrifugé 5 min à 500g. Le surnageant est immédiatement transféré dans un nouveau tube de type eppendorf : il constitue la fraction cytoplasmique.

Le culot est repris dans le tampon d'extraction de la membrane (MEB = Membrane Extraction Buffer) froid contenant des inhibiteurs de protéases, vortexé 5 secondes et incubé 10 min à 4°C sous agitation. Cette étape dissout les membranes plasmique, des mitochondries, du réticulum endoplasmique et du Golgi mais ne solubilise pas les membranes nucléaires. Après centrifugation 5 min à 3000g, le surnageant est transféré dans un nouveau tube eppendorf : il constitue la fraction membranaire.

Le noyau intact est récupéré dans le culot. Du tampon d'extraction nucléaire (NEB = Nuclear Extraction Buffer) froid et contenant des inhibiteurs de protéases lui est ajouté, puis il est vortexé 15 secondes et incubé 30 min à 4°C sous agitation. Cette étape libère l'extrait nucléaire soluble. Après centrifugation 5 min à 5000 g, le surnageant est transféré dans un nouveau tube eppendorf : il constitue la fraction nucléaire soluble, tandis que le culot insoluble contient les protéines du cytosquelette et de la chromatine. Une extraction avec la nucléase micrococcale est réalisée sur le culot afin de relarguer les nucléoprotéines liées à la chromatine. Le tampon permettant de récupérer cette fraction chromatinienne est préparé en ajoutant 5 µL de $CaCl_2$ (100 mM) et 3 µL de Micrococcal Nucléase (300U) pour 100 µL de tampon NEB et des inhibiteurs de protéases. Le culot est ainsi repris dans ce tampon non froid, vortexé 15 secondes et incubé 5 min à 37°C. Après incubation le culot est vortexé 15 secondes puis centrifugé 5 min à 16 000 g. Le surnageant est transféré dans un nouveau tube eppendorf : il constitue la fraction chromatinienne. Enfin, le culot sauvegardé est repris dans du tampon d'extraction de culot (PEB = pellet extraction buffer) contenant des inhibiteurs de protéases, vortexé 15 secondes et incubé 10 min à température ambiante. Cette étape solubilise les protéines cytosquelettiques. Après centrifugation 5 min à 16 000 g, le surnageant est transféré dans un nouveau tube eppendorf : il constitue la fraction du cytosquelette. Tous les extraits protéiques sont dosés par BCA (voir ci-après) puis conservés à -80°C.

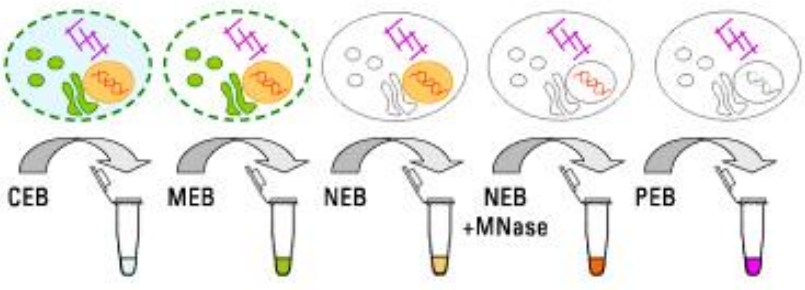


Figure 21. Principe du fractionnement cellulaire réalisé à l'aide du kit Subcellular Protein Fractionation Kit for Cultured Cells (Thermo Scientific, ref. 78840). Les compartiments cellulaires sont extraits séquentiellement par incubation des cellules dans du tampon d'extraction cytoplasmique (CEB) libérant les protéines du cytoplasme (en bleu clair) suivi du tampon d'extraction membranaire (MEB) libérant les protéines des membranes plasmique, des mitochondries, du RE et du Golgi (en vert) et du tampon d'extraction nucléaire (NEB) libérant les protéines nucléaires solubles (en jaune) . L'ajout de la nucléase micrococcale (MNase) au tampon NEB sert à extraire les protéines liées à la chromatine (en orange), puis le tampon PEB est ajouté afin de récupérer les protéines du cytosquelette (en violet).

4.2. Dosage colorimétrique des protéines par la méthode à l'acide bicinchoninique (BCA)

Le dosage protéique de type BCA est une extension de la méthode du Biuret, plus sensible et plus linéaire. Cette méthode porte sur le dosage des liaisons peptidiques. En présence de peptides contenant au moins trois acides aminés et dans un milieu alcalin, les ions Cu^{2+} sont réduits en l'ion métallique cuivreux Cu^{1+}. Les ions Cu^{1+} sont détectés de manière sélective et sensible par l'acide bicinchoninique, formant un complexe BCA/cuivre hydrosoluble de couleur pourpre ayant une absorption maximale à 562 nm. Ce dosage est réalisé avec les réactifs du kit BCA (Thermo Scientific, ref. 23225) et quantifié par colorimétrie avec le spectrophotomètre Nanodrop 2000 (Thermo scientific). Tout d'abord, une gamme d'étalonnage est réalisée par dilution en série d'albumine bovine dans une plage de linéarité de 20 à 2000 µg/ml (Sigma). Puis les protéines sont quantifiées à une DO de 562 nm. On utilise l'équation de la droite de régression linéaire obtenue grâce aux concentrations des standards pour calculer la concentration protéique dans les échantillons. Les concentrations obtenues par le kit BCA serviront à normaliser les concentrations de protéines spécifiques mesurées par ELISA (Enzyme Linked Immuno Sorbent Assay).

4.3. ELISA (Enzyme Linked Immuno Sorbent Assay) sandwich

La technique d'ELISA est un dosage immuno-enzymatique qui permet de détecter la présence de protéines, d'anticorps ou d'antigènes dans un échantillon. L'ELISA sandwich utilise, dans un premier temps, un anticorps spécifique de la protéine d'intérêt qui est immobilisée sur un support en plastique. Dans un deuxième temps, un second anticorps spécifique couplé à une enzyme, la Horseradish Peroxidase (HRP), est fixé sur la protéine d'intérêt. Le système est ensuite révélé par l'addition d'un substrat incolore mais chromogène, le 3,3',5,5'-TetramethylBenzidine (TMB). Le produit de la réaction est de couleur bleu, absorbant à 370 nm ou entre 620 et 655 nm. L'intensité de la couleur sera proportionnelle à la quantité de protéine présente dans l'échantillon. La réaction est stoppée en présence d'un acide fort colorant en jaune, le produit de la réaction, qui peut être quantifié à 450 nm.

Dans ce projet, j'ai testé l'expression des cytokines sécrétées IL27 (Human IL-27 DuoSet ELISA, rndsystems, ref. DY2526) et INFγ (Human IFN-γ DuoSet ELISA, rndsystems, ref. DY285) dans les échantillons de 311 surnageants des 53 lignées lymphoblastoides de la cohorte T1DGC. Les expériences ont été réalisées en 2 réplicats techniques.

Le protocole détaillé utilisé est le suivant. L'anticorps de capture est déposé dans une plaque en polystyrène 96 puits (rndsystems, ref. DY990) qui est mise à incuber toute la nuit à température ambiante sur un agitateur de plaques. La plaque est lavée trois fois avec 400 µL de tampon de lavage (PBS+0,05% Tween20). Une étape de blocage est réalisée en déposant 300 µL de Block buffer (PBS+1% BSA) dans chaque puits pendant 1h à température ambiante. Cette étape est nécessaire pour bloquer les liaisons non spécifiques de protéines. Elle est suivie de trois lavages. Les échantillons sont centrifugés à 13000 rpm à 4°C afin de culoter les débris. Pour chaque échantillon, 100 µL sont déposés puis la plaque est incubée 2h sur un agitateur. Après trois lavages, 100 µL d'anticorps de détection sont déposés dans chaque puits. Après une incubation de 2h à température ambiante, la plaque est lavée trois fois. Ensuite, 100 µL par puits de Streptavidin-HRP sont ajoutés. La plaque est protégée de la lumière et incubée 20 min à température ambiante et lavée trois fois. Puis 100 µL/puits de substrat (TMB et H_2O_2, rndsystems,

ref. DY999) sont ajoutés et la plaque est incubée 20 min à l'abri de la lumière. Enfin, 50 µL par puits de solution stop (2 N H_2SO_4, rndsystems, ref. DY994), pour arrêter la réaction, sont ajoutés.
Les absorbances à 450 nm et 620 nm sont mesurées à l'aide du lecteur de plaques Mithras LB940 (Berthold). Les concentrations obtenues sont normalisées par rapport à la quantité de protéines totales mesurées par BCA pour un volume équivalent de cellules.

4.4. Western blot

Des expériences de western blot ont été réalisées sur les lignées lymphoblastoïdes COX, PGF, QBL et APD. L'échantillon protéique (5 à 15 µg) est préparé dans un volume total de contenant 20 µL de protéine, 5 µL de LDS (dodécyl-sulfate de lithium) 4X (Invitrogen, ref. NP0007), 2 µL d'agent réducteur 10X (Invitrogen, NP0009) afin de dénaturer les protéines et complété avec de l'eau. Les échantillons ainsi que 10 µL d'échelle de poids moléculaire (Invitrogen, ref. LC5800) sont chargés sur un gel d'acrylamide pré-coulé à gradient 4-12% Bis-Tris, épais de 1 mm (Invitrogen, ref. NP0321), qui permet de détecter des protéines entre 15 et 260 kDa. Il est mis à migrer 15 min à 120V puis 1h20 à 160V dans le tampon de migration contenant du SDS (Invitrogen, ref. NP0050) en présence d'une solution anti-oxydante (N,N-Diméthylformamide et bisulfite de sodium) (Invitrogen, NP0005) afin de conserver l'état réduit des protéines. Le gel est transféré sur une membrane PVDF (Polydifluorure de vinylidène) (Millipore, ref. IPVH00010), qui a la propriété d'avoir une très bonne capacité de liaison pour les protéines, pendant 1h à 30V dans le tampon de transfert (Invitrogen, ref. NP00061). La membrane est incubée pendant 1h dans le tampon de blocage (Lait-TBS-T, lait 5% - Tampon Tris Salin (TBS)-Tween 20 (0,1%), TBS = 50 mM Tris-Cl, pH 7,6 ; 150 mM NaCl) puis 1h avec l'anticorps primaire (**Annexe 4**) dilué dans du lait-TBS-T. Elle est ensuite lavée 3 fois pendant 10 min avec du TBS-T. La membrane est enfin incubée 1h en présence de l'anticorps secondaire couplé à la HRP dans du lait-TBS-T, puis de nouveau lavée 3 fois pendant 10 min avec du TBS-T avant d'être révélée avec le kit ECL chemiluminescence (Invitrogen, ref.32109). L'ECL est un substrat chemiluminescent de la HRP et dont le signal va être capté par un film autoradiographique. Pour cela, la membrane est mise en contact avec l'ECL 5 min à l'abri de la lumière puis égouttée et transférée dans une pochette en plastique. La pochette est placée dans une cassette pour film sensible aux rayons X, côté protéines vers le haut. Puis le film est placé sur le dessus de la membrane et la cassette est fermée. Le temps d'exposition peut varier pour optimiser les résultats. Le film est retiré et placé dans une solution de fixation puis de révélation et enfin rincé à l'eau et mis à sécher.

4.5. Cytométrie en flux

Des expériences de cytométrie en flux ont été conduites sur les lignées lymphoblastoïdes COX, PGF, QBL et APD, ainsi que sur les leucocytes issus d'échantillons de sang total prélevé le jour même auprès des volontaires de la cohorte CIG. Pour les LCLs, on collecte un nombre de cellules d'environ 450 000 par tube à passer au cytométre de flux. Elles sont centrifugées à 1500 rpm pendant 10 min. Pour les leucocytes, le sang total est fraichement collecté (< 2h) sur tube EDTA (7,5 ml) à raison de deux volontaires par jour. Chaque tube de sang est centrifugé à 1500 rpm pendant 10 min puis le plasma est aliquoté dans des tubes eppendorf et conservé à -80°C.
Les cellules sont reprises dans un volume de tampon de fixation (PBS 1X+SVF 1%+Formaldéhyde 1%) équivalent au volume de sang restant après centrifugation (sans le plasma), puis incubées 4 min à 37°C. Elles sont ensuite diluées au 1/10 dans du tampon de lyse (Lysing buffer 10X 500 mL :

5g de $KHCO_3$, 6,05g de Tris base, 41,5g de NH_4Cl, qsp 400 mL H_2O, pH7,4 (à peu près 11mL de HCl 1N) qsp 500mL H_2O) pour lyser les érythrocytes. Après incubation 10-12 min à 37°C, les cellules sont lavées deux fois dans 10 mL de PBS 1X (centrifugation 10 min à 1500 rpm) et resuspendues dans 1 mL de PBS 1X. Puis, 150µL de cellules sont incubées avec 60µL du mix contenant les anticorps de marquage de surface cellulaire (Lymphocyte T CD8+ (BD PerCP CD8), Lymphocyte CD4+ (BD APC CD4), Lymphocyte B (Biolegend brillant violet 785 CD19), monocytes (Biolegend brillant violet 605 CD14)) (**Annexe 5**) à l'abri de la lumière pendant 20 min. Les cellules sont lavées deux fois avec 500 µL de PBS (centrifugation 10 min-1500rpm).

Pour le marquage nucléaire des LCLs et des leucocytes, il est nécessaire de fixer et de perméabiliser les cellules. Pour cela, 2 mL de tampon A du kit Human FoxP3 Buffer (BD) sont ajoutés pour fixer les cellules. Après incubation 10 min à l'abri de la lumière, elles sont lavées avec 2mL de tampon de marquage (PBS, BSA 1%, EDTA 2mM, Sodium Azide 0,05%), centrifugées 10 min à 1500 rpm et le surnageant est jeté. Le culot est vortexé, 500 µL de tampon C du kit sont ajoutés au goutte à goutte puis il est incubé 30 min à l'abri de la lumière afin de perméabiliser les cellules.

Après deux lavages avec 1ml de tampon de marquage (centrifugation 10 min-1500 rpm), les cellules sont incubées 1 heure avec 200 µL d'anticorps primaire (Anticorps anti-ZFP57 D à 40µg/mL) sous agitation. Après deux lavages avec 1ml de tampon de marquage (centrifugation 10min-1500rpm), 200 µL d'anticorps secondaire (Donkey anti-rabbit abcam à 2 µg/mL) dilué dans de tampon de marquage sont ajoutés au culot qui est incubé 30 min. Après deux derniers lavages avec 1 mL de tampon de marquage (centrifugation 10min-1500rpm), et resuspension dans 200 µL de tampon de marquage, on procède à l'acquisition des cellules sur l'appareil de cytométrie BD LSR II.

Les données de fluorescence sont analysées avec le logiciel BD FacsDiva. Chaque population cellulaire a été « gatée » sur la base de la taille et de la granulosité. Les intensités de fluorescence moyenne (MFI) des cellules ont été déterminées après soustraction des intensités détectées en absence d'anticorps primaire. Parmi chaque population, les pourcentages de cellules exprimant ZFP57 au-delà du seuil d'auto-fluorescence ont aussi été déterminés.

4.6. Microscopie confocale

Les LCLs étant des cellules en suspension, 200 µL de poly-lysine sont ajoutés au goutte à goutte sur chaque lame afin de les fixer. Les lames sont ensuite incubées 1h à 37°C, puis le surnageant est retiré et elles sont lavées avec 200µL de PBS. Le surnageant est retiré puis 200µL de cellules sont incubées 1h à 37°C. Après avoir enlevé le surnageant, 200 µL de tampon A du kit Human FoxP3 Buffer (BD) sont ajoutés pour fixer les cellules. Après incubation 10 min à l'abri de la lumière, les cellules sont lavées avec 200 µL de tampon de marquage (PBS, BSA 1%, EDTA 2mM, Sodium Azide 0,05%). Puis, 200 µL de tampon C du kit sont ajoutés au goutte à goutte et les lames sont incubées 30 min à l'abri de la lumière afin de perméabiliser les cellules. Après deux lavages avec 200 µl de tampon de marquage, les cellules sont incubées 1 heure avec 200 µL d'anticorps primaire (Anticorps anti-ZFP57 D à 40 µg/mL). Après deux lavages avec 200 µL de tampon de marquage, 200 µL d'anticorps secondaire (Donkey anti-rabbit abcam à 2 µg/mL) dilués dans le tampon de marquage sont ajoutés et mis à incuber 30 min. Après deux lavages avec 200 µL de tampon de marquage, 200 µL de DAPI (noyau) et phalloïdine (actine) sont ajoutés. Enfin, après deux derniers lavages au PBS les lames sont montées.

4.7. Spectrométrie de masse

Pour identifier les protéines reconnues par l'anticorps D anti-ZFP57, deux fractions nucléaires, l'une soluble et l'autre insoluble, ont été préparées à partir de cellules de la lignée COX. Dans chaque cas, une condition « lysat » et une condition « lysat suivie d'une immunoprécipitation avec l'anticorps » ont été préparées. Ces extraits ont été ensuite migrés en SDS-PAGE et colorés au bleu de Coomassie puis confiés à la plateforme de spectrométrie de masse de l'Institut Jacques Monod. Pour aider au repérage des bandes à analyser, un Western-blot a été effectué en parallèle sur les mêmes gels.

4.7.1. Immunoprécipitation

Cette étape est réalisée à l'aide du « Immunoprécipitation Kit Dynabeads Protein A » (10006D, ThermoFisher Scientific). Les billes magnétiques associées à la protéine A sont préparées (resuspension par vortex 30 sec, lavage par aimentation). Puis elles sont incubées sous rotation durant 10 min à température ambiante avec 10µg d'anticorps pré-dilué dans 200µL de PBS+0,1% Tween20. Après retrait du surnageant, les complexes billes-anticorps sont resuspendus dans 200 µL de PBS-0,1% Tween20 et lavées. Les anticorps sont ensuite liés de façon covalente aux billes à l'aide d'une solution de 250 µl de 5 mM de BS3 (Bis(sulfosuccinimidyl)suberate, ThermoFisher Scientific) dans le tampon Conjugation Buffer (20mM Sodium Phosphate, 0,15M NaCl pH 7-9). Après 30 minutes à température sous rotation, cette étape est stoppée avec une incubation 15 min sous rotation dans 12,5 µL de Quenching Buffer (1M Tris HCl pH7,5). Les complexes «crosslinkés» sont lavés trois fois avec 200µL de PBS-0,1%Tween20. L'échantillon contenant l'antigène d'intérêt est ensuite incubé 10 min à température ambiante sous rotation. Après trois lavages avec 200µL de PBS-0,1% Tween20, les complexes billes-anticorps-antigène sont repris dans 100µL de PBS-0,1% Tween20 et transférés dans un nouveau tube (pour éviter l'élution de protéines adsorbées aux parois du tube). Après retrait du surnageant, 20µL d'Elution Buffer, 2,5µL de LDS Sample Buffer (Invitrogen), 1µL de Sample Reducing Agent (Invitrogen) et 6,5µL d'H_2O sont ajoutés. Le complexe est bien homogénéisé, chauffé 10 min à 70°C, placé sur l'aimant et le surnageant est déposé sur un gel d'acrylamide pré-coulé à gradient 4-12% Bis-Tris (Invitrogen). Le gel est mis à migrer 15 min à 120V puis 1h20 à 160V dans du tampon de migration en présence d'une solution anti-oxydante. Une coloration à l'argent ou au coomassie est ensuite réalisée.

4.7.2. Coloration à l'argent

La coloration à l'argent est réalisée à l'aide du SilverXpress Silver Staining Kit (LC6100, ThermoFisher Scientific) qui a une sensibilité de quelques nanogrammes de protéines. L'ensemble est réalisé sous agitation. Le gel est trempé dans 200 mL de fixateur 10 min (90mL d'H_2O, 100 mL de méthanol, 20 mL d'acide acétique) puis deux fois 30 min dans 100 mL de sensibilisateur (52,5mL d'H_2O, 50mL de méthanol, 2,5 mL de sensibilisant). Il est ensuite lavé deux fois 10 min dans 200mL d'H_2O puis coloré 15 min avec un mélange de 5mL de colorant A (contenant du nitrate d'argent), 5mL de colorant B (contenant de l'ammonium et du sodium hydroxide) complétés avec 90mL d'H_2O. Le gel est lavé deux fois 5 min dans 200mL d'H_2O puis révélé 10 min avec 5 mL de révélateur complété avec 95 mL d'H_2O. La réaction est stoppée en ajoutant 5mL de stoppeur 10 min puis trois derniers lavages de 10 min avec 200 mL d'H_2O sont effectués.

4.7.3. Coloration au bleu de Coomassie

La coloration au bleu de Coomassie est réalisée à l'aide du SimplyBlue SafeStain (LC6060, Thermo Fisher Scientific) qui a une sensibilité de 500 ng de protéines. L'ensemble est réalisé sous agitation. Le gel est rincé 3 fois 5 min dans 100 mL d'eau déminéralisée afin d'éliminer tous les sels et le SDS qui peuvent interférer avec le colorant. Chaque gel est mis à incuber avec le colorant 1h à température ambiante. Les bandes commencent à apparaitre au bout de quelques minutes.
Le gel est ensuite lavé avec 100mL d'H_2O entre 1-3h. Pour éliminer le maximum de bruit de fond, il est nécessaire de réaliser un second lavage 1h avec 100mL d'H_2O.

4.7.4. Spectrométrie de masse en tandem

Après découpage des bandes de gel, les échantillons ont été digérés en présence de trypsine couplée à des billes et analysés sur un spectromètre de masse en tandem (nanoESI-Orbitrap) à la plateforme de Protéomique Structurale et Fonctionnelle de l'Institut Jacques Monod (C GARCIA, directeur : JM Camadro). L'interprétation des résultats a été réalisée au moyen du logiciel Mascott.

5. Analyse statistique des données

5.1. Contrôle de qualité des données génotypiques

Pour chaque polymorphisme, nous avons vérifié que la distribution génotypique respectait l'équilibre de Hardy-Weinberg qui repose sur le caractère aléatoire de la transmission allélique dans une population « panmictique ». L'écart des proportions génotypiques prédit par la loi de Hardy-Weinberg a été testé par un χ^2 de conformité.
Nous avons détecté les erreurs de transmission mendélienne au sein des pedigrees avec PEDSTATS v0.6.12 (Wigginton et Abecasis 2005). Ces erreurs ont été corrigées en annulant les génotypes concernés à l'aide d'un script python ou en retirant la totalité du pedigree lorsque deux erreurs ou plus étaient identifiées.

5.2. Estimation du déséquilibre de liaison

Le déséquilibre de liaison allélique a été évalué avec le coefficient r^2, calculé soit dans l'environnement R, soit avec le logiciel PLINK v1.07 (http://pngu.mgh.harvard.edu/purcell/plink/) (Purcell et al. 2007).

5.3. Etudes d'association familiales

Les études familiales d'association ont été réalisées avec le Test de Déséquilibre de Transmission (TDT) qui évalue la proportion d'allèles transmis d'un parent hétérozygote à un enfant atteint, la proportion attendue dans l'hypothèse nulle de non association étant de 50%.
Ce test a été réalisé à l'aide de PLINK avec l'option –tdt. Les TDT avec biais de transmission parental ont été effectués avec PLINK v1.07 avec l'option –poo (parent of origin). Les résultats d'association rapportés sont donnés en référence à l'allèle mineur chez les fondateurs.

La visualisation de l'association est représentée sous forme d'un manhatttan plot avec le package « qqman » sous R (https://github.com/stephenturner/qqman). Le zoom sur les régions d'intérêt se fait grâce au logiciel LocusZoom (Pruim et al. 2010).
Les études d'association conditionnées sur les allèles de risque ont été testées avec le logiciel UNPHASED v3.1.7 (Dudbridge 2008) qui infère les haplotypes par maximum de vraisemblance, tout en autorisant certaines phases incertaines et génotypes manquants. Les options utilisées sont –condition sur le marqueur « conditionnel » et –condspecific sur l'allèle associé de ce marqueur.

5.4. Cartographie d'expression (eQTL) de *ZFP57*

La cartographie d'association du trait quantitatif de l'expression de *ZFP57* a été réalisée à l'aide du logiciel PLINK v1.07 par régression linéaire avec l'option –linear selon un modèle additif. Les données sont rapportées pour l'effet de l'allèle mineur.

5.5. Analyses statistiques des données d'expression génique (qPCR et ELISAs)

Les étapes de prétraitements des données, les contrôles de qualité et la totalité des analyses statistiques ont été effectuées avec l'environnement R. La normalité des traits quantitatifs a été testée par un test de Shapiro. L'homoscédasticité (homogénéité des variances) selon les niveaux des facteurs testés a été vérifiée par un test de Bartlett.
Selon les conditions d'application, des tests paramétriques ou non paramétriques ont été menés pour comparer les niveaux d'expression entre deux populations (test t de Welch ou test de Wilcoxon) ou pour corréler les variables quantitatives entre elles (test de chi² de Pearson ou de Spearman). Différents modèles de régression linéaire univariés et multivariées ont été testés (modèles multivariés spécifiés dans les tableaux de résultats) et les diagnostics appropriés effectués, dont la vérification de la normalité des résidus. Les analyses de corrélations partielles ont été effectuées avec le package « ppcor » sous R (Kim 2015).

Résultats

1. Etude génétique : « French-GWAS »

Si de nombreuses études d'association génétiques au DT1 ont été menées dans différentes cohortes internationales, aucune à ce jour n'avait inclus de patients français. L'étude « French-GWAS » menée par notre laboratoire est la première étude réalisée sur des familles françaises grâce au recrutement d'une cohorte familiale de patients de DT1 et de leurs apparentés au 1er degré en collaboration avec les Hospices civils de Lyon. Elle se décompose en une cohorte de « découverte », appelée Lyon 1 et une cohorte de « réplication », Lyon 2. Cette première partie de mon projet de recherche visait à répliquer dans la cohorte Lyon 2 les résultats d'association génétique détectés dans la cohorte de « découverte » Lyon 1.

1.1. Données préliminaires : étude d'association pangénomique dans la cohorte de « découverte » Lyon 1

La cohorte de « découverte » Lyon 1 comprend 685 sujets, 241 patients de DT1 et 444 contrôles apparentés (parents et germains) au sein de 237 familles nucléaires. Elle a été génotypée à l'aide de l'immunochip, une puce incluant 196 524 SNPs sur l'ensemble du génome, ciblant particulièrement la région du CMH mais également des régions associées à différentes maladies auto-immunes et inflammatoires (Trynka et al. 2011 ; Parkes et al. 2013). L'ensemble de ces génotypages et des analyses a été réalisé avant mon arrivée au laboratoire. Une analyse d'association familiale par test du déséquilibre de transmission (TDT) avait été conduite. Ce test d'association en présence de liaison génétique compare les fréquences avec lesquelles un allèle particulier est transmis ou non transmis par les parents hétérozygotes à leur enfant atteint. La signification de l'association est testée sur l'ensemble des familles nucléaires par un test de comparaison de fréquences (Chi2). Si l'allèle est associé à la maladie, un excès de transmission de cet allèle sera observé dans plus de la moitié des cas. Le seuil statistique de la p-value permettant de déterminer si l'association d'un SNP avec le DT1 est significative est de 5.10^{-8} pour une association pangénomique. Comme attendu, après analyse des résultats, le locus principal qui présentait le plus grand nombre d'associations avec le DT1 était le CMH avec l'effet majeur du locus *HLA-DRB1-DQB1* au niveau des deux tag-SNPs : rs2187668 pour DR3 et rs7454108 pour DR4-DQ8 (**Figure 22 et Tableau 6**).

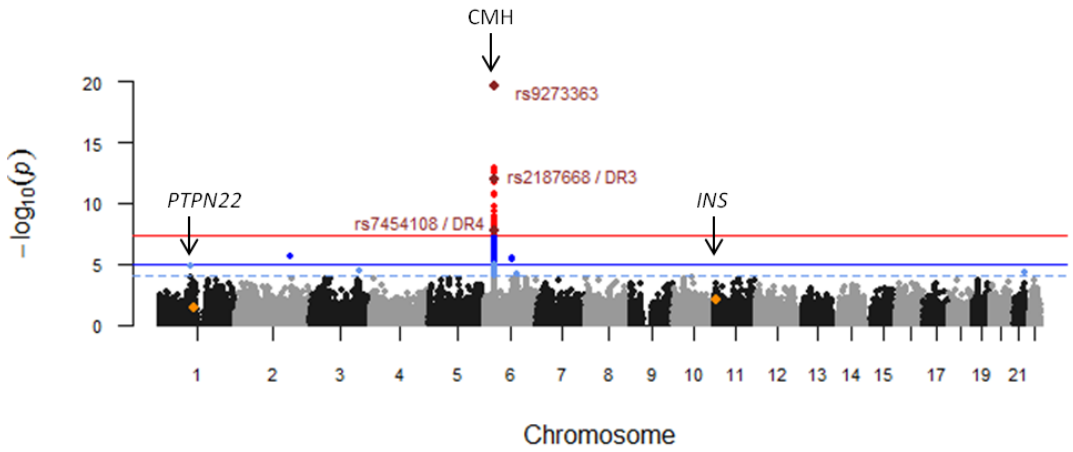


Figure 22. Manhattan plot de l'étude d'association au DT1 des SNPs de l'immunochip dans la cohorte Lyon 1. Représentation du niveau d'association (-log10 de la valeur p du test de chaque SNP) en ordonnée pour chaque SNP génotypé dans la cohorte Lyon 1 selon leur position le long du génome en abscisse, avec une alternance de points noirs ou gris pour les SNPs des chromosomes impairs et pairs. La ligne rouge indique le seuil de signification statistique pangénomique 5.10^{-8}, celle en bleu foncé le seuil d'association suggestif 10^{-5} et celle en bleu clair pointillé celui à 10^{-4}. Les SNPs présentant une valeur p en dessous de ces seuils sont colorés selon ces trois seuils. Les trois SNPs du CMH significatifs au seuil pangénomique 5.10^{-8} discutés sont représentés par des losanges rouges foncés. Les SNPs de l'*INS* et de *PTPN22* associés dans les autres cohortes, sont ici non significatifs et sont indiqués par des losanges orange sur les chromosomes 11 et 2 respectivement.

De plus, un SNP du CMH, rs9273363, est retrouvé avec un effet plus important. Il n'avait été rapporté associé au DT1 à ce jour que dans la publication du Wellcome Trust Case Control Consortium (Wellcome Trust Case Control Consortium et al. 2010). Il s'agit donc de la première réplication de l'association principale dans une cohorte Française. Cependant, aucune autre région connue comme étant associée au DT1 n'a été retrouvée comme associée de manière significative dans cette cohorte Lyon 1. Un problème de puissance dû à l'effectif limité de la cohorte conjugué aux amplitudes modestes des effets connus pour des variants fréquents peut l'expliquer (**Figure 23**).

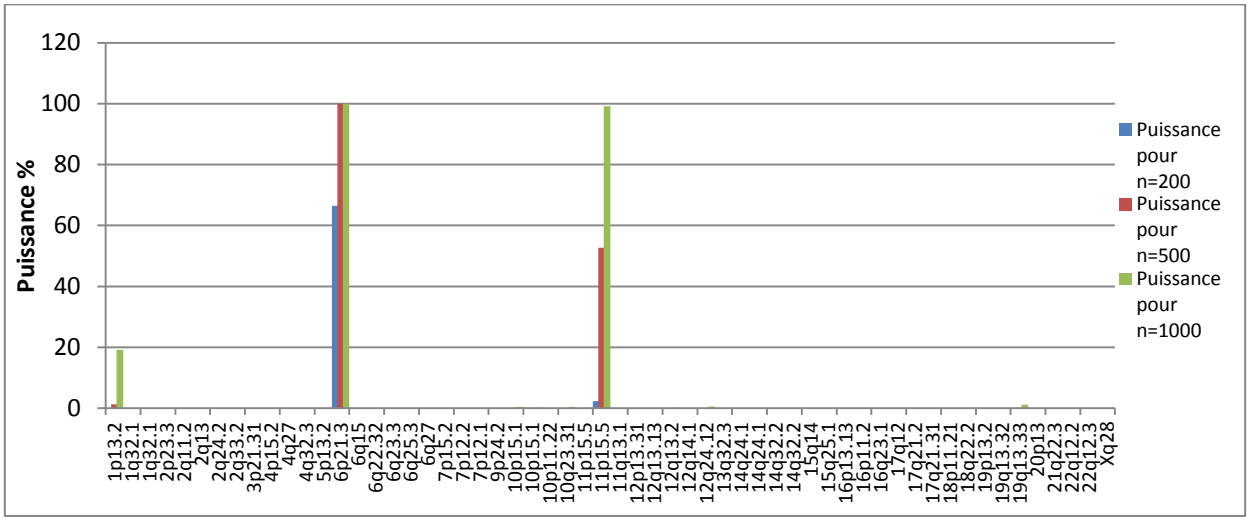


Figure 23. Puissance de détection des 58 régions déjà impliquées dans le DT1 dans des études d'association allélique selon la taille de l'échantillon. Les puissances ont été calculées pour des échantillons incluant 200, 500 ou 1000 cas et autant de contrôles. Plus le nombre de sujets inclus au sein de la cohorte augmente, plus la puissance de détecter une association augmente. On peut constater que seules les régions du CMH (6p21.3) et de l'*INS* (11p15.5) ont 100% et 99,2% de chances d'être détectées avec une cohorte de 1000 cas/contrôles. Il est intéressant de noter que *PTPN22* (1p13.2) a seulement une puissance de 19,2% pour n=1000.

Tableau 6. Résultats de l'étude d'association au DT1 dans la cohorte de Lyon. Les résultats sont donnés pour chaque SNP pour lequel une association a été trouvée dans la cohorte de « découverte » Lyon 1, la cohorte de « réplication » Lyon 2 ou les cohortes combinées. Pour chaque SNP, la bande cytogénétique, la position chromosomique (hg19), le nom du gène associé et les allèles mineur et majeur sont indiqués. Pour les cohortes Lyon 1 et Lyon 2 sont renseignés la fréquence de l'allèle mineur (MAF) chez les fondateurs, le nombre de transmissions et de non transmission (T:U) de l'allèle mineur, la p-value et l'odds ratio correspondant donné pour l'allèle mineur avec son intervalle de confiance à 95%. Pour la cohorte combinée, la p-value et l'odds ratio sont indiqués. Chr : Chromosomique, T:U : Transmitted:Untransmitted (Transmis:Non transmis), MAF : Minor Allele Frequency (Fréquence de l'allèle mineur) chez les fondateurs, OR : Odds Ratio, IC : Intervalle de Confiance.

						Cohorte Lyon 1				**Cohorte Lyon 2**				**Cohortes Lyon 1 et 2**	
SNP	**Bande cytogénétique**	**Position chr (hg19)**	**Gène(s)**	**Allèle mineur**	**Allèle majeur**	**MAF**	**T:U**	**p-value**	**OR [IC 95%]**	**MAF**	**T:U**	**p-value**	**OR [IC 95%]**	**p-value**	**OR [IC 95%]**
rs9273363	6p21.3	32626272	CMH	T	G	0,45	108:12	1,89E-18	9,00 [4,96;16,34]	0,48	289:43	1,54E-41	6,72 [4,88;9,26]	8,54E-59	7,18 [5,43;9,49]
rs2187668 (DR3)	6p21.3	32605884	CMH	T	C	0,243	70:10	1,97E-11	7,00 [3,61;13,58]	0,23	198:50	5,56E-21	3,96 [2,90;5,40]	4,24E-31	4,52 [3,42;5,97]
rs7454108 (DR4)	6p21.3	32681483	CMH	G	A	0,175	45:6	4,73E-8	7,50 [3,20;17,58]	0,20	164:34	2,50E-20	4,82 [3,33;6,98]	9,24E-27	5,15 [3,68;7,19]
rs689 Immunochip	11p15.5	2182224	*INS*	A	T	0,041	3:13	1,24E-2	0,23 [0,067;0,81]						
Taqman						0.219	16:37	3,92E-3	0,43 [0,24;0,78]	0,24	64:131	1,60E-6	0,49 [0,36;0,66]	2,3E-8	0,48 [0,36;0,62]
rs17638639	2q31.2	179026118	*OSBPL6, RBM45, PDE11A*	G	C	0,141	10:42	9,10E-6	0,24 [0,12;0,48]	0,10	41:67	1,24E-2	0,61 [0,42;0,90]	3,32E-6	0,46 [0,33;0,46]
rs2476601	1p13.2	114377568	*PTPN22*	A	G	0,084	22:10	3,39E-2	2,20 [1,04;4,65]	0,09	67:35	1,53E-3	1,91 [1,27;2,88]	1,08E-4	2,00 [1,40;2,86]
rs7739286	6q21	106433575	*Promoteur de PRMD1*	C	A	0,220	17:49	8,19E-5	0,35 [0,20;0,60]	0,27	110:118	5,96E-1	0,93 [0,72;1,21]	1,74E-2	0,78 [0,60;0,95]
rs1604643	3q25	158653634	*MFSD1, IQJC*	C	T	0,150	12:42	4,46E-5	0,29 [0,15;0,54]	0,14	75:69	6,17E-1	1,09 [0,78;1,51]	8,97E-2	0,79 [0,59;1,04]
rs2838514	21q22.3	45608337	*ICOSL*	C	T	0,291	21:56	6,65E-5	0,38 [0,22;0,62]	0,27	123:115	6,04E-1	1,07 [0,83;1,38]	1,45E-1	0,85 [0,68;1,06]
rs62408234	6q15	90978160	*BACH2*	T	C	0,206	17:55	7,52E-6	0,31 [0,18;0,53]	0,17	94:80	2,89E-1	1,18 [0,87;1,58]	1,64E-1	0,84 [0,65;1,08]
rs62408223	6q15	90949195	*BACH2*	G	A	0,210	18:56	9,99E-6	0,32 [0,19;0,55]	0,17	95:81	2,91E-1	1,17 [0,87;1,58]	1,68E-1	0,84 [0,66;1,08]
rs75460852	1p21.2	101539691	*DHP5*	A	T	0,045	1:20	3,38E-5	0,05 [0,01;0,37]	0,06	39:26	1,07E-1	1,50 [0,91;2,46]	4,19E-1	0,87 [0,57;1,33]

Au seuil de suggestivité classique de 1.10^{-5}, on trouve une association suggestive de 40 SNPs. La majorité (37 SNPs) demeure dans le CMH. Ces niveaux d'associations sont dus au déséquilibre de liaison avec les allèles de risque majeur, DR3 et DR4-DQ8. Les trois autres SNPs sont en dehors du CMH : rs17638639 en 2q31.2, et rs62408234 et rs62408223 en 6q15.

Le premier SNP, rs17638639, n'est pas localisé dans une région déjà associée au DT1 ni à aucune autre maladie auto-immune ou inflammatoire (d'après ImmunoBase (immunobase.org) et le GWAS catalog (ebi.ac.uk/gwas/). Seuls 19 autres SNPs ont été génotypés dans une fenêtre de 500 kb autour de ce SNP, aucun n'étant en déséquilibre de liaison avec ce SNP (**Figure 24**). Il convenait donc de répliquer cette nouvelle association suggestive dans une cohorte de « réplication ».

Les deux autres SNPs associés de manière suggestive, rs62408234 et rs62408223, n'ont jamais été associés au DT1, mais ils sont situés au centre d'une région (chr6 : 90806835-91046297) déjà identifiée comme associée au DT1 (**Tableau 4 et Figure 4**) et contenant le gène candidat *BACH2* (Barrett et al. 2009 ; Cooper et al. 2008). Ils sont en déséquilibre de liaison presque parfait entre eux (r^2 = 0,987). En revanche, le SNP rs72928038 qui est le variant connu le mieux associé dans cette région n'est pas du tout associé dans notre étude (p = 0,1235) et n'est pas en déséquilibre de liaison avec rs62408223 et rs62408234 (r^2 =0,018 pour les deux).

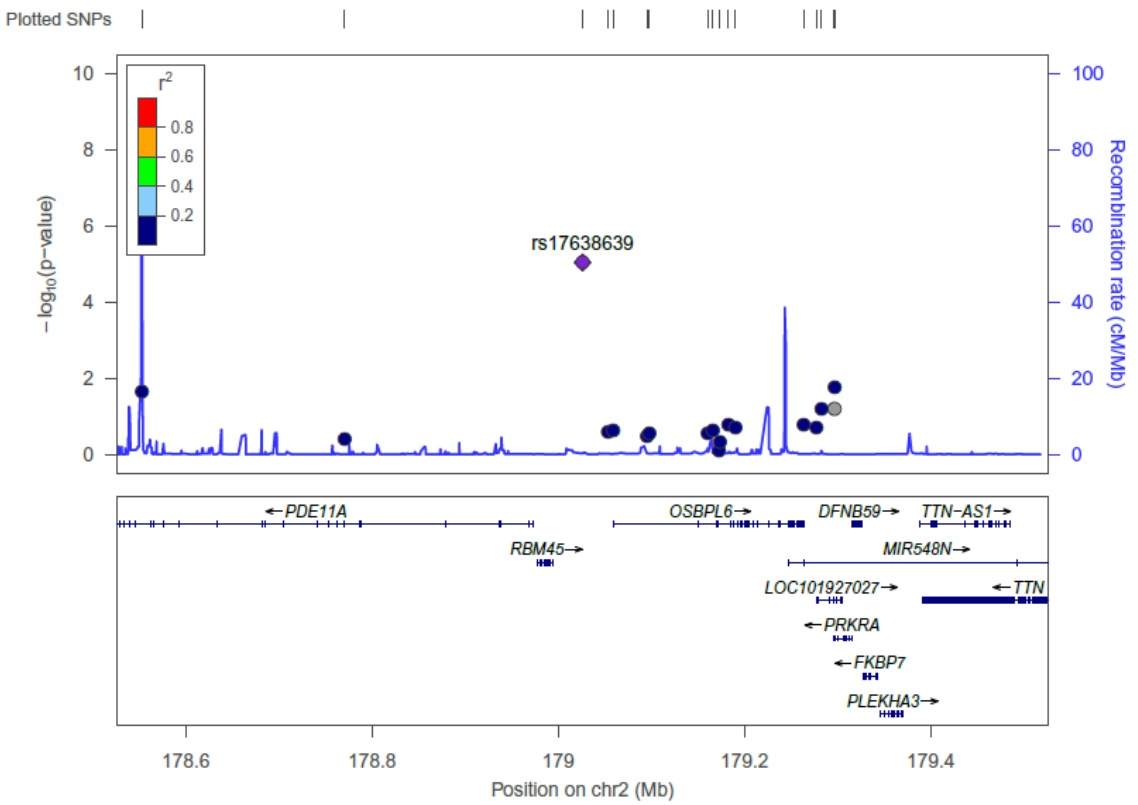


Figure 24. Zoom sur la région 2q31.2 du SNP rs17638639. La couverture en SNPs génotypés de la région est représentée en haut de la figure (traits fins verticaux). L'encadré au centre représente le niveau d'association pour chaque SNP de la région. La couleur de chaque SNP indique le niveau de déséquilibre de liaison (r^2) avec le SNP le mieux associé, rs17638639 (losange violet). La ligne bleue plus claire indique le taux de recombinaison. L'encadré du bas représente les gènes de la région selon leur orientation.

Il n'y a pas d'association pour les autres régions déjà associées au DT1, en particulier pour le premier locus après le CMH, l'*INS* (p = 0,012), ni pour le deuxième locus, *PTPN22* (p = 0,039). De ce fait, nous avons également considéré un seuil moins strict d'association suggestive à 1.10^{-4}.
Au total, 78 SNPs sortent sous ce seuil de 10^{-4} dont 74 sont présents dans le CMH tandis que 4 SNPs sont nouveaux : rs75460852 en 1p21.2, rs1604643 en 3q25, rs2838514 en 21q22.3 et rs7739286 en 6q21. Il s'agit de 4 régions encore inconnues dans le DT1.
Le premier SNP, rs75460852, se situe dans une région associée à la sclérose en plaques mais l'association n'est pas connue avec ce SNP. La région du deuxième SNP, rs1604643, n'est associée à aucune maladie auto-immune ou inflammatoire connues à ce jour. Le troisième SNP, rs2838514, est situé dans une région associée à plusieurs maladies auto-immunes ou inflammatoires (maladie cœliaque, maladie de Crohn, spondylarthrite ankylosante, immunobase.org) mais pas directement avec ce SNP. Il est intéressant de noter que dans cette région se trouve *ICOSL* (21q22.3) et que le gène codant le récepteur de ce gène, appelé *ICOS*, est situé dans une région qui est associée au DT1 même si le gène candidat le plus probable dans la région est *CTLA4* (**Tableau 4**). Le quatrième SNP, rs7739286, est situé dans une région associée à la maladie de Crohn mais pas directement avec ce SNP. On recense 11 autres SNPs avec une p-value inférieure à 10^{-3} dans une fenêtre de 28 kb autour de ce SNP. Il est donc peu probable qu'il s'agisse d'un artefact de génotypage.

A l'issu de ce travail, il était donc nécessaire de répliquer les associations suggestives et de clarifier les contributions relatives des régions habituellement les mieux associées (CMH, *INS*, *PTPN22*) dans cette cohorte, ce à quoi je me suis attelée.

1.2. Ré-évaluation de l'association de la région de l'insuline dans la cohorte Lyon 1

Nous avons souhaité comprendre l'absence d'association du polymorphisme de l'insuline puisque les résultats de la puce n'ont montré aucune association significative au seuil pangénomique de 5.10^{-8} ni même suggestive aux seuils de 10^{-5} ou 10^{-4}. Cette absence peut être due à un problème de puissance puisque celle-ci est comprise entre 2,3% et 11,6% pour un effectif de cas et de contrôles respectif de 200 et 300. Bien que ce SNP, rs689, fût en équilibre de Hardy-Weinberg (p= 0,0792) et ne laissât présager aucun problème de génotypage, nous avons vérifié le clustering de ce SNP obtenu par génotypage de l'immunochip. Il en résulte que ce SNP souffrait d'un mauvais clustering (**Figure 25**) qui n'avait pourtant pas conduit à une distorsion des fréquences attendues sous l'équilibre de Hardy-Weinberg (p= 0,0792) et avait permis à ce SNP de franchir les critères de contrôle de qualité habituels.
Un nouveau génotypage du SNP rs689 à l'aide d'une sonde TaqMan s'est donc avéré nécessaire pour la cohorte de « découverte » Lyon 1. Les résultats de l'étude d'association familiale pour ce nouveau génotypage sont également présentés dans le **Tableau 6**. La fréquence allélique de l'allèle mineur chez les fondateurs est désormais conforme à celle attendue d'environ 20 à 30%. Cependant, l'association allélique détectée n'est pas significative au seuil pangénomique avec une valeur p = $3,92.10^{-3}$, possiblement par manque de puissance. Il convenait donc de tester aussi l'association de ce polymorphisme dans la cohorte Lyon 2.

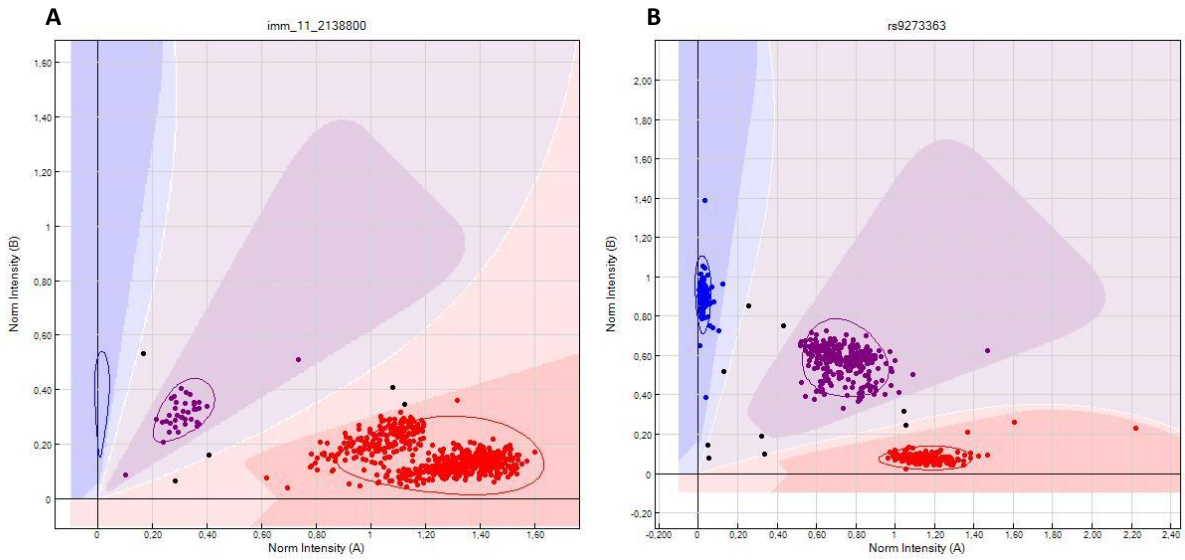


Figure 25. Données de génotypage de l'immunochip de la cohorte Lyon 1 des SNPs rs689 (alias imm_11_213880) (A) et rs9273363 (B). Chaque point correspond au génotype d'un individu. Les points rouges correspondent aux sujets homozygotes pour l'allèle A, les points bleus aux sujets homozygotes pour l'allèle B et les points violets aux sujets hétérozygotes pour les allèles A et B selon la fluorescence détectée au niveau des sondes des allèles A et B. Une séparation nette entre les nuages de points de couleur différente est représentative d'un bon génotypage. On peut voir dans le panneau A pour le SNP rs689 que les génotypes des sujets n'ont pas été correctement attribués : les hétérozygotes ont été génotypés homozygotes pour l'allèle A et les homozygotes pour l'allèle B ont été génotypés hétérozygotes. Ce génotypage est donc ininterprétable. En contre-exemple, le génotypage du SNP rs9273363, figure B, donne un bon clustering puisque les nuages de points sont bien séparés.

1.3. Etude d'association dans la cohorte de « réplication » Lyon 2

Afin de confirmer les résultats obtenus dans la cohorte de « découverte », 12 SNPs ont été génotypés avec des sondes TaqMan dans la cohorte de « réplication » Lyon 2 qui ne pouvait être génotypée par l'immunochip pour des raisons de coût. En effet, la cohorte de « réplication » Lyon 2 inclut près de trois fois plus de patients que la cohorte Lyon 1 avec 2779 sujets au sein de 825 familles nucléaires incluant 710 patients diabétiques et 2061 contrôles apparentés (parents et germains) dont 8 de phénotype inconnu.

Les résultats au seuil pangénomique de 5.10^{-8} répliquent ceux obtenus pour la cohorte Lyon 1 et montrent une association majeure significative des marqueurs testés du CMH, le « top » SNP étant une nouvelle fois rs9273363, suivi du SNP taggant DR3 puis de celui taggant DR4-DQ8 (**Tableau 6**). La taille de l'échantillon étant plus grande que celle de la cohorte Lyon 1, la puissance de détecter une association était plus élevée. On peut constater que les résultats sont plus significatifs et les estimations par intervalle de confiance sont plus précises. Les mêmes allèles sont associés à un risque. Aucune autre association significative n'a été trouvée au seuil pangénomique.

L'étude de la réplication sur les 12 SNPs génotypés porte le seuil de correction multiple à $4,17.10^{-3}$. A ce seuil, on observe une association significative de l'*INS* et de *PTPN22* avec les mêmes allèles de risque que ceux décrits dans les associations déjà connues des autres cohortes internationales.

Au seuil alpha de 5%, le SNP rs17638639 en 2q31.2 est répliqué avec le même allèle de risque que dans la cohorte Lyon 1, l'allèle majeur C. L'amplitude de l'effet de l'allèle de risque donné par son Odds Ratio (OR) égal à 1,63 (1/0,6119) est inférieure à celle de l'*INS* (OR=2,047 pour l'allèle de risque) et à celle de *PTPN22* (OR=1,914) mais elles sont inclues dans son IC à 95% = [1,108 ; 2,410]. De plus cet intervalle de confiance (IC) à 95% chevauche celui de l'étude de la cohorte Lyon 1 suggérant la possibilité d'un effet similaire entre les deux cohortes.

En revanche, les SNPs de *BACH2* ne sont pas du tout associés dans la cohorte Lyon 2. Les 4 autres SNPs suggestifs au seuil 10^{-4} dans la cohorte Lyon 1 ne sont pas non plus associés.

1.4. Analyse d'association combinée des cohortes Lyon 1 et Lyon 2

Une étude combinant les deux cohortes ensemble a été réalisée, portant ainsi la taille de l'échantillon à 3464 sujets, 951 patients de DT1, 2505 contrôles apparentés (parents et germains), 8 sujets apparentés sans phénotype connu, au sein de 1062 familles nucléaires.
Les résultats du Test de Déséquilibre de Transmission (TDT) (**Tableau 6**) indiquent une association significative au seuil pangénomique 5.10^{-8} des SNPs du CMH avec une précision des estimations des OR : 4,517 avec un IC95% [3,415 ; 5,974] pour rs2187668 (DR3), 5,146 avec un IC95% [3,683 ; 7,191] pour rs7454108 (DR4) et 7,179 avec un IC95% [5,428 ; 9,494] pour rs9273363. De plus, rs689 du gène *INS* est également associé significativement au seuil 5.10^{-8} ($p=2,3.10^{-8}$) et devient le second locus de risque dans cette cohorte avec un OR de 2,1 avec un IC95% de [1,61 ; 2,74] pour l'allèle de risque qui est l'allèle majeur.
Surtout, nous rapportons pour la première fois une nouvelle association suggestive au seuil 10^{-5} pour rs17638639 ($p=3,32.10^{-6}$) qui avait été détectée dans la cohorte Lyon 1 et répliquée dans celle de Lyon 2. Cette association a une forte amplitude pour une maladie multifactorielle puisque l'OR de l'allèle majeur atteint 2,157 avec un IC à 95% de [1,55 ; 3,01].
Le SNP de *PTPN22* occupe donc le 4ème rang mais n'atteint pas le seuil de suggestivité de 10^{-5} ($p=1,075.10^{-4}$) dans cette double cohorte.

1.5. Etude d'association avec biais de transmission parental

Nous avons complété l'analyse par une étude du biais de transmission parental, selon que les allèles morbides sont transmis préférentiellement par le père ou par la mère. A ce jour, un seul SNP était connu pour présenter un biais d'origine paternelle dans la cohorte du T1DGC, rs941576 situé en 14q32.2 dans la région soumise à empreinte parentale *DLK1-MEG3* (Wallace et al. 2010).
Avant mon arrivée au laboratoire, ce SNP avait été testé au moyen de l'immunochip sur la cohorte Lyon 1 et ne montrait ni association ni biais de transmission parental. En revanche, un SNP voisin, le rs11160607 présentait un biais de transmission paternel. Cependant, nous n'avons répliqué aucune de ces deux associations dans la cohorte Lyon 2 (**Tableau 7**). Il faut noter que trois SNPs présentent un biais significatif au seuil de 5%, dans l'une ou l'autre ou la combinaison des deux cohortes. Il s'agit du SNP rs17638639 nouvellement associé sur le chromosome 2 avec un biais de transmission maternel dans la cohorte Lyon 1, du SNP de *PTPN22* avec un biais de transmission paternel dans la cohorte Lyon 2, et du SNP tagguant DR4 rs7454108 avec un biais de transmission paternel dans les cohortes combinées. Ces résultats doivent toutefois être pris avec précaution et nécessitent une réplication indépendante.

Tableau 7. Résultats de l'étude du biais de transmission parental dans la cohorte de Lyon. Les résultats sont donnés pour chaque SNP pour lequel une association a été trouvée, rs941576 (Wallace et al. 2010), rs11160607 et rs17638639 dans la cohorte Lyon 1, rs2476601 dans la cohorte Lyon 2 et rs7454108 dans les cohortes combinées.
Pour chaque SNP, la bande cytogénétique, les allèles mineur et majeur, le nombre d'allèle mineur paternel (pat) ou maternel (mat) transmis et non transmis, la p-value et l'odds ratio correspondant sont indiqués. Chr : Chromosomique, T:U : Transmitted:Untransmitted (Transmis:Non transmis). La valeur « p-value pool » indique s'il existe une différence significative des transmissions selon qu'elles sont d'origine maternelle ou paternelle. Les valeurs significatives sont indiquées en gras.

				Analyse TDT avec biais de transmission parental Cohorte Lyon 1					**Analyse TDT avec biais de transmission parental Cohorte Lyon 2**					**Analyse TDT avec biais de transmission parental Cohortes Lyon 1 et 2**				
SNP	**Bande cytogénétique**	**Allèle mineur**	**Allèle majeur**	**T:U pat**	**p-value pat**	**T:U mat**	**p-value mat**	**p-value pool**	**T:U pat**	**p-value pat**	**T:U mat**	**p-value mat**	**p-value pool**	**T:U pat**	**p-value pat**	**T:U mat**	**p-value mat**	**p-value pool**
rs941576	14q32.2	C	T	16:22	0,33	24:19	0,45	0,22	73:83	0,42	79:75	0,75	0,43	90:107	0,23	104:95	0,524	0,191
rs11160607	14q32.2	A	G	10:22	**0,03**	19:09	0,06	**0,01**	55:45	0,32	54:43	0,26	0,93	67:68	0,93	73:53	0,075	0,179
rs17638639	2q31.2	G	C	08:17	0,07	02:25	**9,58E-06**	**0,04**	21:34	0,08	20:33	0,07	0,96	29:51	**0,01**	22:59	**3,938E-05**	0,216
rs2476601	1p13.2	A	G	08:04	0,25	14:06	0,07	0,84	40.5:13.5	**2,37E-04**	26.5:21.5	0,47	**0,04**	49.5:17.5	**9,25E-05**	40.5:27.5	0,115	0,079
rs7454108 (DR4)	6p21.3	G	A	29.5:1.5	**4,93E-07**	15.5:4.5	**0,01**	0,08	83.5:13.5	**1,18E-12**	80.5:20.5	**2,37E-09**	0,24	114:15	**2,87E-18**	97:26	**1,535E-10**	**0,044**

1.6. Discussion et perspectives

Réplication dans une cohorte française de l'association connue du CMH avec les facteurs de risque majeur DR3 et DR4

Comme attendu, l'association la plus forte détectée porte sur le CMH et notamment le locus *HLA-DRB1-DQB1*. Compte tenu de la taille des échantillons et des associations déjà rapportées pour cette région, nous avions une puissance de détecter cette association supérieure à 66% pour la cohorte Lyon 1 et de 100% pour la cohorte Lyon 2. Cependant, le SNP le mieux associé, rs9273363, n'est pas le « tagging » SNP de DR3 et DR4. Il n'est d'ailleurs qu'en déséquilibre de liaison partiel avec DR3 et DR4 (r^2=0,336 et r^2=0,274 respectivement) et il demeure associé de manière suggestive ($p=1,29.10^{-3}$) après analyse d'association conditionnée sur DR3 et DR4.

Il reste à comprendre pourquoi il n'a été précédemment associé au DT1 que dans la cohorte du Wellcome Trust Case Control Consortium (WTCCC) (Nejentsev et al. 2007). Dans la base de données « ImmunoBase » (www.immunobase.org), les données d'associations de ce SNP sont rapportées pour 6 maladies auto-immunes et inflammatoires dont trois de manière significative ($p=1.10^{-323}$ pour la maladie cœliaque, $p=2.10^{-46}$ pour la polyarthrite rhumatoïde et $p=1,9.10^{-23}$ pour la thyroïdite auto-immune). En revanche, aucune donnée n'est rapportée pour ce SNP dans le DT1 dans cette base de données, ni dans la base de données « sœur » dédiée exclusivement au DT1, T1Dbase (t1dbase.org). Ceci pourrait être dû soit à un problème de qualité de génotypage dans la cohorte du T1DGC, soit à une élimination de ce SNP lors d'une étape de contrôle de qualité de l'analyse, parmi lesquels un taux de génotypage insuffisant ou trop différent entre cas et contrôles, ou des proportions alléliques et génotypiques observées en non adéquation avec les proportions à l'équilibre de Hardy-Weinberg. Dans la cohorte française, le génotypage du SNP au moyen de l'immunochip est correct avec trois nuages de points bien distincts pour l'intensité de fluorescence (**Figure 25B**) comparable aux « clustering » rapportés pour ce SNP par ImmunoBase et T1DBase sur des sujets contrôles. De plus, son taux de génotypage est satisfaisant (99,65%) mais il est à la limite de l'équilibre de Hardy-Weinberg toléré avec une p-value=$7,5.10^{-4}$ pour la cohorte Lyon 1 et $6,8.10^{-6}$ pour celle de Lyon 2.

En fait, *HLA-DRB1* n'a pas été directement génotypé. Ce sont des SNPs étiquettes de DR3 et de DR4 en déséquilibre de liaison avec ces allèles de *HLA-DRB1* qui ont été génotypés. Cette méthode de typage a déjà été évaluée au laboratoire sur les cohortes internationales du T1DGC avec une très forte sensibilité (corrélation de Pearson, r^2=0,9955 pour DR3 et r^2=0,9923 pour DR4 dans la cohorte EUR) sans toutefois être parfaite. Le génotypage HLA par PCR-SSCP (Réaction d'amplification en chaine-Polymorphisme de conformation des simples brins) est coûteux en ADN et nous ne pourrons pas le réaliser. Dans nos résultats, on remarque dans la cohorte Lyon 1 qu'en plus du SNP rs9273263, d'autres SNPs sont mieux associés que les tagging-SNPs de DR3 et DR4 qui ne sont qu'aux rangs respectivement 8 et 29. Il est possible que la combinaison de leurs génotypes permette de mieux prédire le typage du gène *HLA-DRB1*. Il a notamment été décrit un algorithme de typage de DR3 et DR4 au moyen de deux SNPs supplémentaires en complément du rs9273363 (Nguyen et al. 2013) : rs3104413 et rs2854275 qui ont obtenu respectivement les rangs 26 et 4 dans notre étude d'association dans la cohorte Lyon 1. D'autres algorithmes plus sophistiqués reposant sur le déséquilibre de liaison existant entre les allèles HLA et les allèles de nombreux SNPs de la région du CMH permettent de tagger l'ensemble des allèles HLA tels que

HLA*IMP:02 (Dilthey et al. 2013) ou SNP2HLA (Jia et al. 2013). Grâce au génotypage du CMH à haute densité effectué avec l'immunochip (près de 10 000 SNPs dans la région), nous envisageons d'utiliser ce dernier algorithme. En plus du typage HLA au niveau nucléotidique, ce logiciel permet de déterminer la séquence en acide aminé des molécules HLA, qui peuvent à leur tour être traitées comme polymorphismes multi-alléliques et testées par association génétique. Ce type d'approche a été réalisé pour la sclérose en plaques, le DT1, la polyarthrite rhumatoïde, le psoriasis, la maladie cœliaque et l'achalasie idiopathique (Hu et al. 2015 ; Lenz et al. 2015) et permet de préciser les effets prédisposants ou protecteurs des différentes molécules HLA selon des modèles additifs ou d'interaction. Nous espérons ainsi pouvoir mieux comprendre l'association majeure du locus *HLA-DRB1-DQB1* dans cette cohorte française.

Réplication dans la cohorte française des deux facteurs de risques les plus importants recensés en dehors du CMH : insuline et PTPN22

Après le CMH, la seconde région que nous pouvions détecter en termes de puissance était celle du gène codant l'insuline. Pour la cohorte Lyon 1, la puissance de détection était de seulement 2,3% et nous n'avons effectivement pas réussi à détecter cette association au seuil pangénomique. La puissance dans la cohorte Lyon 2 augmentait à 52,7% et à 99,2% dans les cohortes combinées. Effectivement, l'association est plus significative dans la cohorte Lyon 2 que la cohorte Lyon 1 et elle atteint le seuil de significativité pangénomique pour les cohortes combinées. L'effet observé est comparable à celui observé dans les autres cohortes internationales.
La troisième région connue pour être associée dans les populations caucasiennes est celle du gène *PTPN22*. Nous ne disposions cependant d'aucune puissance pour la cohorte Lyon 1, et d'une puissance limitée à 1,3% pour la cohorte Lyon 2 et 19,2% pour les cohortes combinées. Sur l'ensemble des deux cohortes, nous parvenons à une association à peine suggestive ($p=1,08.10^{-4}$) avec un OR pour l'allèle de risque estimé à 2 avec un IC95% [1,398 ; 2,86] incluant la valeur de l'OR estimé à 1,89 dans les cohortes internationales. Dans la cohorte française, la région du gène *PTPN22* n'est cependant pas la troisième région en termes de significativité et d'amplitude contrairement aux associations observées dans les cohortes internationales.
Sans surprise compte tenu de la taille de l'échantillon limitée, nous ne parvenons pas à détecter les autres associations connues par manque de puissance : puissance nulle pour toutes les autres régions dans chaque cohorte seule, et puissance inférieure à 1% pour les autres régions connues (**Figure 23**).

Nouvelle association suggestive sur le chromosome 2 avec le SNP rs17638639

En revanche, nous détectons une nouvelle association suggestive dans la cohorte Lyon 1 avec le SNP rs17638639 en 2q31.2. Nous répliquons cette association dans la cohorte Lyon 2 avec le même allèle associé et le même effet. Ainsi, dans les cohortes combinées, nous atteignons le seuil habituel des associations suggestives pangénomiques de 10^{-5} avec une p-value de $3,32.10^{-6}$. En effectuant une combinaison des valeurs p de chaque cohorte par la méthode de Fisher, nous obtenons une valeur $p = 1,91.10^{-6}$.
De façon remarquable, l'effet observé est très similaire à celui du SNP rs689 de l'insuline en terme d'amplitude avec un OR de 2,157 avec un IC95% [1,548 ; 3,007] et d'allèle de risque qui est l'allèle majeur. L'allèle majeur est plus fréquent que celui de l'insuline (90% versus 76%), ce qui explique une p-value moins significative que pour le SNP de l'insuline. Ce SNP est situé en 2q31.2. Parmi les

gènes à moins de 400kb de rs17638639, nous pouvons citer les gènes *OSBPL6*, *RBM45* et *PDE11A* (**Figure 24**). Un autre gène de la famille de *RBM45* a déjà été identifié comme associé au DT1 : le gène *RBM17* (région 24 ; **Tableau 4**). Par ailleurs, le gène *OSBPL6* code un récepteur lipidique intracellulaire. Enfin, le gène *PDE11A*, quant à lui, lié à la maladie de Cushing, code une protéine de la famille des phosphodiestérases dont les membres sont connus pour être impliqués dans un très grand nombre de phénomènes biologiques dont l'activation de l'immunité. Ce gène est connu comme potentiellement impliqué dans des phénomènes d'asthmes (DeWan et al. 2010, 11).
Il convient évidemment de répliquer cette nouvelle association dans une cohorte indépendante et d'affiner la localisation de cette association.

Association cohorte-spécifique de la région de susceptibilité BACH2 ?

Concernant la région du gène *BACH2* (6q15) déjà connue pour être associée au DT1 (**Tableau 4**) et correspondant au locus IDDM15, nous avons trouvé une association suggestive au seuil 1.10^{-5} uniquement dans la cohorte Lyon 1. Toutefois, les deux SNPs associés et en déséquilibre de liaison, rs62408234 et rs62408223, n'avaient encore jamais été associés au DT1.
Cependant cette association n'a pas été retrouvée dans la cohorte Lyon 2 ($p=2,9.10^{-1}$ pour rs62408223 et rs62408234) ni dans les cohortes combinées ($p=1,67.10^{-1}$ et $p=1.64.10^{-1}$ respectivement pour rs62408223 et rs62408234). Nous n'excluons pas qu'il s'agissait de vraies associations suggestives pour Lyon 1. Plusieurs pistes peuvent expliquer cette différence constatée entre les deux cohortes.
Le manque de puissance peut être une des causes essentielles, puisqu'elle n'était que de 38% dans la cohorte Lyon 2 pour un seuil alpha de 5%. De plus, ce calcul de puissance suppose un déséquilibre de liaison complet entre les variants génotypés et le variant causal. Or le variant causal de *BACH2* reste encore à identifier et le déséquilibre de liaison peut beaucoup varier d'une cohorte à l'autre.
Une troisième hypothèse serait que l'association détectée dans la cohorte Lyon 1 soit spécifique de la population dont cette cohorte est originaire. Il est en effet important de rappeler que les cohortes de Lyon incluent des Caucasiens originaires de France pour 2/3 ou du Maghreb pour 1/3. La structure de population des deux cohortes est peut être différente et pourrait engendrer des résultats divergents. Une stratification de population avait d'ailleurs pu être observée pour la cohorte Lyon 1 à partir des génotypes de l'immunochip (**Figure 26**) mais elle ne peut être étudiée dans la cohorte Lyon 2 faute d'un nombre suffisant de SNPs génotypés. Toutefois, on pourrait réanalyser les résultats obtenus pour la cohorte Lyon 1 dans chaque sous-groupe afin de vérifier si l'association observée est spécifique d'un seul sous-groupe.

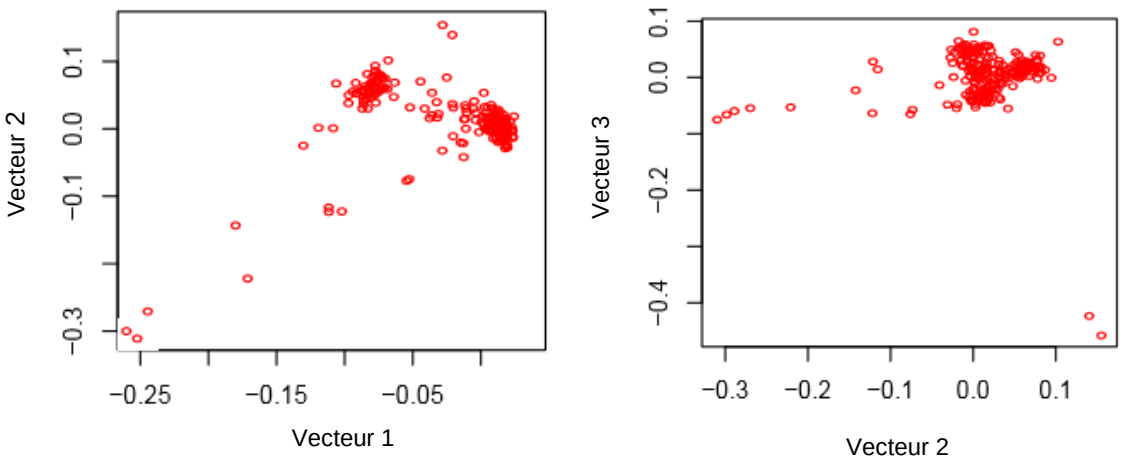


Figure 26. Structure de la population des patients de la cohorte Lyon 1. Analyse de l'hétérogénéité des patients de la cohorte Lyon 1 par positionnement multi-dimensionnel (MDS) à partir des génotypes de 13 000 SNPs en équilibre de liaison. Le panneau de gauche représente la dispersion des sujets selon les deux premiers « vecteurs propres » du MDS, celui de droite selon le 2ème et le 3ème « vecteur propres » de l'analyse MDS. Ces trois premiers « vecteurs propres » sont les trois axes qui ont les effets les plus forts sur la variabilité génétique. Le panneau de gauche montre deux groupes bien distincts, témoignant d'une stratification de population, le panneau de droite montre quelques sujets aberrants restants.

Absence de réplication des trois autres régions suggestives identifiées dans la cohorte Lyon 1

Les autres associations suggestives dans la cohorte Lyon 1 n'ont pu être répliquées dans celle de Lyon 2. Il est possible qu'il s'agisse de faux positifs dans la cohorte Lyon 1. Cependant, la puissance de reproduire ces associations était également limitée dans la cohorte Lyon 2 (entre 5% et 65% selon le SNP) et il n'est donc pas exclu qu'il s'agissait de vraies associations suggestives. Il pouvait également s'agir de vrais positifs spécifiques de la population dont était issue la cohorte Lyon 1.
L'ensemble de ces résultats pourrait être approfondi par une étude de type méta-analyse ou méta-régression prenant en compte les effets de chaque cohorte, leur hétérogénéité éventuelle et pondérant la statistique finale selon la taille de chaque cohorte.

2. Etude transcriptomique : « T1DGC-Express »

Plus de 50 locus sont associés au DT1 mais très peu d'entre eux ont été validés comme fonctionnellement responsables de la maladie. De plus, les locus impliqués dans la prédisposition au DT1 n'expliquent pas la totalité de l'héritabilité manquante. Des approches fonctionnelles peuvent aider, d'une part à préciser les variants et gènes causaux au sein des régions déjà associées à la maladie, et d'autre part à identifier de nouveaux gènes candidats à la maladie. Le projet « T1DGC-Express » développé au laboratoire s'inscrit dans un projet collaboratif international financé par le NIH et repose sur les cohortes internationales du T1DGC. Il a comme objectif principal d'identifier de nouveaux gènes de prédisposition au DT1 en comparant l'expression génique entre patients et leurs contrôles apparentés, porteurs des mêmes allèles de risque majeurs du CMH de classe II, *HLA-DR3* et *HLA-DR4-DQ8*, afin de s'affranchir de leur effet. Pour cela, une sélection originale des paires de germains au sein des familles a été effectuée : les patients diabétiques et leurs germains non atteints, bien que partageant les mêmes allèles de risque au locus *HLA-DRB1-DQB1*, ont hérité d'au moins un haplotype parental différent pour l'ensemble du CMH (cf. Matériel et Méthodes section 1.3 et **Figure 27**). Cette stratégie particulière permet de masquer l'effet du locus majeur *HLA-DRB1-DQB1* afin de démasquer de nouveaux gènes de prédisposition.

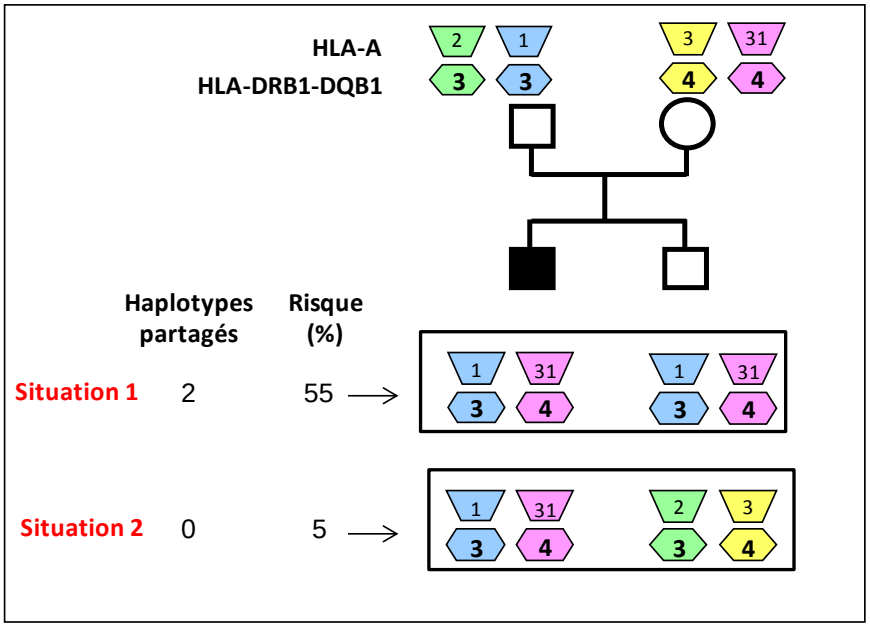


Figure 27. Stratégie de sélection des familles dans le cadre du projet « T1DGC-Express ». Dans cet exemple, l'arbre généalogique représente deux parents et leur deux fils, un atteint (carré noir) et un sain (carré vide). Les génotypes pour le locus *HLA-DRB1-DQB1* et pour le gène *HLA-A* sont donnés pour chaque sujet. Les allèles en phase sur les mêmes haplotypes sont de la même couleur. Dans la situation 1, les deux fils sont hétérozygotes DR3/DR4 et ils ont hérité du même haplotype paternel HLA-A*1-DR3 en bleu et du même haplotype maternel HLA-A*31-DR4 en rose. Dans la seconde situation, les enfants sont également hétérozygotes DR3/DR4 mais ces allèles n'ont pas été transmis par les mêmes haplotypes parentaux : le patient a hérité des haplotypes bleus du père et roses de la mère, tandis que son frère non atteint a hérité de l'haplotype HLA-A*2-DR3 du père en vert et de l'haplotype HLA-A*3-DR4 de la mère en jaune. Une étude prospective réalisée sur plusieurs familles a estimé un risque pour le germain non atteint de développer un DT1 à son tour de 55% si les deux enfants ont hérité des deux mêmes haplotypes parentaux avec l'allèle de risque DR3/DR4 comme dans la situation 1. Dans le situation 2 où ils n'ont pas hérité des deux mêmes haplotypes parentaux, le risque chute à seulement 5% (Aly et al. 2006). Ainsi, des facteurs du CMH autres que DR3 ou DR4 sont impliqués dans le DT1. Notre stratégie vise à recruter de telles familles ayant eu des enfants discordants pour le DT1 porteurs des allèles à risque DR3/DR4 mais ne partageant pas les mêmes haplotypes du CMH.

Avant mon arrivée au laboratoire, un transcriptome à partir des lignées lymphoblastoïdes (lymphocytes B immortalisés avec le virus Epstein-Barr) de 45 sujets (20 patients et 25 germains) de la cohorte européenne (EUR) du T1DGC avait ainsi été réalisé. Au total, 264 puces d'expression (Illumina Beadchip Human_HT12 V4) ont été hybridées correspondant, pour rappel, à deux réplicats de culture avec 3 conditions de stimulation avec l'agent mitogène PMA (0h : non stimulé ; 6h de PMA ; 24h : 6h de stimulation puis lavage et collecte après 18h). L'analyse de l'expression a permis d'identifier 211 jeux de sondes correspondant à 199 gènes montrant une expression différentielle statistiquement significative entre patients et contrôles au seuil ajusté de 10^{-3} (**Tableau 8**). De façon remarquable, cette liste de gènes différentiellement exprimés est significativement enrichie en gènes appartenant à la voie de l'interferon-γ ($p=2.10^{-5}$) (**Figure 28**), avec en particulier une diminution significative de l'expression de l'*IL27* et de l'*IFNγ*. Par ailleurs, au sein du CMH, 9 gènes étaient différentiellement exprimés entre patients et contrôles : *HLA-DPB1*, *HLA-DRB3*, *HLA-DRB6*, *HLA-DOA*, *TAP1*, *CUTA*, *KCTD20*, *IER3* et *UBD*. Ce dernier gène codant une ubiquitine avait déjà été associé au DT1 indépendamment de l'effet de la région de CMH de classe II (Baschal et al. 2011 ; Aly et al. 2008), mais aucune évidence fonctionnelle chez l'homme de cette association génétique n'avait été rapportée. Ces résultats préliminaires encourageants devaient donc être confirmés.

Tableau 8. Dix gènes différentiellement exprimés de l'étude transcriptomique. Parmi les 199 gènes différentiellement exprimés entre patients et contrôles, ce tableau liste les cinq gènes les plus significatifs (fond blanc) et les cinq gènes d'intérêt retenus pour la validation, *IFNγ*, *IL27*, *IRF1*, *STAT1* et *UBD* (fond gris*)*. Le rang de chaque gène est indiqué ainsi que les coordonnées du jeu de sondes sur la séquence de référence du génome version hg38. Pour le gène *STAT1*, trois jeux de sondes d'expression étaient présents sur la puce Illumina et donnent des résultats concordants. La différence d'expression entre patients et contrôles est exprimée en log2. Une valeur positive indique un gène plus exprimé chez les patients que les contrôles, tandis qu'une valeur négative indique un gène moins exprimé chez les patients que chez les contrôles. La valeur p de l'analyse statistique d'expression différentielle est donnée après ajustement pour correction multiple de type Benjamini-Hochberg.

Rang	Gène cible	Localisation chromosomique du jeu de sondes (hg38)	Expression différentielle Patients/ Contrôles	p-value ajustée
1	*KIAA1671*	chr22:25,197,125-25,197,174	-2,415	$2,17.10^{-12}$
2	*TIMM10*	chr11:57,528,544-57,528,593	0,844	$3,94.10^{-09}$
3	*SDSL*	chr12:113,438,114-113,438,000	-1,438	$7,36.10^{-09}$
4	*PCBP4*	chr3:51,957,634-51,957,683	1,756	$7,36.10^{-09}$
5	*TSPAN12*	chr7:120,787,589-120,787,000	-4,817	$4,47.10^{-08}$
35	*IRF1*	chr5:132,483,180-132,483,000	-1,226	$1,96.10^{-05}$
57	*UBD*	chr6:29,556,023-29,556,072	-2,635	$5,68.10^{-05}$
65	*IFNγ*	chr12:68,154,957-68,155,006	-3,681	$6,55.10^{-05}$
82	*STAT1* (jeu de sondes 1)	chr2:190,975,857-190,976,000	-0,753	$1,00.10^{-04}$
189	*STAT1* (jeu de sondes 2)	chr2:190,975,654-190,975,000	-0,625	$7,62.10^{-04}$
220	*STAT1* (jeu de sondes 3)	chr2:190,969,248-190,969,000	-0,817	$1,08.10^{-03}$
166	*IL27*	chr16:28,499,579-28,499,628	-3,996	$5,59.10^{-04}$

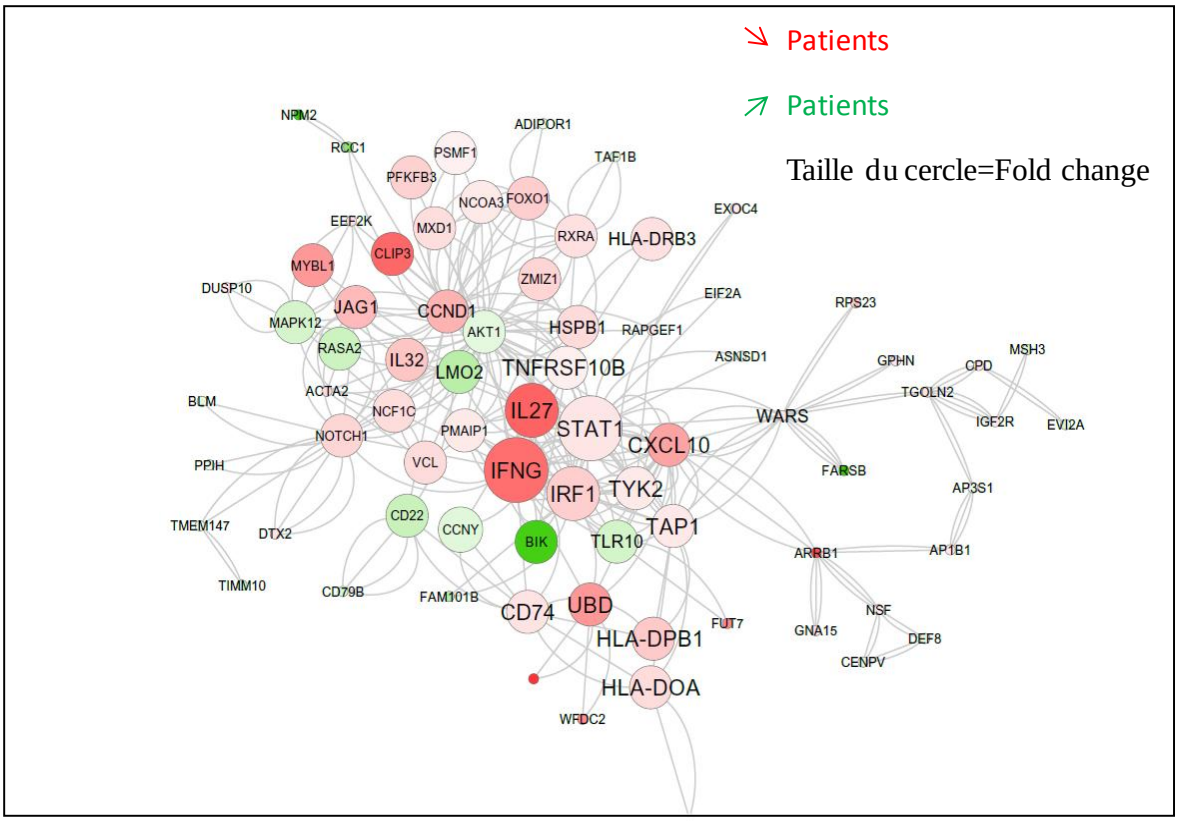


Figure 28. Réseau des gènes de la voie de l'IFNγ différentiellement exprimés entre patients et contrôles. Chaque cercle représente un gène de la voie de l'IFNγ significativement enrichie étudiée avec le logiciel Genomatix. La diminution de l'expression du gène chez les patients comparés aux contrôles est représentée par un cercle rouge et l'augmentation par un cercle vert. Plus le diamètre du cercle est grand, plus la différence d'expression (fold change) est importante. Représentation du réseau à l'aide du logiciel Cytoscape.

Mon objectif en rejoignant le laboratoire a ainsi consisté à tenter de valider ces résultats par des approches d'expression génique, tant au niveau des transcrits que des protéines. Pour cette étape de validation, cinq gènes d'intérêt ont été sélectionnés parmi les 199 différentiellement exprimés : 4 appartenant à la voie de l'interféron gamma (*IFN* γ, *IL27*, *IRF1* et *STAT1*), voie qui se distingue majoritairement au sein du réseau et un cinquième gène, *UBD*, excellent candidat au sein du CMH. Dans un premier temps, ma mission était de valider ces résultats par qPCR sur les ARNs des 264 mêmes échantillons (45 sujets) et de les répliquer sur 162 autres échantillons du T1DGC : 8 sujets EUR (48 échantillons) et 57 NAm (114 échantillons).

Dans un second temps, le but était de comparer les résultats du transcriptome à l'expression protéique par ELISA des protéines IFNγ et IL27 sur les 311 surnageants de culture cellulaire des lignées lymphoblastoïdes à disposition.

Enfin, le projet s'étend aussi aux cellules mononucléées du sang périphérique de patients et de contrôles du T1DGC des cohortes EUR et NAm (39 et 49 sujets respectivement) (cf. **Figure 17**) dont j'ai extrait l'ARN et dont le transcriptome est en cours par la technologie RNA-Seq.

2.1. Validation de l'étude transcriptomique par qPCR

Pour répondre au premier objectif, j'ai extrait les ARNs des 162 échantillons additionnels EUR et NAm et j'ai effectué leur contrôle qualité. J'ai ensuite préparé les cDNA des 264 ARNs utilisés pour la puce d'expression et des 162 échantillons additionnels, dessiné des oligonucléotides pour les cinq gènes cibles à l'aide du logiciel Primer 3 Plus et quantifié l'expression de ces gènes et de 3 gènes de ménage (*GAPDH*, *BACT*, *RPL30*) par PCR quantitative (qPCR) en temps réel avec du SyberGreen dans des plaques de 384 puits. Toutes les amplifications ont été réalisées en triplicats techniques et incluent les contrôles négatifs sans reverse transcriptase. Au total, 48 plaques de 384 puits ont été réalisées. L'analyse de ces résultats, au vu du nombre important d'échantillons et de gènes, nécessite d'être faite informatiquement et a été réalisé par ma tutrice scientifique, Claire Vandiedonck.

2.1.1. Traitement des données brutes et contrôle de qualité

Les données d'expression de type Cq brutes ont été modifiées pour tenir compte de l'efficacité des amorces (**Annexe 2**) ainsi que des facteurs de dilution différents utilisés entre les gènes de ménage et les gènes cibles. Au total, 17 626 mesures de qPCR ont été analysées (**Figure 29**). Pour chaque échantillon plusieurs contrôles de qualité ont été réalisés : identification des températures de fusion anormales ; vérification du Cq des échantillons sans reverse transcriptase par rapport à ceux obtenus avec l'enzyme pour identifier toute contamination avec de l'ADN génomique ; coefficient de variation entre les réplicats techniques qui doit être <5%. Les échantillons ne respectant pas ces critères de qualité ont été retirés de l'analyse.

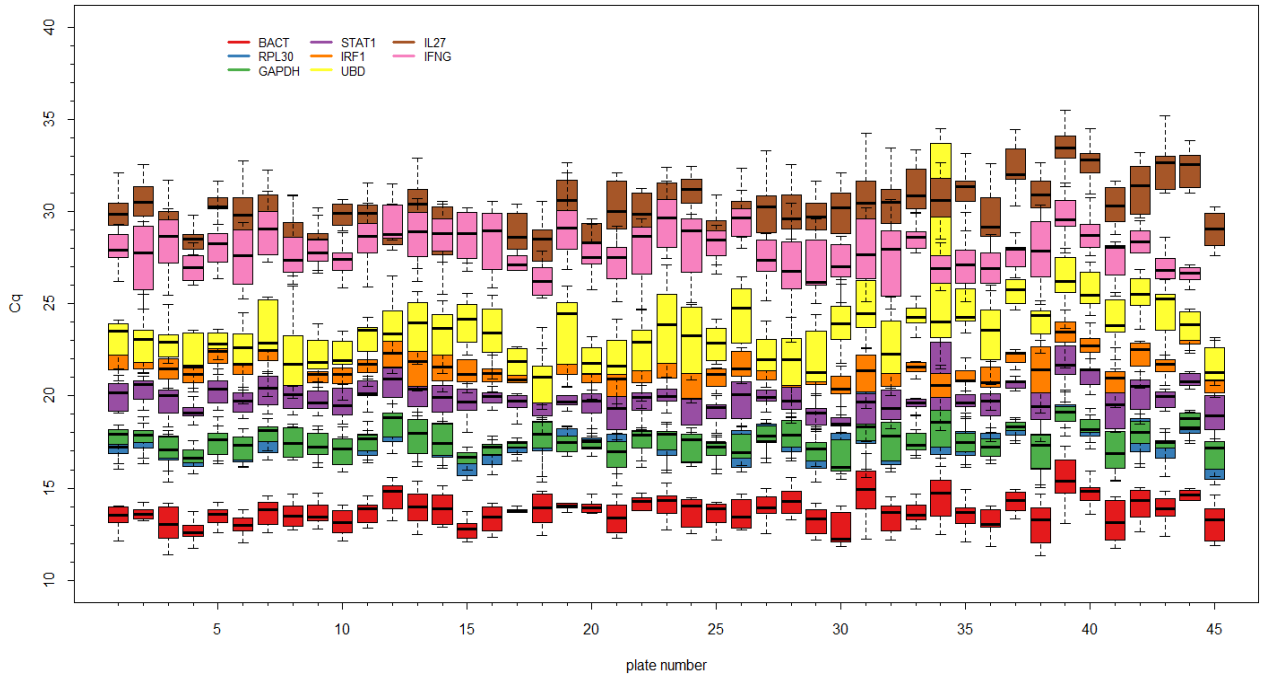


Figure 29. Distribution du niveau d'expression des gènes dans chacune des plaques de qPCR. Le niveau d'expression de chaque gène, représenté par une couleur différente, est donné par la valeur Cq non normalisée. Plus l'expression est forte, plus le Cq est faible. La représentation de la distribution sous forme d'une boite à moustache indique les valeurs minimales et maximales (hors points aberrants) aux extrémités des moustaches, le 1er et le 3ème quartile aux extrémités de la boîte et la médiane représentée par un trait épais horizontal.

La reproductibilité des mesures était excellente quel que soit le niveau d'expression du gène puisque le coefficient de variation (écart-type rapporté à la moyenne) se situait entre 0,00001 % et 8,87% (moyenne = 0,67%, médiane=0,48%) (**Figure 30**).

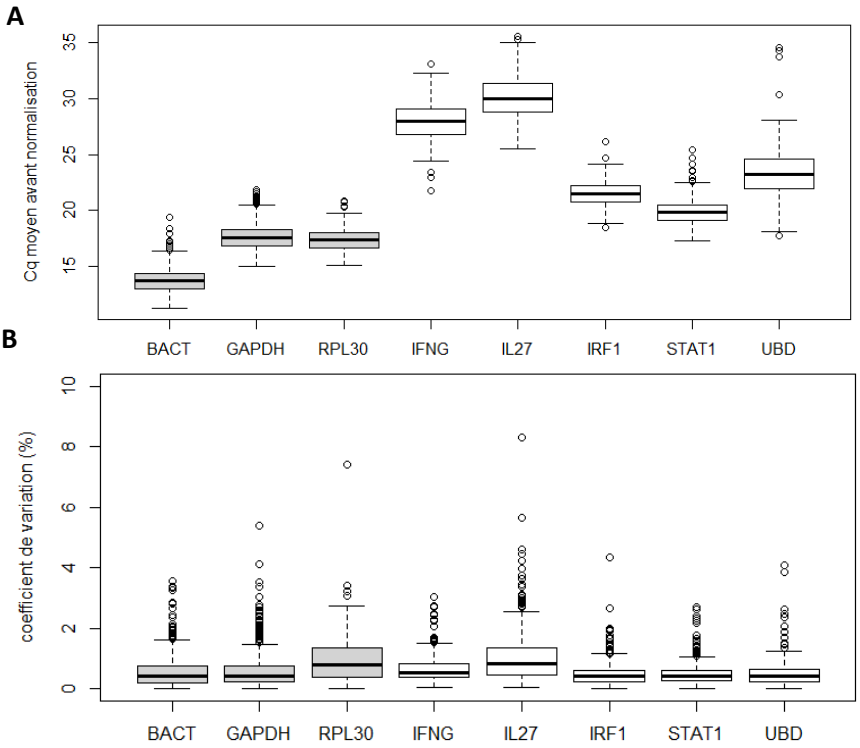


Figure 30. Distribution du niveau d'expression et du coefficient de variation par gène. Les boites à moustaches en ordonnée ont été générées à partir des données obtenues pour l'ensemble des plaques de qPCR pour les trois gènes de ménages (en gris) et les cinq gènes cibles (en blanc). Pour chaque gène en abscisse, le niveau d'expression en Cq **(A)** et le coefficient de variation **(B)** sont calculés à partir des triplicats techniques de chaque échantillon. Les points aberrants sont représentés par un cercle vide. Par convention, ils sont définis comme les valeurs au-delà du quartile +/- 1,5 fois la distance interquartile (=3ème quartile – 1er quartile).

Afin d'étudier la variabilité entre les plaques, 22 échantillons ont été quantifiés dans une seconde plaque de qPCR pour les différentes cibles et gènes de ménage, se répartissant sur 9 plaques différentes. Les valeurs de Cq étaient très bien corrélées avec un coefficient de corrélation de Pearson (r) sur l'ensemble des gènes de 0,9952 (Intervalle de Confiance (IC) à 95% = [0,9928-0,9967], p-value = 1,82x10^{-106}) (**Figure 31A**). De plus, la corrélation du Cq des échantillons entres les plaques a été calculée pour chaque gène et est significative quel que soit le gène (**Figure 31B**) : *BACT* : r=0,77 [0,52 ; 0,90], p=2,3x10^{-5} ; *RPL30* : r=0,82 [0,61 ; 0,92], p=2,8x10^{-6} ; *GAPDH* : r=0,80 [0,57 ; 0,91], p=7.9x10^{-6} ; *STAT1* : r=0,82 [0,28 ; 0,97], p=0,012; *IRF1* : r=0,73 [0,05 ; 0,94], p=0,039; *UBD* : r=0,86 [0,41 ; 0,98], p=0,0054 ; *IL27* : r=0,94 [0,72 ; 0,99], p=4,02x10^{-4} ; *IFNγ* : r=0,97 [0,83 ; 0,99], p=7,3x10^{-5}. Il n'y a donc pas d'impact du « facteur » plaque (analyse de variance multi-variée à 3 facteurs : « plaque » = non significatif, « cible » p=8,1x10^{-173} et « échantillon-condition » p=9,23x10^{-28}).

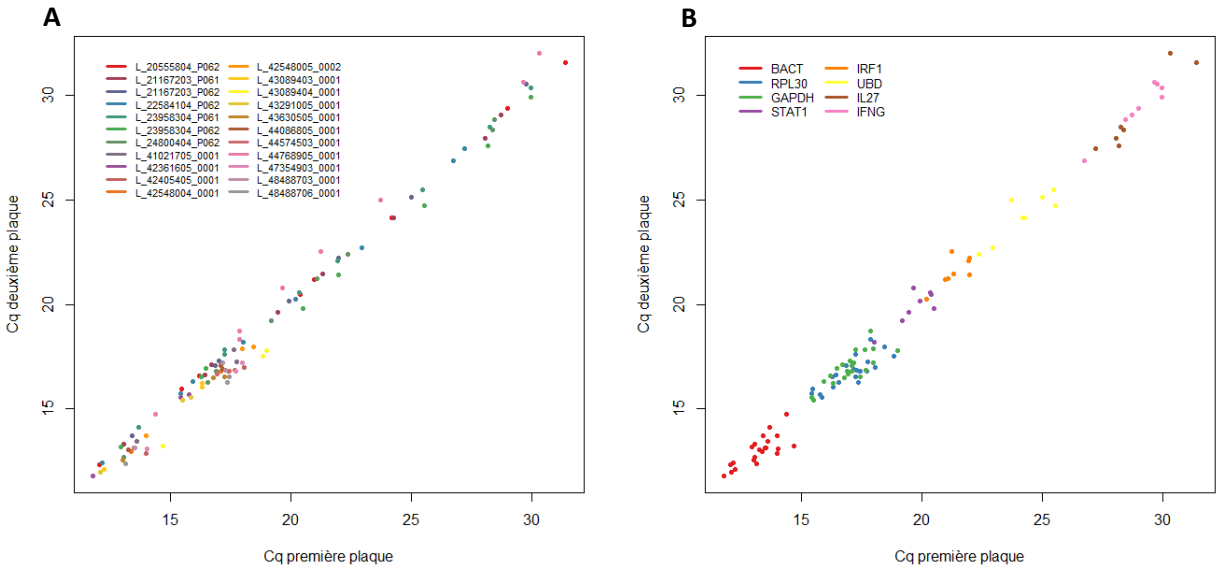


Figure 31. Corrélation inter-plaques des échantillons répétés pour chacun des gènes d'intérêts et de ménage. Chaque point (n=106) représente le Cq obtenu pour un échantillon dans la première plaque en abscisse et celui obtenu dans la seconde plaque où il a été répété en ordonnée. En **(A)** les points sont colorés en fonction de l'échantillon-condition, c'est-à-dire de l'identifiant de l'échantillon combiné à la condition de stimulation ; en **(B)** les points sont colorés selon le gène cible (22 mesures par gène de ménage, 8 mesures par gène cible).

Il est important de vérifier que l'expression de chaque gène de ménage varie de la même manière entre les échantillons quelles que soient les conditions. Par exemple, un échantillon X qui exprimera plus *BACT* qu'un échantillon Y exprimera également d'avantage *GAPDH* et RPL30. La **figure 32** montre que c'est globalement le cas sur l'ensemble des échantillons pour les 3 gènes de ménage, ce qui est confirmé par une étude de corrélation des gènes de ménage deux à deux (**Figure 33**). On obtient une très bonne corrélation (Pearson) significative pour chaque paire : r=0,903 avec un IC à 95% =[0,883 ; 0,918] et p=5,27x10^{-165} pour BACT/GAPDH ; r=0,921 [0,906 ; 0,934], p=1,07x10^{-184} pour BACT/RPL30 ; et r=0,932 [0,919 ; 0,943], p=1,42x10^{-198} pour *GAPDH*/*RPL30*) avec notamment à noter un niveau d'expression similaire des transcrits de *RPL30* et *GAPDH*, respectivement 12 et 14 fois inférieur à celui des transcrits de *BACT*. Ainsi, ces résultats montrent que nous pouvons utiliser la moyenne du Cq de ces 3 gènes de ménage pour effectuer la normalisation de chacun des gènes cibles.

L'expression relative dite « deltaCq » est donc calculée sur la moyenne du Cq des 3 gènes de ménage. Enfin, l'expression relative est normalisée sur l'échantillon ayant la plus faible expression, qui peut être différent d'un gène à l'autre. Les « deltadeltaCq » ainsi obtenus sont ainsi directement comparables entre conditions et jeux de données.

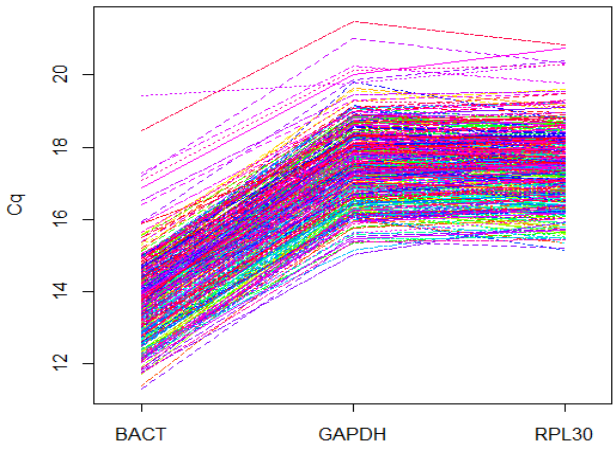


Figure 32. Niveau d'expression de chaque échantillon pour les 3 gènes de ménage. Chaque trait de couleur représente un échantillon différent et relie le Cq obtenu pour les 3 gènes de ménage.

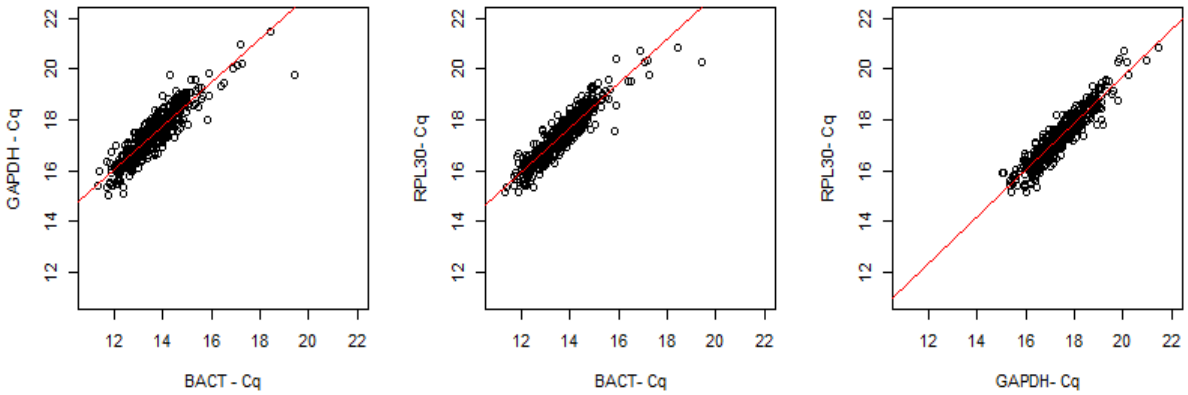


Figure 33. Corrélation deux à deux du niveau d'expression des gènes de ménage sur l'ensemble des échantillons. Chaque point correspond en abscisse à la valeur Cq obtenue pour un échantillon pour un des trois gènes de ménage et en ordonnée par sa valeur obtenue pour un second gène de ménage. La droite en rouge correspond à la droite de régression linéaire.

2.1.2. Facteurs de variation de l'expression génique

Parmi les sources de variations d'intérêt sur l'expression des gènes, nous souhaitons étudier l'impact de trois facteurs : la cohorte, le statut vis-à-vis de la maladie et la stimulation. Le plan expérimental est déséquilibré puisque le facteur « stimulation » ne peut être étudié que sur les échantillons EUR, les échantillons NAm étant tous non stimulés.

La première source de variation testée est donc la cohorte. Sur l'ensemble des échantillons, le niveau d'expression des gènes cibles est toujours inférieur à celui des 3 gènes de ménage selon un gradient décroissant : *STAT1* (ΔΔCq moyen =2,32), *IRF1* (ΔΔCq moyen =4,02), *UBD* (ΔΔCq moyen =5,74), *IFNγ* (ΔΔCq moyen = 10,47) et *IL27* (ΔΔCq moyen =12,56). On observe cependant une différence d'expression significative selon la cohorte : les échantillons EUR non stimulés ont une expression relative plus élevée que les échantillons NAm pour les gènes *IL27* et *UBD* et plus faible pour *IRF1* (**Figure 34**).

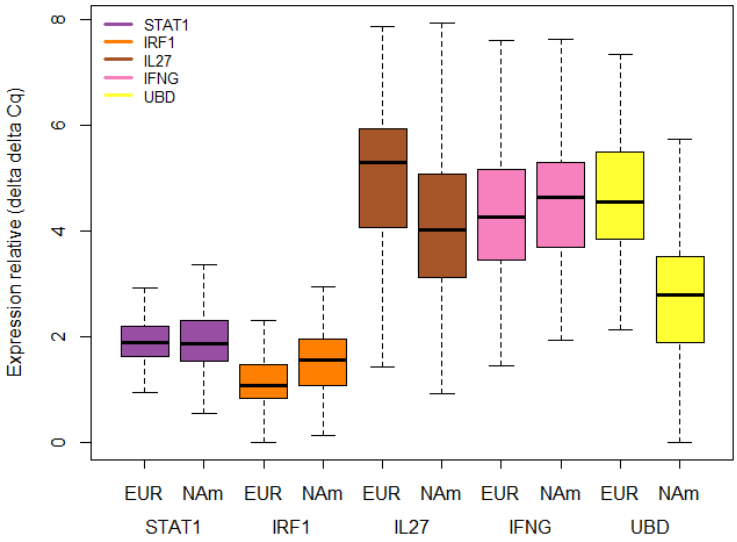


Figure 34. Niveau d'expression relative de chaque gène cible selon la cohorte dans les échantillons non stimulés. Boîtes à moustache de la distribution des expressions relatives ΔΔCq par gène (indiqué selon une couleur différente) et par cohorte. Un test de Welch de comparaison des moyennes entre EUR et NAm a été effectué pour chaque gène et est significatif pour I*RF1* ($p=6,55x10^{-7}$), *IL27* ($p=2,15x10^{-6}$) et *UBD* ($p<2.2x10^{-16}$).

L'étude de l'impact du statut est donc menée sur chaque cohorte indépendamment. L'analyse est faite par régression linéaire. Ici, la variable dite « expliquée » est le niveau d'expression du gène (ΔΔCq). Les variables dites « explicatives » sont les facteurs d'intérêt : le « statut » vis-à-vis de la maladie, le « pedigree » puisque les patients et contrôles sont des germains et pour la cohorte EUR, la « stimulation ».

Pour la cohorte Nord Américaine, un seul modèle prenant en compte le statut et le pedigree a donc été testé. L'étude par régression linéaire du facteur « statut » n'a donné aucun résultat significatif entre patients et contrôles pour les cinq gènes d'intérêt. Le tableau ci-dessous résume les résultats obtenus pour chaque gène après analyse (**Tableau 9**).
Pour la cohorte européenne, deux modèles ont été testés par régression linéaire. Le premier prend en compte les facteurs « statut », « pedigree » et « stimulation » sur l'ensemble des échantillons. Il met en évidence une différence significative entre patients et contrôles pour les cinq gènes d'intérêt (**Tableau 10**). En revanche, le facteur stimulation n'influence pas significativement l'expression génique. Le deuxième modèle prend en compte les facteurs « statut » et « pedigree » dans quatre sous-groupes définis par le temps de stimulation : non stimulés, stimulés à 6h, stimulés à 24h et stimulés à 6 ou 24h. On constate que la différence d'expression entre patients et contrôles est d'autant plus marquée que les échantillons sont stimulés. Il s'agit toujours d'une diminution chez les patients avec une amplitude variant de 45,3% (*UBD* à 6h) à 84,3% (*STAT1* à 6h). Si la différence est déjà significative pour *IL27* et *UBD* sans stimulation, il faut attendre 6h de stimulation pour *IFNγ* ou *STAT1* et 24h pour *IRF1*. Les résultats les plus significatifs sont observés en agrégeant les échantillons stimulés à 6 et 24h. Les résultats, gène par gène, sont détaillés dans le **tableau 10** et la **figure 35** ci-dessous. Sur la cohorte EUR, nous validons donc les résultats de l'étude du transcriptome avec une diminution de l'expression des transcrits chez les patients pour les cinq gènes cibles testés.

Tableau 9. Résultats obtenus après étude par régression linéaire de la différence d'expression entre patients et contrôles de la cohorte Nord Américaine. Ce tableau liste la moyenne de l'expression relative (ΔΔCq) obtenue pour les contrôles et les patients, ainsi que le rapport de l'expression des patients comparés aux contrôles ($2^{-\Delta\Delta Cq}$) pour chacun des 5 gènes d'intérêt. Un ratio positif est représentatif d'une expression plus élevée chez les patients comparé aux contrôles. L'impact du facteur « statut » est donné par sa pente β estimée avec son intervalle de confiance à 95%. Si la pente est nulle, il n'y a pas d'effet du facteur. Dans les deux dernières colonnes, le modèle est testé : ici la variable expliquée « expression du gène » est influencée par les variables explicatives « statut » et « pedigree ». Le r^2 donne la proportion de la variation de l'expression du gène expliquée par le modèle.

	Modèle statut + pedigree						
	Contrôles	**Patients**	**Ratio de l'expression (patients / contrôles)**	**Facteur statut**		**Test du modèle**	
	Moyenne ΔΔCq	**Moyenne ΔΔCq**	**$2^{-\Delta\Delta Cq}$**	**β Statut [IC95%]**	**p-value**	**r^2**	**p-value**
IFNγ	4,311	4,563	1,191	0,239 [-0,162 ; 0,640]	NS	0,325	$3{,}04.10^{-5}$
IRF1	1,547	1,629	1,059	0,080 [-0,184 ; 0,345]	NS	0,084	NS
STAT1	1,825	2,047	1,166	0,204 [-0,063 ; 0,470]	NS	0,072	NS
IL27	4,23	3,932	0,814	-0,283 [-0,769 ; 0,203]	NS	0,298	$1{,}20.10^{-4}$
UBD	3,061	2,679	0,767	-0,361 [-0,731 ; 0,008]	NS	0,468	$7{,}56.10^{-9}$

Tableau 10. Résultats obtenus après étude par régression linéaire de la différence d'expression entre patients et contrôles de la cohorte européenne. La partie de gauche du tableau liste la moyenne de l'expression relative (ΔΔCq) obtenue pour les contrôles et les patients, la pente β estimée avec son intervalle de confiance à 95% et le r^2 pour le modèle statut+pedigree+stimulation. La partie de droite donne la pente β estimée et la valeur p du modèle testé dans la condition de stimulation indiquée.

			Modèle statut + pedigree + stimulation							**Modèle statut+pedigree**							
	Contrôles	**Patients**	**Facteur statut**		**Facteur stimulation**			**Test du modèle**		**à 0h**		**à 6h**		**à 24h**		**à (6h+24h)**	
	ΔΔCq	**ΔΔCq**	**β statut [IC95%]**	**p -value**	**β 6h**	**β 24h**	**p-value**	**r^2**	**p -value**	**β statut [IC95%]**	**p-value**	**β statut [IC95%]**	**p-value**	**β statut [IC95%]**	**p-value**	**β statut [IC95%]**	**p-value**
INFγ	4,487	3,561	-0,919 [-1,227;-0,610]	1,25E-8	-0,402	-0,467	NS	0,228	7,29E-12	-0,489 [-1,04;0,067]	NS	-1,072 [-1,623;-0,520]	2,1E-4	-1,115 [-1,707; -0,523]	3,3E-4	-1,096 [-1,479;-0,714]	5,7E-8
IRF1	1,542	1,268	-0,308 [-0,450;-0,167]	2,50E-5	0,414	0,391	NS	0,166	3,29E-8	-0,142 [-0,373;0,090]	NS	-0,445 [-0,718;-0,172]	NS	-0,324 [-0,572;-0,075]	0,011	-0,388 [-0,567;-0,210]	2,8E-5
STAT1	1,913	1,762	-0,186 [-0,299;-0,073]	1,37E-3	-0,15	0,053	NS	0,157	1,06E-7	-0,117 [-0,328;0,094]	NS	-0,236 [-0,437;-0,036]	0,022	-0,204 [-0,391;-0,017]	NS	-0,229 [-0,365 ; -0,091]	1,2E-3
IL27	5,707	5,137	-0,692 [-0,967;-0,416]	1,29E-6	1,131	0,17	NS	0,39	1,11E-25	-0,525 [-1,016;-0,034]	0,036	-0,777 [-1,226;-0,328]	8,9E-4	-0,775 [-1,314;-0,236]	5,3E-3	-0,741 [-1,107 ; -0,375]	9,2E-5
UBD	4,925	4,069	-0,987 [-1,280;-0,694]	1,67E-10	-0,162	-0,217	NS	0,322	2,23E-19	-0,756 [-1,191;-0,320]	8,9E-4	-1,198 [-1,756;-0,640]	4,9E-5	-0,957 [-1,553;-0,360]	2.1E-3	-1,090 [-1,477 ; -0,702]	9,5E-8

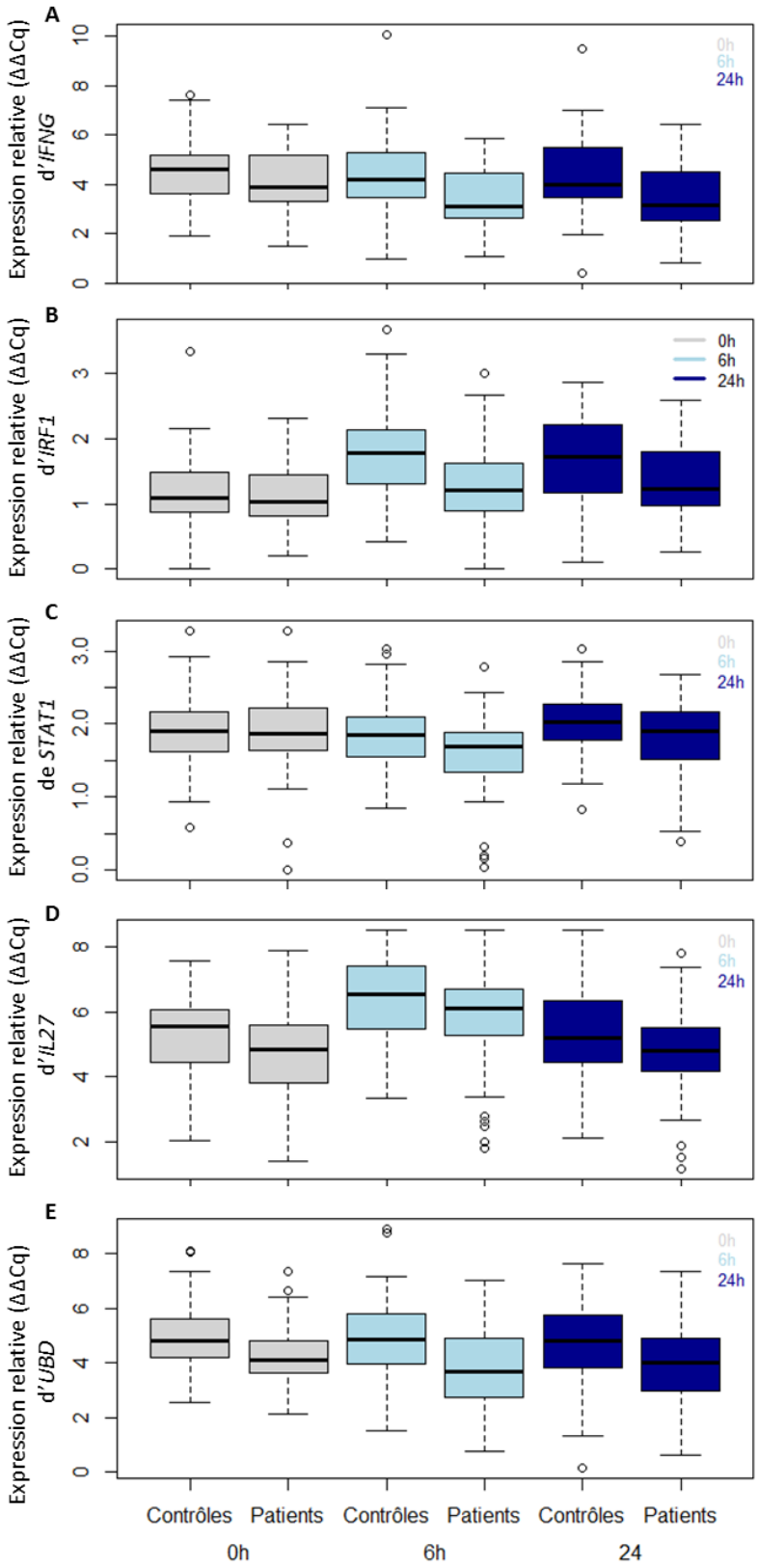


Figure 35. Expression relative des cinq gènes d'intérêts, chez les patients et les contrôles de la cohorte européenne avec ou sans stimulation au PMA. Boites à moustaches de la distribution des expressions relatives pour chacun des cinq gènes d'intérêt à des temps de stimulation différents (indiqués par une couleur différente, gris=0h de stimulation, bleu clair=6h et bleu foncé=24h). Une diminution significative de l'expression des patients comparée aux contrôles est constatée pour les cinq gènes. Cette diminution est d'autant plus marquée avec la stimulation à 6 ou 24h.

2.1.3. Corrélation avec les puces d'expression

Pour conforter la validation, nous avons regardé si les résultats des tests de qPCR échantillons individuels étaient corrélés à ceux qui avaient été obtenus avec les puces. Cela était possible pour 48 sujets étudiés avec les deux méthodes. Une seule sonde Illumina permettait de quantifier chaque gène sauf pour *STAT1* pour lequel 3 sondes Illumina permettaient de quantifier son expression (**Tableau 11**).

Tableau 11. Exons ciblés par qPCR et les puces d'expression pour les 5 gènes d'intérêt

Gène	Exon(s) ciblé(s) pour les puces d'expression	Exon(s) cible(s) qPCR
UBD	exon 2	exon 1-2
IFNγ	exon 4	exon 4
IL27	exon 5	exon 3-4
IRF1	exon 10	exon 3-4
STAT1	exon 22-23 (1ère sonde) exon 23 (2ème sonde) exon 25 (3ème sonde)	exon 18-19

Toutes les corrélations (Pearson) des données de qPCR avec celles des sondes sont positives et très significatives (**figures 36 et 37**). Il est important de noter que pour *STAT1*, les corrélations sont équivalentes pour chacune des sondes (**Figure 37**). Pour l'*IL27* la puce n'a pu détecter l'expression dans tous les échantillons contrairement à la qPCR, cette technique semble donc plus sensible.

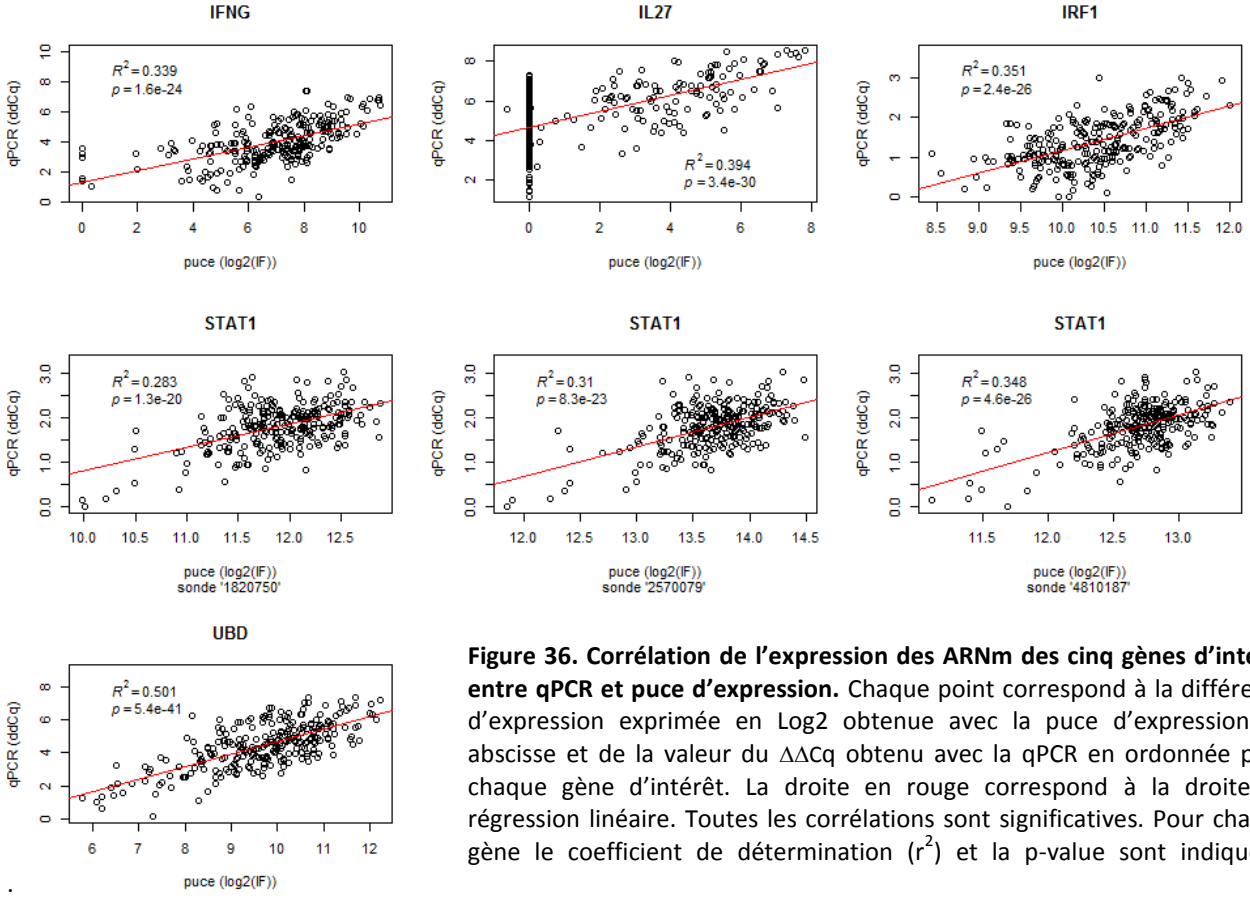


Figure 36. Corrélation de l'expression des ARNm des cinq gènes d'intérêt entre qPCR et puce d'expression. Chaque point correspond à la différence d'expression exprimée en Log2 obtenue avec la puce d'expression en abscisse et de la valeur du ΔΔCq obtenu avec la qPCR en ordonnée pour chaque gène d'intérêt. La droite en rouge correspond à la droite de régression linéaire. Toutes les corrélations sont significatives. Pour chaque gène le coefficient de détermination (r^2) et la p-value sont indiquées.

.

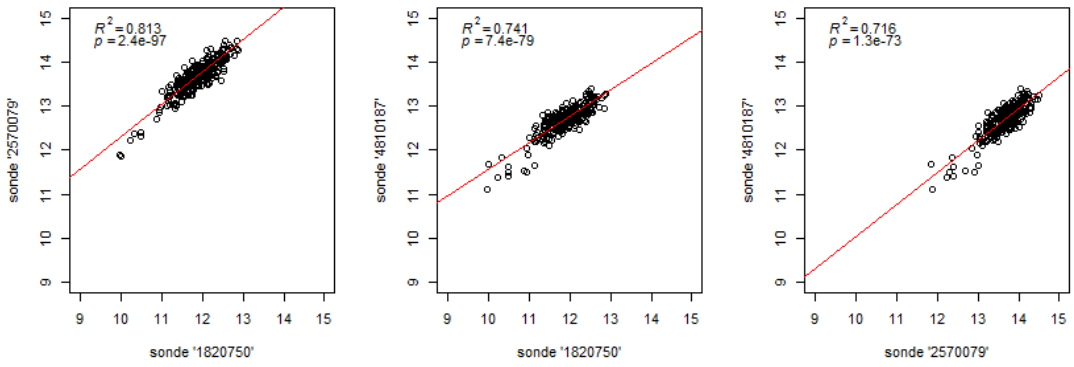


Figure 37. Corrélation de l'expression de STAT1 entre les 3 sondes de la puce. Chaque point correspond à la différence d'expression exprimée en Log2 obtenue pour une des 3 sondes de *STAT1*. La droite en rouge correspond à la droite de régression linéaire.

2.2. Etude de l'expression protéique

Nous avons voulu également tester si la variation de l'expression génique entre patients et contrôles concernait exclusivement les transcrits ou si elle se manifestait également au niveau protéique. Pour répondre à cette question, l'expression de deux cytokines sécrétées, l'IL27 et l'IFNγ, a été dosée par ELISA dans 311 surnageants disponibles de la cohorte EUR. Ces 311 surnageants ont été récupérés à partir de la culture des 53 lignées lymphoblastoides avec ou sans stimulation au PMA précédemment étudiées en qPCR, correspondant à 24 patients et 29 contrôles. Les expériences d'ELISA ont été réalisées en 2 réplicats techniques. Les concentrations obtenues ont été normalisées par rapport à la quantité totale de protéines mesurées par BCA.

2.2.1. Corrélation de l'expression des ARNm et des protéines secrétées

Concernant l'IFNγ dans l'ensemble de la cohorte européenne, les résultats avec la qPCR et la puce d'expression montraient une diminution d'expression chez les patients et les contrôles au cours du temps après stimulation au PMA (**Figure 38A**). La technique ELISA n'a pas permis de détecter la protéine secrétée sans stimulation. Elle n'est détectable qu'à 24h dans l'ensemble de la cohorte (**Figure 38B**).

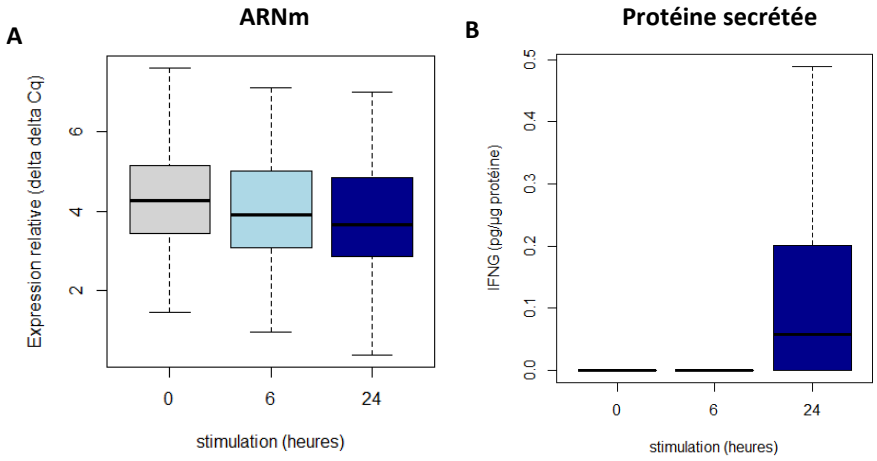


Figure 38. Expression relative et dosage protéique par ELISA de l'IFNγ dans la cohorte européenne, avec ou sans stimulation au PMA. Boites à moustache de l'expression relative obtenue par qPCR **(A)** montrant une diminution de l'expression qui s'accentue au cours du temps avec la stimulation et du dosage de l'IFNγ par ELISA **(B)** qui montre une expression de la protéine secrétée qui n'est détectable qu'après 24h de stimulation.

L'expression des transcrits et des protéines n'est donc comparable qu'à ce temps de stimulation. On observe une bonne corrélation entre les deux avec un coefficient de Spearman r=0,878 et une p-value de $1{,}73 \times 10^{-34}$ (**Figure 39**).

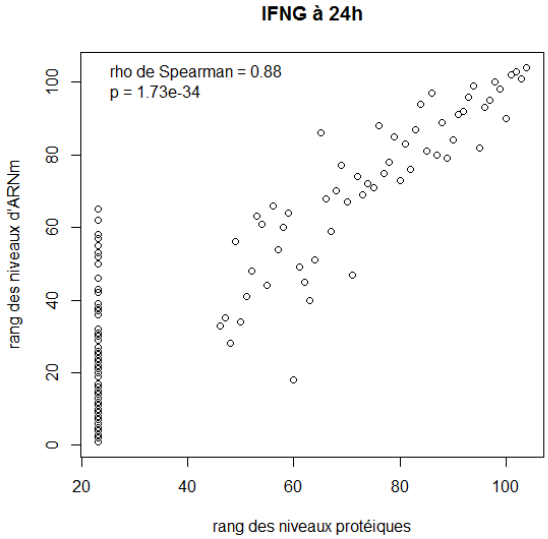


Figure 39. Corrélation de l'expression des ARNm et des protéines de l'IFNγ après 24h de stimulation. Chaque point correspond aux rangs des niveaux d'expression des protéines en abscisse et des transcrits en ordonnée. Une bonne corrélation est retrouvée entre l'expression des ARNm et des protéines de l'IFNγ avec r=0,878 et une p-value correspondante de $1{,}73 \times 10^{-34}$.

L'impact du statut sur l'expression de la protéine secrétée a donc été testé (par test de rang non paramétrique de Wilcoxon) et nous constatons que la médiane des patients (médiane= 0,00000) est significativement inférieure à celle des sujets sains (médiane=0,09956) avec une p-value de 0,00679 (**Figure 40**).

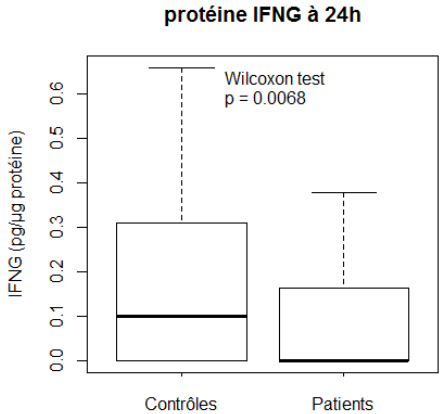


Figure 40. Comparaison de l'expression protéique de l'IFNγ secrété entre patients et contrôles après 24h de stimulation. Boites à moustache de l'expression protéique des contrôles en comparaison de celle des patients à 24h de stimulation.

Concernant l'IL27, on constatait pour l'ARNm un pic d'expression à 6h dans l'ensemble de la cohorte européenne (**Figure 41A**). Au niveau protéique, l'ELISA montre une expression de l'IL27 qui augmente à 6h et plus encore à 24h (**Figure 41B**) mais qui semble très variable selon les sujets, c'est-à-dire qu'elle diminue chez certains à 24h après avoir été exprimée à 6h, tandis que d'autres

sujets ont un niveau d'expression qui continuent d'augmenter à 24h ou que d'autres commencent seulement à exprimer la protéine secrétée à 24h (**Figure 42**).
Le test non paramétrique de Spearman montre qu'il n'y a pas de corrélation entre l'expression des ARNm et des protéines, même en prenant 6h pour les ARNm et 24h pour les protéines (**Figure 43**).
L'impact du statut sur l'expression protéique testé par test de rang non paramétrique de Wilcoxon montre qu'il n'y a pas de différence statistiquement significative (p-value=0,568) (**Figure 44**). La médiane des contrôles est de 1,460 contre 1,554 pour les patients.
Ainsi, si cette analyse ne montre pas de différence d'expression de l'IL27 secrétée entre patients et contrôles, en revanche, pour l'IFNγ, la variation d'expression génique de l'IFNγ mise en évidence au niveau des transcrits se reflète aussi au niveau protéique.

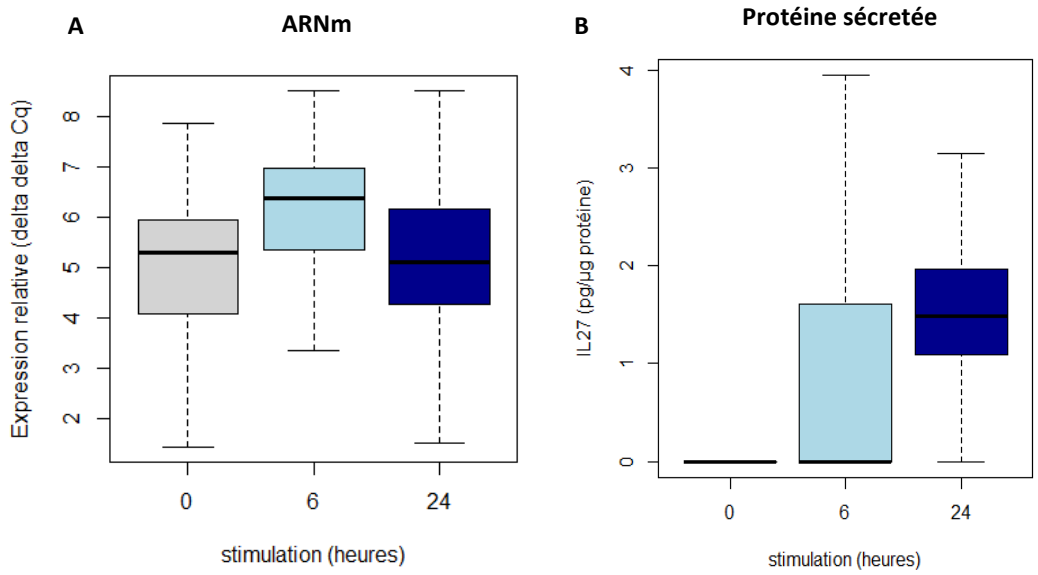


Figure 41. Expression relative et dosage protéique par ELISA de l'IL27 secrétée dans la cohorte européenne, avec ou sans stimulation au PMA. Boites à moustache de l'expression relative obtenue par qPCR **(A)** montrant un pic d'expression à 6h, ou de l'expression protéique par ELISA **(B)** détectable uniquement après 6 et 24h de stimulation.

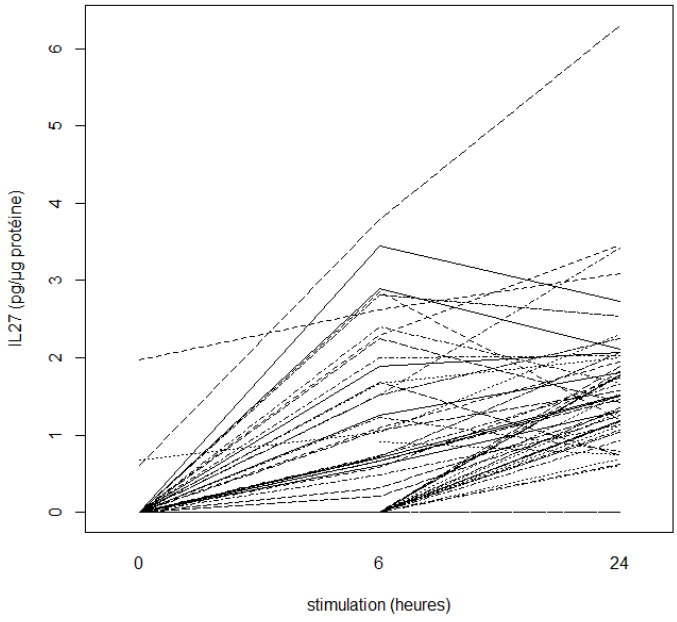


Figure 42. Dosage par ELISA de l'IL27 pour chaque individu en fonction du temps de stimulation. Chaque trait représente la concentration obtenue pour l'IL27 à 0,6 et 24h de stimulation pour un individu.

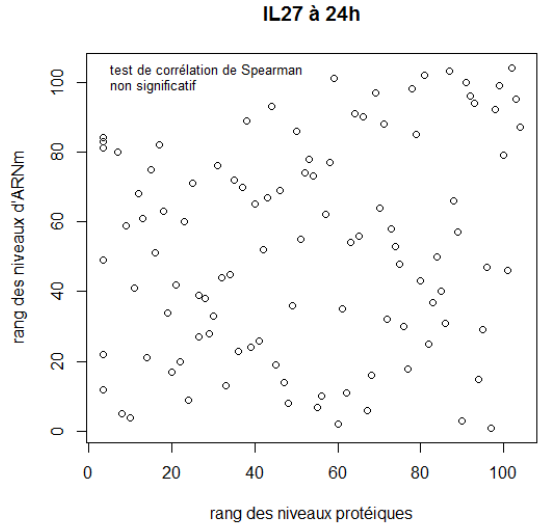


Figure 43. Corrélation de l'expression des ARNm et des protéines secrétées de l'IL27. Chaque point correspond aux rangs des niveaux d'expression des protéines en abscisse et des transcrits en ordonnée. Le test de corrélation est non significatif.

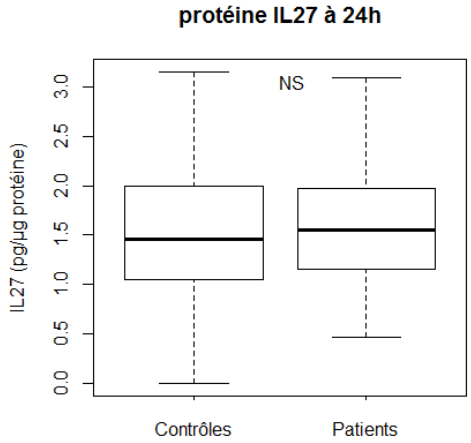


Figure 44. Comparaison de l'expression protéique de l'IL27 sécrétée entre patients et contrôles après 24h de stimulation. Boîtes à moustache de l'expression protéique de l'IL27. Un test de rang non paramétrique de Wilcoxon montre que la différence entre patients et contrôles n'est pas statistiquement significative (NS : Not Significant).

2.3. Etude de la corrélation entre les gènes

Au-delà de la variation individuelle des gènes, il est important de considérer cette variation de façon intégrée et d'examiner si elle se fait de façon corrélée entre les gènes pris deux à deux. Cette analyse intégrative se justifie d'autant plus que des relations fonctionnelles ont été mises en évidence entre ces différents gènes. En effet, à l'aide de l'outil informatique PCViz (http://www.pathwaycommons.org/pcviz/) qui collecte des informations venant de plusieurs bases de données publiques (Reactome, Panther Pathway ou encore KEGG pathway) (Cerami et al. 2011), on met en évidence plusieurs interactions entre les gènes d'intérêt (**Figure 45**). Ainsi, on retrouve la voie de signalisation connue selon laquelle IFNγ active le facteur de transcription STAT1 en le phosphorylant. A son tour celui-ci active la transcription d'IRF1. L'IL27 interagit avec ces trois partenaires : elle module la phosphorylation de STAT1 et contrôle l'expression de l'INFγ, tandis que sa propre expression est contrôlée par IRF1. Il est à noter que PCViz associe UBD à cette voie sur la base de la présence d'un motif de liaison pour IRF1 dans le promoteur de ce gène (Oliva et al. 2010).

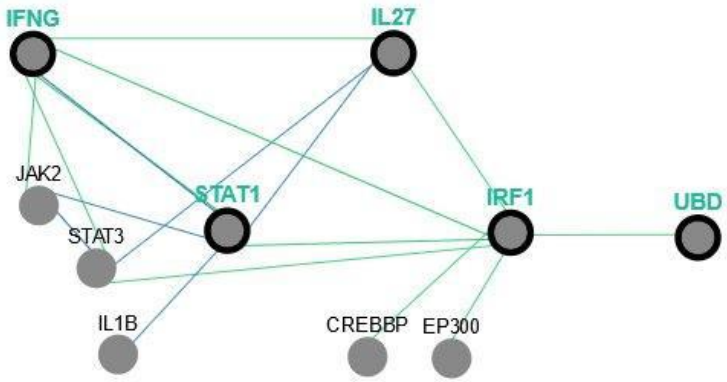


Figure 45. Voie de signalisation de l'IFNγ connectant les gènes d'intérêt selon l'outil PCViz. Chaque cercle représente un gène. Un trait vert représente le contrôle de l'expression d'un des deux gènes reliés à l'autre et un trait bleu au contrôle du changement d'état d'un des deux gènes par l'autre, comme une phosphorylation par exemple. Selon la voie de signalisation connue, l'IFNγ active STAT1 par phosphorylation (trait bleu) qui active l'expression d'IRF1 (trait vert). L'expression de l'IL27 est contrôlée à son tour par l'IRF1, tandis que l'IL27 contrôle l'expression de l'IFNγ (trait vert) et active la phosphorylation de STAT1 (trait bleu). De plus, selon PCViz, l'expression d'UBD serait contrôlée par IRF1.

Une étude des corrélations deux à deux de l'expression des transcrits quantifiés par qPCR a donc été réalisée pour les quatre gènes de la voie de l'IFNγ entre eux (**Figure 46**) et également avec *UBD* (**Figure 47**). La Figure 46 montre des corrélations très significatives entre le niveau des transcrits d'*IFNγ*, *STAT1* et *IRF1* avec des coefficients de détermination compris entre 38 et 55,5%. L'expression d'*IL27* est également significativement corrélée à celle d'IRF1 (19,4%) et dans une moindre mesure à celle de *STAT1* et d'*IFNγ* (autour de 10%). De façon impressionnante, les niveaux de transcrits d'*UBD* sont très fortement et très significativement corrélés avec les trois premiers gènes et particulièrement avec l'expression d'*IRF1* (r^2=0,621 p=1,4x10^{-66}) et surtout avec celle de l'*IFNγ* (r^2=0,704 p=7,01x10^{-81}). Toutes ces corrélations étant significatives, nous avons procédé à une analyse des corrélations partielles, qui élimine les corrélations indirectes. Cette étude précise la hiérarchie des corrélations directes (**Figure 48**). Les corrélations partielles les plus fortes sont notées entre l'IFNγ et UBD (r^2=0,183, p=3,52x10^{-15}) et entre STAT1 et IRF1 (r^2=0,171, p=3,57x10^{-14}). En revanche et en accord avec la voie de signalisation connue (phosphorylation), *STAT1* n'est plus corrélé ni à *IFNγ* ni à *IL27*.

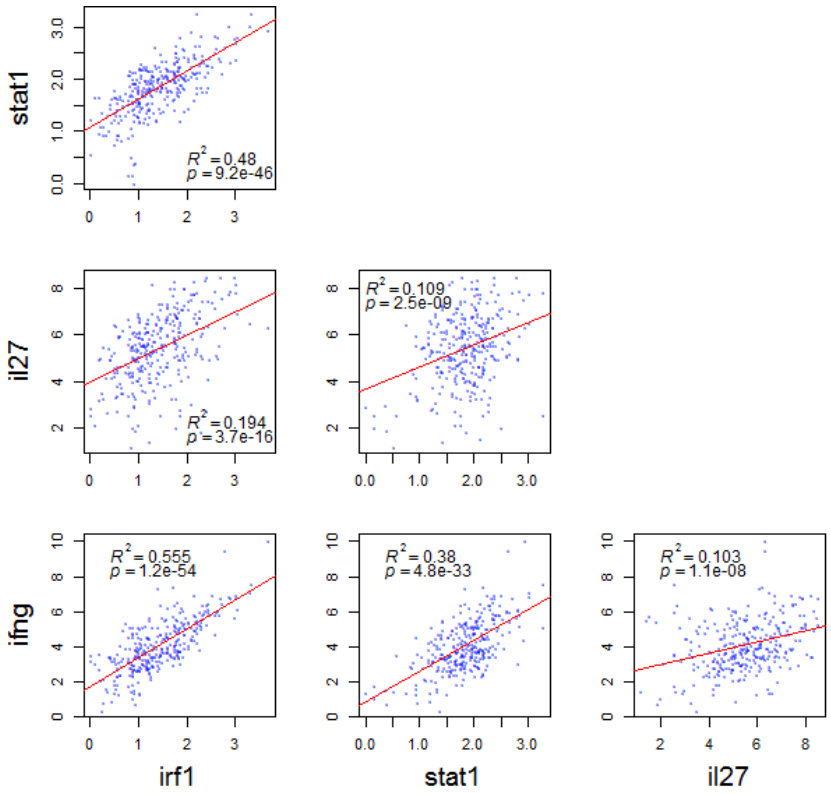


Figure 46. Corrélation deux à deux de l'expression relative des gènes de la voie de l'IFNγ (IFNγ, IL27, IRF1 et STAT1). Chaque point bleu correspond à l'expression relative d'un des quatre gènes en abscisse et d'un second gène en ordonnée. Le coefficient de détermination r^2 de Pearson et la p-value correspondante sont indiquées pour chaque paire de gènes. La droite en rouge correspond à la droite de régression linéaire.

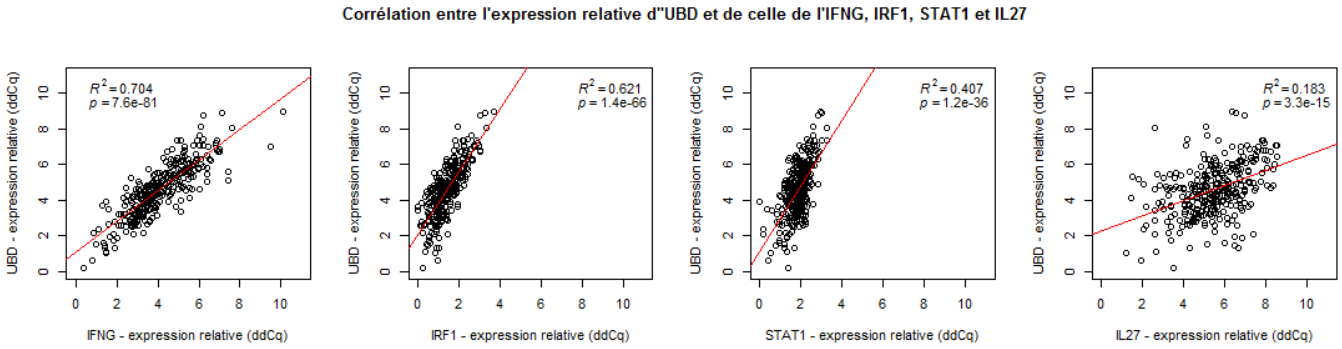


Figure 47. Corrélation de l'expression relative d'UBD avec celle d'IFNγ, IRF1, STAT1 et IL27. Chaque point correspond à l'expression relative d'UBD en ordonnée et d'un des quatre autres gènes de la voie de l'IFNγ en abscisse (*IFNγ*, *IRF1*, *STAT1* ou *IL27*). Le r^2 et la p-value correspondante sont indiquées pour chaque paire de gène et la droite de régression linéaire est tracée en rouge.

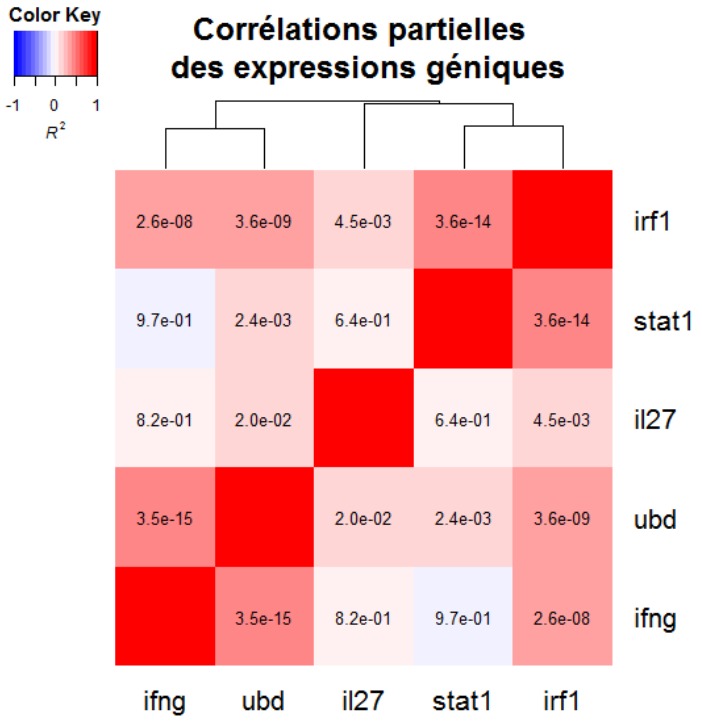


Figure 48. Matrice de corrélations partielles deux à deux des expressions des transcrits des cinq gènes d'intérêts, *UBD*, *IFNγ*, *IL27*, *IRF1* et *STAT1*. Carte de chaleur des coefficients de corrélations partielles de Pearson des gènes deux à deux selon la clé de couleur indiquée en haut à gauche. La valeur inscrite dans chaque case correspond à la valeur p. Une corrélation partielle entre deux gènes prend en compte les corrélations intermédiaires existantes avec les autres gènes. Par exemple si un gène A est corrélé à deux gènes B et C, la corrélation entre A et C a deux composantes : l'une dépendante de B et l'autre indépendante de B. La corrélation partielle entre A et C correspond à la composante indépendante de B.

2.4. Discussion et perspectives

Validation de la diminution d'expression des transcrits des cinq gènes chez les patients de la cohorte EUR

Les puces d'expression avaient permis de mettre en évidence une variation de nombreux gènes entre patients et contrôles avec un enrichissement significatif de gènes de la voie de l'interferon-γ, avec notamment une diminution chez les patients de l'expression de 4 gènes appartenant à cette voie (*IFNγ*, *STAT1*, *IRF1* et *IL27*). De plus neuf gènes du CMH étaient différentiellement exprimés, dont le gène *UBD*, un très bon candidat qui avait été associé à la prédisposition au DT1.

Nous avons tenté de valider et de répliquer par qPCR ces résultats pour ces cinq gènes d'intérêt sur les mêmes échantillons que ceux analysés avec la puce d'expression mais aussi sur de nouveaux sujets appartenant aux cohortes européenne et nord-américaine. Sur l'ensemble des échantillons de la cohorte européenne, nous avons retrouvé une différence significative d'expression pour ces cinq gènes, la variation se produisant dans la même direction. De plus, il existait une corrélation significative entre les quantifications individuelles d'expression obtenues par puce et par qPCR. Ceci valide complètement les résultats obtenus par puce.

Par contre, ces résultats n'ont pu être répliqués dans la cohorte nord-américaine. Différents facteurs pourraient l'expliquer. La cause majeure est probablement l'absence de stimulation des lignées lymphoblastoïdes nord-américaines par le PMA. En effet, dans la cohorte européenne, les différences d'expression sont visibles essentiellement après stimulation des échantillons. De ce fait, nous manquons de puissance pour détecter des variations d'expression sur les seuls

échantillons non stimulés, le nombre de personnes recrutées étant trop faible. Enfin, on pourrait invoquer les conditions de recrutement différentes entre les deux cohortes du T1DGC : l'âge moyen des individus de la cohorte américaine est un peu plus élevé, avec un traitement du diabète qui est déjà engagé depuis plus longtemps comparé à celui de la cohorte européenne.

Extension des résultats au niveau protéique

De façon remarquable, nous avons mis en évidence une corrélation entre les niveaux d'expression des transcrits et celui des protéines sécrétées pour l'IFNγ. Nous avons également retrouvés une différence d'expression de cette cytokine avec des patients qui la sécrètent moins que les contrôles. Ceci mérite d'être souligné car il existe rarement une corrélation entre la variation d'expression des transcrits d'un gène et celle de son produit protéique (Wu et al. 2013). De fait, pour l'IL27, nous n'avons pas retrouvé de corrélation avec les niveaux d'expression protéique et de transcrits, ni de différence des niveaux de protéines sécrétées entre patients et contrôles. Il est possible que ce soit le reflet d'un défaut de sécrétion protéique. Des modifications post-traductionnelles ou la durée de vie des protéines généralement plus longue que celle des messagers pourraient aussi être en cause. De plus, l'IL27 est une molécule hétérodimérique associant deux chaînes : le produit du gène *IL27* associé au produit de l'Epstein-Barr virus induced gene 3 (*EBI3*). Il est donc possible que l'expression de ce complexe mesuré par ELISA soit fortement modifiée par le niveau d'expression d'EBI3 ou par d'autres facteurs cellulaires ou biochimiques régulant leur association. Nous n'avons pas pu quantifier par ELISA les protéines codées par les trois autres gènes, car elles ne sont pas sécrétées. Il faudrait recourir à des méthodes d'immunophénotypage par cytométrie en flux qui s'adressent à des échantillons cellulaires fraîchement préparés, ce qui n'était pas réalisable dans cette cohorte du T1DGC.

Perspectives : RNASeq sur cellules primaires

Pour renforcer l'ensemble de ces résultats, la validation des puces d'expression et des qPCR est en cours par RNASeq sur des cellules mononucléées du sang périphérique de patients et de contrôles du T1DGC européen et nord-américain. Les résultats sont en cours d'analyse au laboratoire. La technologie RNASeq présente de nombreux avantages par rapport aux puces et à la qPCR. Elle permet, entre autres, une quantification « absolue » des transcrits et une caractérisation fine des formes épissées alternativement ainsi que la découverte de nouveaux transcrits. De plus cette étude porte sur des cellules primaires qui ne sont pas sujettes aux biais liés à l'immortalisation des cellules par le virus Epstein-Barr. En contrepartie, les cellules mononuclées représentent un mélange complexe de différents types cellulaires qui peut conduire à masquer des différences spécifiques de cellules. De plus, elles n'ont pas été stimulées.

Signification physiopathologique

Notre étude met en évidence une diminution chez les patients de l'expression de la voie de l'interferon-γ, et avant tout de l'IFNγ lui-même. Cette diminution est paradoxale puisque l'IFNγ est essentiellement connue comme une cytokine pro-inflammatoire et que le diabète de type 1 met en jeu un processus inflammatoire au niveau des îlots de Langerhans. Toutefois, il est à noter qu'une telle diminution a été mise en évidence dans une autre maladie de nature auto-immune et inflammatoire, la spondylarthrite ankylosante (Smith et al. 2008 ; Fert et al. 2014). En fait, il semble que l'INFγ puisse aussi avoir un effet anti-inflammatoire par la stimulation de l'expression

de gènes tels que l'antagoniste du récepteur de l'IL1 (*IL1RN*) ou l'IL18 binding protein (*IL18BP*) (Mühl et Pfeilschifter 2003 ; Kelchtermans, Billiau, et Matthys 2008). Il a été également montré que l'IFNγ stimule le développement de cellules T régulatrices capables de contrôler les cellules T CD4+ proinflammatoires (Hall et al. 2012). Il n'est donc pas si surprenant de constater une diminution de l'IFNγ dans une maladie inflammatoire comme le diabète de type 1. De plus, cette observation a peut-être été rendue possible par le « design » inédit de l'étude qui masque les effets des gènes de classe II du CMH.

Parmi les gènes testés, l'*IL27* présente un intérêt particulier puisque des polymorphismes de la région de ce gène ont été génétiquement associés au DT1 (**Tableau 4**). Il est donc possible que ces variants génétiques soient aussi responsables en partie de la diminution de l'expression du gène observée chez les patients mais cela reste à tester. Comme l'IFNγ, l'IL27 présente des propriétés tantôt pro-inflammatoires tantôt anti-inflammatoires (Hall et al. 2012). Sa diminution chez les patients diabétiques suggère que ce sont ses propriétés anti-inflammatoires qui sont mises en jeu. Ce serait donc une justification pour une immunothérapie reposant sur l'administration de cette cytokine (Meka et al. 2015 ; Zamani et al. 2015).

Le gène *UBD* quant à lui était un excellent candidat du fait de son appartenance au complexe majeur d'histocompatibilité dans la région de classe I étendue. C'était précisément l'objectif du «design» que d'identifier de tels gènes. Il code une protéine de type ubiquitine-like capable de se lier de façon covalente à une protéine cible et d'entrainer sa dégradation par le protéasome. Elle est notamment impliquée dans la mitose, l'apoptose et la réponse immunitaire. Sa surexpression serait impliquée dans la progression de nombreux cancers, tels que l'adénocarcinome du colon (Yan et al. 2010) ou l'hépatocarcinome (Lee et al. 2003). Il a été initialement impliqué dans le DT1 par des études d'association génétique visant à identifier des polymorphismes du CMH autres que les gènes *HLA* de classe II (Baschal et al. 2011 ; Aly et al. 2008). Dans un modèle de diabète viro-induit chez le rat, *UBD* apparait également comme un gène candidat (Blankenhorn et al. 2009). Sur le plan fonctionnel, aucune étude n'avait été conduite jusque très récemment avec la mise en évidence d'une action pro-apoptotique dans les cellules β des îlots de Langerhans (Brozzi et al. 2016). En revanche, la variation du niveau des transcrits de ce gène n'avait pas été étudiée chez des patients diabétiques avant notre travail. De plus, nous montrons que dans une population humaine l'expression de ce gène est fortement dépendante de l'expression de l'*IFNγ* et d'*IRF1*, le rattachant ainsi à la voie de signalisation de l'IFNγ. Nos résultats renforcent donc l'intérêt de ce gène comme candidat à la prédisposition au DT1.

Perspectives d'analyse

Afin d'approfondir ces résultats, il serait intéressant de relier les variations d'expression génique observés à la variation génétique grâce au génotypage des mêmes sujets qui a été réalisé par l'immunochip. Les eQTLs identifiés pourraient constituer de nouveaux variants candidats à la prédisposition au DT1, qu'il faudrait ensuite tester par association génétique. Cette stratégie serait à l'inverse de la démarche récemment appliquée au DT1, dans laquelle les variants connus pour être associés au DT1 ont été testés comme eQTLs (Ram et al. 2016). Enfin, il serait possible de tester les eQTLs identifiés dans des modèles multivariés comportant des termes d'interaction. L'analyse de variables quantitatives comme l'expression génique bénéficie en effet d'une puissance accrue comparée aux traits binaires comme le statut « patient-contrôle » pour identifier des termes d'interaction.

3. Etude génétique et fonctionnelle d'un gène candidat, *ZFP57*

ZFP57 est un facteur épigénétique de répression de la transcription qui agit au cours du développement embryonnaire principalement au niveau des régions soumises à empreinte parentale en maintenant la méthylation.
Le laboratoire s'est intéressé à ce gène pour plusieurs raisons (Cf. Introduction). Tout d'abord, il est situé au sein du Complexe Majeur d'Histocompatibilité (CMH), facteur de risque génétique majeur au DT1 et mais dont seule la moitié de la composante génétique est expliquée. La recherche de nouveaux gènes du CMH autres que les gènes de classe II *HLA-DRB1-DQB1* est un thème de recherche central du laboratoire. Ensuite, des mutations de *ZFP57* ont été impliquées dans des formes néonatales transitoires de diabète. Enfin, notre équipe avait mis en évidence que ce gène du développement était aussi exprimé au niveau transcriptionnel dans les cellules sanguines de volontaires sains et avait montré que pour son expression était significativement plus élevée chez des individus porteurs de l'haplotype porteur de l'allèle DR3 à risque de développer un diabète. Ainsi, nous avons émis l'hypothèse que l'expression de ZFP57 puisse être un facteur épigénétique important de la susceptibilité au DT1.
Notre but est donc de tester si des polymorphismes de *ZFP57* prédisposent au DT1 et d'élucider les mécanismes à travers lesquels la protéine codée par ce gène pourrait être impliquée.

3.1 Cartographier les variants causaux de *ZFP57* dans la cohorte de patients diabétiques de Lyon

Un premier travail réalisé dans le laboratoire sur le DT1 avait montré une association génétique très significative du DT1 avec deux SNPs de *ZFP57*, rs3870968 et rs3129054, dans la cohorte du consortium génétique international du diabète (T1DGC) ($p=1,4x10^{-11}$, OR=1,61 [IC95% : 1,39 ; 1,86] et $p=6,10^{-8}$, OR=1,43 [IC95% : 1,25 ; 1,64]). Ces deux SNPs sont indépendants l'un de l'autre et aussi des allèles HLA de classe II à risque. En effet, ils ne sont pas en déséquilibre de liaison avec DR3 ni avec à DR4-DQ8 ($r^2=0,004$ avec DR3 et $r^2=0,001$ avec DR4-DQ8 pour rs3870968; $r^2=0,000$ pour DR3 et DR4-DQ8 avec rs3129054) et, fait capital, les associations persistent après une analyse conditionnelle sur DR3 ($p=3x10^{-5}$, OR=2,04 [IC95% : 1,25 ; 3,33] et $p=7x10^{-5}$, OR=3,125 [IC95% : 1,92 ; 5]).
Mon premier objectif était donc de répliquer l'association génétique des polymorphismes de *ZFP57* au DT1 dans la cohorte française de Lyon (French GWAS) et d'étudier si ces mêmes variants pouvaient influencer l'expression de ZFP57.
Lorsque j'ai rejoint le laboratoire, la cohorte Lyon 1 (685 sujets) avait été génotypée pour près de 200 000 SNPs au moyen de l'immunochip mais le reste de la cohorte, Lyon 2 (2779 sujets) n'était pas encore caractérisé. Les résultats avaient montré des associations significatives ou suggestives pour 12 SNPs (cf. Projet « French-GWAS »). Concernant les polymorphismes du gène candidat *ZFP57,* cinq d'entre eux, rs2235383, rs7741807, rs7759272, rs9257932 et rs2076177, étaient associés au seuil de 5%. Nous avons donc entrepris de génotyper ces 5 SNPs dans la cohorte Lyon 2.
Par ailleurs, nous avions quantifié l'expression de *ZFP57* par qPCR dans la cohorte Lyon 1. Grâce au génotypage parallèle de cette cohorte au moyen de l'immunochip, nous avions conduit une analyse d'association pangénomique de cartographie d'expression de *ZFP57* de type eQTL (**cf. Introduction**). Une étude similaire avait été conduite dans une cohorte de volontaires sains à

Oxford génotypés avec une puce à plus haute densité. Ces deux études avaient identifié 9 SNPs situés dans *ZFP57* ou à proximité et fortement associés à la variation de son expression : rs396660, rs445150, rs375984, rs2535238, rs2747421, rs2747429, rs2747431, rs365052 et rs416568. Les 4 derniers n'étaient pas inclus dans l'immunochip. Enfin, un polymorphisme plus distant du gène, situé dans la région du CMH de classe II, et présent sur l'immunochip, rs2269422 avait aussi été associé à une modulation de l'expression de ZFP57.
Nous avons donc tenté de confirmer ces associations en génotypant à l'aide de sondes Taqman d'une part les 4 SNPs non inclus dans l'immunochip dans la cohorte Lyon 1, et d'autre part l'ensemble des 17 SNPs dans la cohorte Lyon 2 (**Figure 49**). Au total j'ai participé au génotypage de 13 de ces SNPs. Le reste du génotypage a été réalisé par des étudiantes de Master 2, Tharshana Stephen et Neda Rezaei.

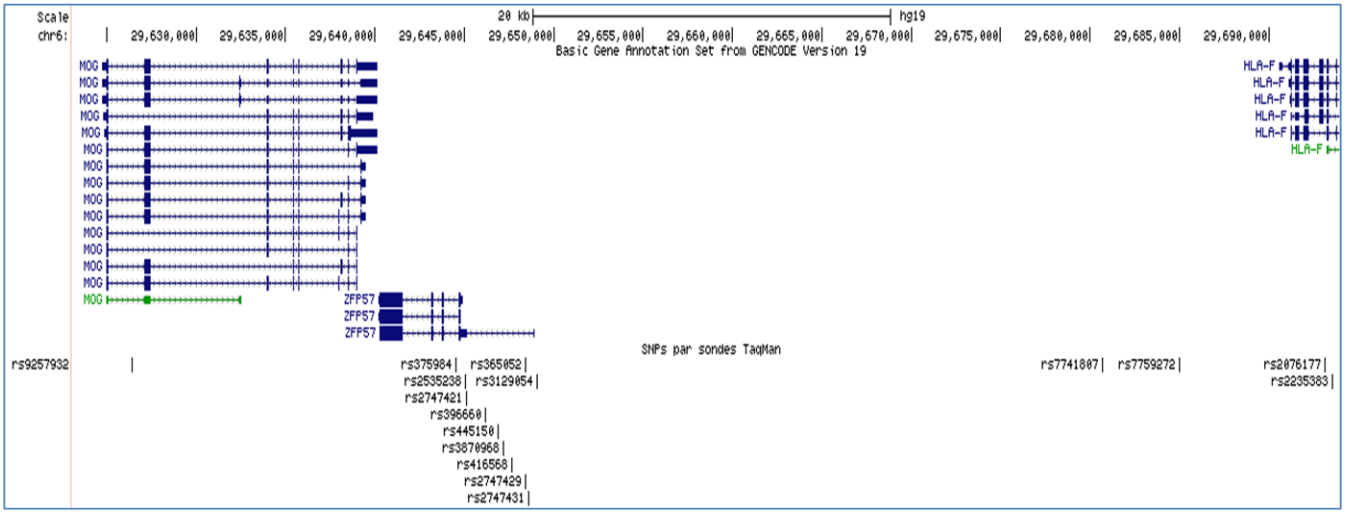


Figure 49. Localisation génomique du gène *ZFP57* et des SNPs génotypés en Taqman selon hg19/ncbi37. Le gène *ZFP57* est situé dans la région du CMH de classe I entre les gènes HLA-F et MOG. Les annotations des transcrits sont celles de la version 19 de ENCODE. Les rectangles bleus épais représentant les exons codants, les rectangles moins épais correspondent aux 5' et 3'UTR. Les flèches indiquent l'orientation de la phase de lecture. *ZFP57* est codé sur le brin moins. Les différents SNPs génotypés au moyen de sondes Taqman sont positionnés dans la piste inférieure (traits fins verticaux). Les coordonnées génomiques sont indiquée dans la piste supérieure selon la version hg19/NCBI 37 du génome humain. Le SNP rs2269422 est situé à plus de 2 Mégabases vers le centromère et ne figure pas sur le schéma.

Nous avons mené une analyse d'association familiale simple-point par déséquilibre de transmission (TDT) sur l'ensemble de la cohorte de Lyon (Lyon 1 et Lyon 2 combinées) afin d'avoir une puissance maximale de détecter des associations (**Tableau 14**).
Nous répliquons l'association initialement trouvée dans la cohorte du T1DGC pour les marqueurs rs3870968 et rs3129054 ($p=6,5x10^{-3}$ et $1,34x10^{-2}$ respectivement) avec des amplitudes comparables (OR= 2,18 [IC95% : 1,23 ; 3,87] pour l'allèle majeur de rs3870968 et OR= 1,31 [IC95% : 1,06 ; 1,62] pour l'allèle majeur de rs3129054). Le premier SNP, rs3870968, est situé dans le premier intron de *ZFP57*, le second est situé 169 bases en 5' du gène.
En revanche, nous ne répliquons au seuil 5% qu'un seul des cinq SNPs que nous avions précédemment détectés à ce même seuil dans la cohorte Lyon 1 : rs7741807.
Enfin, parmi les 10 SNPs connus pour être associés à une variation de l'expression de *ZFP57*, deux SNPs, rs396660 et rs445150, sont pour la première fois associés au DT1 au seuil 5%.

Afin de clarifier si ces associations sont indépendantes des facteurs de risque majeurs de classe II du CMH, nous avons entrepris une analyse conditionnelle de ces associations. Le principe est de détecter si une association persiste pour un SNP « test » donné lorsqu'on fixe l'allèle de risque du marqueur sur lequel on effectue la condition, c'est-à dire en comparant les haplotypes porteurs de l'allèle de risque au marqueur « conditionnel ». Cette analyse conditionnelle effectuée sur l'allèle de risque DR3 révèle une association de 10 des 17 SNPs testés de *ZFP57* (**Tableau 12**). L'association la plus significative est obtenue pour le SNP rs7741807 ($p=4,96x10^{-4}$) déjà détecté précédemment dans la cohorte Lyon 1 par une analyse similaire. Mais surtout, nous obtenons pour la première fois une association au DT1 de SNPs connus pour influencer l'expression de *ZFP57* et en déséquilibre de liaison modéré à fort, tel que rs396660 dont l'allèle mineur est protecteur avec un OR=0,416 [IC95% : 0,229 ; 0,758]. La même analyse conditionnée sur l'allèle de risque du meilleur SNP du CMH selon notre étude « French-GWAS », rs9273363*C, révèle une association de tous les SNPs de *ZFP57* ou voisins du gène qui étaient précédemment connus pour leur association à l'expression de *ZFP57*, en particulier pour l'allèle mineur C de rs2535238 (OR=2,12 [IC95% : 1,34 ; 3,35, $p=1,3x10^{-3}$] (**Tableau 12**). En revanche, aucune association n'est significative en conditionnant sur l'allèle DR4.

Comme nous avons répliqué l'association de *ZFP57* avec le DT1 dans cette cohorte, nous avons entrepris dans un second temps une cartographie de l'expression de *ZFP57* dans la cohorte Lyon 1 déjà quantifiée pour l'expression des messagers de *ZFP57*. Grâce au génotypage au moyen de sondes TaqMan, nous disposions de 4 SNPs supplémentaires. Trois d'entre eux se sont montrés plus fortement associés que les SNPs initiaux avec une p-value très significative de $3,74x10^{-17}$ pour le meilleur, rs2747431 (**Tableau 13**). Ce SNP est en fort déséquilibre de liaison ($r^2=0,899$) avec le meilleur SNP de *ZFP57* associé au DT1, rs396660. Tous les deux sont situés dans le 1^{er} intron, à 2,4 kb l'un de l'autre. Ainsi, nous obtenons une co-localisation de plusieurs SNPs associés au DT1 et à l'expression de *ZFP57*, dans le 1^{er} intron du gène, dans une région caractérisée par la présence de sites connus d'hypersensibilité à la DNAse.

Tableau 12. Résultats des analyses d'associations par TDT et des analyses conditionnelles sur DR3 et rs9273363*C des 17 SNPs sélectionnés sur l'ensemble de la cohorte Lyon 1 et Lyon 2. Pour chaque SNP sont indiqués la position génomique (hg19), l'allèle mineur chez le fondateurs, l'allèle majeur, le nombre de transmissions de l'allèle mineur et de non transmissions, l'odds ratio relatif à l'allèle mineur avec son intervalle de confiance à 95% et la p-value correspondante. L'odds ratio n'est indiqué que pour les SNPs pour lesquels la p-value était significative.

			Test de déséquilibre de Transmission non conditionné						Analyse conditionnée sur l'allèle DR3		Analyse conditionnée sur l'allèle associé C de rs9273363	
	SNP	Position en hg19	Allèle mineur	Allèle majeur	Transmis	Non transmis	OR [IC95%]	p-value	OR [IC95%]	p-value	OR [IC95%]	p-value
SNPs associés au DT1 dans le T1DGC, y compris après condition sur DR3 et DR4	rs3870968	29647148	A	C	17	37	0,460 [0,259 ; 0,816]	6,50E-3		NS		NS
	rs3129054	29649055	T	C	150	196	0,765 [00619 ; 0,947]	1,34E-2		NS		NS
SNPs précédemment associés au seuil 5% au DT1 dans la cohorte Lyon 1	rs2235383	29693498	C	T	77	100		NS		NS		NS
	rs7741807	29680656	C	G	12	24	0,5 [0,25 ; 0,9998]	4,55E-2	22,24 [2,55 ; 193,9]	4,96E-4		NS
	rs7759272	29684994	C	T	77	98		NS		NS		NS
	rs9257932	29626375	C	T	10	18		NS	14,43 [1,452 ; 143,3]	1,32E-2		NS
	rs2076177	29693112	T	C	75	92		NS		NS		NS
SNPs eQTLde *ZFP57* dans différentes cohortes (Oxford, T1DGC et/ou cohorte Lyon1)	rs396660	29646164	T	C	137	175	0,783 [0,626 ; 0,979]	3,15E-2	0,416 [0,229 ; 0,757]	4,43E-3	0,528 [0,343 ; 0,814]	3,32E-3
	rs445150	29646878	C	T	141	180	0,783 [0,628 ; 0,977]	2,95E-2	1,967 [1,093 ; 3,54]	2,58E-2	1,749 [1,141 ; 2,679]	9,54E-3
	rs375984	29644501	T	C	119	145		NS	0,496 [0,268 ; 00917]	2,75E-2	0,474 [0,298 ; 0,755]	1,49E-3
	rs2747429 par Taqman	29648376	C	T	121	134		NS	1,975 [1,026 ; 3,804]	4,36E-2	2,004 [1,195 ; 3,362]	7,68E-3
	rs2535238	29645037	A	C	126	146		NS	2,1 [1,158 ; 3,811]	1,60E-2	2,115 [1,337 ; 3,346]	1,23E-3
	rs2747421	29645117	C	G	123	143		NS	1,935 [1,065 ; 3,515]	3,27E-2	2,019 [1,277 ; 3,192]	2,45E-3
	rs2747431 par Taqman	29648563	T	C	130	162		NS	0,473 [0,245 ; 0,911]	2,63E-2	0,541 [0,332 ; 0,883]	1,28E-2
	rs365052 par Taqman	29648397	C	G	101	118		NS		NS	1,863 [1,008 ; 3,444]	4,29E-2
	rs416568 par Taqman	29647627	A	T	133	157		NS	2,19 [1,132 ; 4,236]	2,08E-2	1,824 [1,115 ; 2,983]	1,56E-2
	rs2269422	32151293	C	T	3	5		NS		NS		NS

Tableau 13. Résultats de l'analyse d'association à l'expression de *ZFP57* dans la cohorte Lyon 1 pour les 17 SNPs sélectionnés. L'association a été testée par régression linaire en fonction de l'allèle mineur selon un mode additif. Pour chaque SNP sont indiqués la position génomique (hg19), la pente β conférée par l'allèle mineur avec son intervalle de confiance à 95% et la p-value correspondante.

			eQTL sur *ZFP57*	
	SNP	**Position en hg19**	**β [IC95%]**	**p-value**
SNPs associés au DT1 dans le T1DGC, y compris après condition sur DR3 et DR4	rs3870968	29647148	72,4 [30,37 ; 114,4]	8,69E-4
	rs3129054	29649055	-18,21 [-34,97 ; -1,464]	0,03418
SNPs précédemment associés au seuil 5% au DT1 dans la cohorte Lyon 1	rs2235383	29693498	8,183 [-14,79 ; 31,16]	1,96E-11
	rs7741807	29680656	8,183 [-14,79 ; 31,16]	0,4858
	rs7759272	29684994	8,781 [-14,1 ; 31,66]	0,4528
	rs9257932	29626375	14,62 [-30,43 ; 59,66]	0,5254
	rs2076177	29693112	5,743 [-17,12 ; 28,61]	0,6231
SNPs eQTL de *ZFP57* dans différentes cohortes (Oxford, T1DGC et/ou Cohorte Lyon1)	rs396660	29646164	73,04 [55,21 ; 90,88]	6,043E-14
	rs445150	29646878	74,31 [56,49 ; 92,13]	2,488E-14
	rs375984	29644501	72,52 [52,39 ; 92,65]	2,145E-11
	rs2747429 par Taqman	29648376	76,43 [55,97 ; 96,88]	5,12E-12
	rs2535238	29645037	72,44 [52,38 ; 92,5]	1,96E-11
	rs2747421	29645117	72,52 [52,39 ; 92,65]	2,145E-11
	rs2747431 par Taqman	29648563	83,11 [65,38 ; 100,8]	3,74E-17
	rs365052 par Taqman	29648397	82,55 [64,73 ; 100,4]	8,397E-17
	rs416568 par Taqman	29647627	82,97 [64,78 ; 101,2]	1,908E-16
	rs2269422	32151293	286,6 [171,8 ; 401,4]	1,92E-6

3.2 Localisation de la protéine dans la cellule

En vue de préciser les mécanismes moléculaires et cellulaires reliant ZFP57 au DT1, il devenait essentiel de caractériser le produit protéique du gène. Mon objectif était donc de mettre en évidence une expression protéique de ZFP57, ce qui à ce jour, n'a encore jamais été réalisé chez l'homme, que ce soit par western blot (WB), par cytométrie en flux ou par immunoprécipitation de la chromatine (ChIP).

3.2.1 Développement d'un anticorps anti-ZFP57 dirigé contre la protéine humaine

Il était tout d'abord primordial de disposer d'un anticorps spécifique dirigé contre la protéine humaine. La mise au point a été réalisée en binôme avec une étudiante de Master 2, Tharshana Stephen.

Pour rappel, ZFP57 présente 3 isoformes protéiques connues. Chacune possède 7 motifs à doigts de zinc permettant de se fixer à l'ADN et un domaine KRAB N-terminal impliqué dans les interactions avec d'autres protéines (**Figure 50**). Chez la souris, il a été montré que ZFP57 interagit au niveau de son domaine KRAB avec la protéine KAP1 afin de recruter une DNA Methyl Transferase.

Au démarrage du projet, il n'existait qu'un seul anticorps anti-ZFP57 humain commercialisé par Abcam (nommé A_1). Cependant, il ne ciblait parfaitement que la région N-terminale de l'isoforme 2, plus précisément ses cinquante premiers acides aminés incluant le domaine KRAB (**Figure 50**). Or il était absolument nécessaire d'avoir un anticorps qui cible toutes les isoformes hors de ses domaines fonctionnels. De ce fait, le laboratoire a fait produire par ProtéoGenix des anticorps polyclonaux chez quatre lapins (**Figure 50**). Deux anticorps, nommés B et C, ciblent chacun deux peptides entre les domaines en doigt de zinc 4 et 5. Cette région centrale est homologue à la région ciblée chez la souris par un anticorps ayant permis d'immunoprécipiter ZFP57 lié à la chromatine dans des cellules souches embryonnaires de souris (Quenneville et al. 2011). Deux autres anticorps, nommés D et E, ciblent chacun deux peptides de la région C-terminale. Entre temps, un second anticorps a été commercialisé par Abcam également, nommé A_2, ciblant un autre peptide entre les doigts de zinc 4 et 5. Nous avions donc six anticorps à tester.

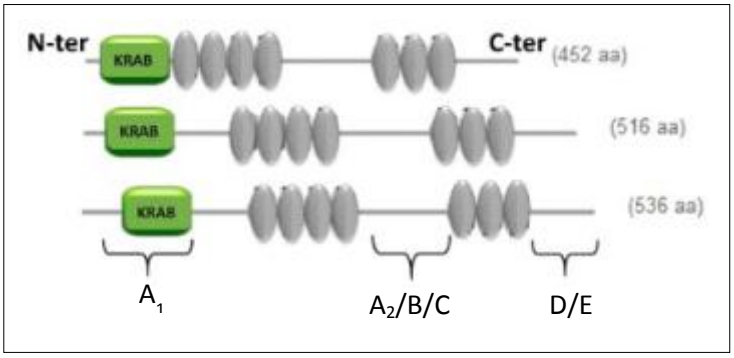


Figure 50. Structure protéique en 3 isoformes de ZFP57 et régions ciblées par les anticorps. Les isoformes 1 à 3 de ZFP57 selon SwissProt (Q9NU63-1, -2 et -3) sont représentées de haut en bas et ont une masse moléculaire de 52, 59 et 61 kDa respectivement. La protéine humaine comprend un domaine KRAB et 7 domaines en doigt de Zinc. L'anticorps commercial A_1 cible la région N-terminale de l'isoforme 2 exclusivement (acides aminés1 à 50). Les autres anticorps testés ciblent les 3 isoformes. L'anticorps commercial A_2 et les anticorps B et C développés par le laboratoire ciblent la région centrale entre les doigts de Zinc 4 et 5 (selon la séquence de l'isoforme 1 : acides aminés 268 à 296 pour A_2, acides aminés 198 à 207 et 247 à 259 pour B et C). Les anticorps D et E développés par le laboratoire ciblent la région C-terminale (acides aminés 415 à 428 et 437 à 452).

Pour tester ces différents anticorps, nous avons choisi de travailler avec quatre lignées lymphoblastoïdes particulières très étudiées au laboratoire pour leur homozygotie tout au long du CMH (Horton et al. 2008 ; Stewart et al. 2004 ; Vandiedonck et al. 2011). Deux de ces lignées expriment *ZFP57* au niveau transcriptionnel (lignées COX et APD) avec des niveaux très élevés pour COX, et intermédiaires pour APD, tandis que les deux autres l'expriment quasiment pas (lignées PGF et QBL). Par exemple, COX exprime 900 fois plus d'ARNm de *ZFP57* que PGF dans nos expériences de RT-qPCR. Il est à noter qu'une seule isoforme d'ARNm est véritablement détectée, celle codant l'isoforme protéique principale de 52 kDa.

Tharshana Stpehen a testé ces différents anticorps par western blot. Aucune protéine n'a été détectée avec les anticorps A_1, A_2, B ou C. Par contre les anticorps D et E ont révélé une bande à 52 kDa, qui semble spécifique car absente des contrôles négatifs (sans anticorps ou avec un isotype contrôle IgG). La bande détectée était très fortement présente dans la lignée COX comparé à la lignée PGF en accord avec les résultats de quantification des transcrits. L'anticorps D a finalement été choisi et utilisé dans la suite du projet.

J'ai réitéré les résultats obtenus par la Tharshana Stephen (**Figure 51**). Pour cela j'ai cultivé les quatre lignées cellulaires, extraits les protéines du cytoplasme et du noyau et réalisé un western blot avec l'anticorps D. La protéine est détectée à 52 kDa conformément à ce qui est attendu pour l'isoforme 1 de la protéine codée par le transcrit détecté en qPCR. De plus, elle est localisée uniquement dans le noyau, ce qui est cohérent avec la fonction attendue de la protéine.

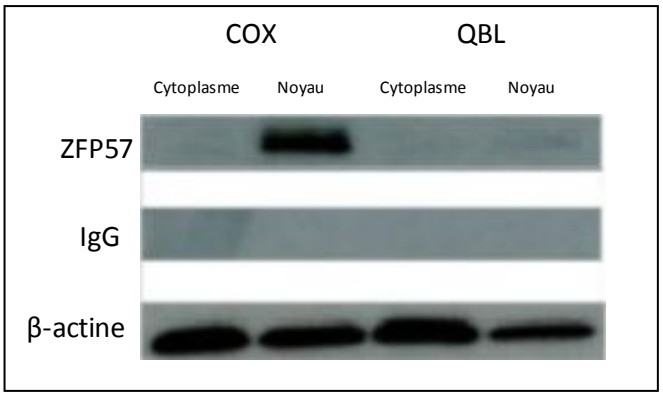


Figure 51. Western blot de l'expression de ZFP57 dans les fractions cytoplasmique et nucléaire des lignées COX et QBL, contrôles positif et négatif de l'expression de ZFP57. 5 µg de chaque extrait cytoplasmique et nucléaire ont été analysés par western blot en utilisant des anticorps dilués au 1:1000 (ZFP57 anticorps D), 1:6000 (IgG) ou 1:10000 (β -actine). Un anticorps de chèvre anti-IgG de lapin dilué au 1:10000 a été utilisé comme anticorps secondaire.

3.2.2 Spécificité de l'anticorps D anti-ZFP57

Après avoir mis au point l'anticorps, il était nécessaire de vérifier la spécificité de l'anticorps. Cela été fait dans la fraction nucléaire de COX. Deux approches ont été envisagées pour établir cette spécificité:

- une compétition avec les deux peptides de la région C-terminale utilisés pour la production de l'anticorps D
- une immunoprécipitation suivie d'une migration sur gel et coloration au bleu de Coomassie suivie de spectrométrie de masse

Pour la compétition de peptides, j'ai réalisé un western blot sur les protéines nucléaires de la lignée COX en utilisant séparément les deux peptides contre lesquels était ciblé l'anticorps. Des quantités croissantes (0,001 à 10 µg) de ces peptides ont été testées. Un seul des peptides, le peptide correspondant aux acides aminés C-terminaux 437 à 452, permet de faire disparaître la bande à 52 kDa qui est donc spécifique de la protéine ZFP57 (**Figure 52**).

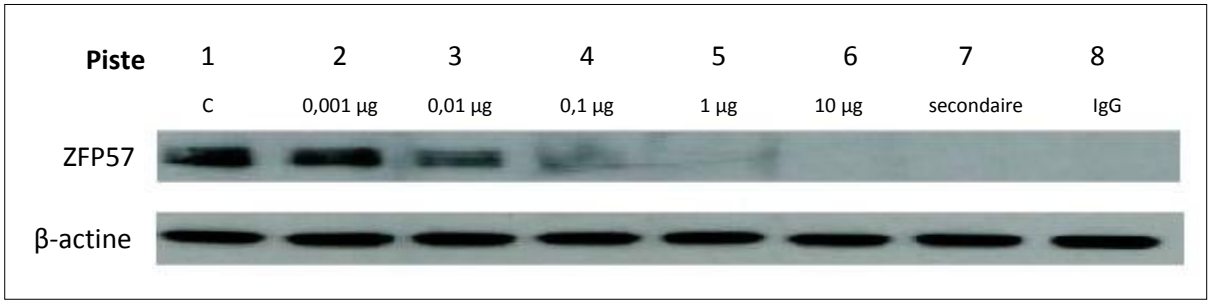


Figure 52. Compétition avec les peptides dans la fraction nucléaire de la lignée COX. Western blot avec l'anticorps primaire D anti-ZFP57 (pistes 1 à 6), sans anticorps primaire (piste 7) et un isotype contrôle (piste 8). La piste 1 ne contient pas de peptide en compétition (contrôle positif). Les bandes ont été découpées et incubées en présence de l'anticorps D avec des quantités croissantes de peptide : 0.001 µg, 0.01 µg, 0.1 µg, 1 µg et 10 µg pour les pistes 2, 3, 4, 5 et 6 respectivement.

En vue d'une expérience de spectrométrie de masse en tandem, j'ai également réalisé une immunoprécipitation à l'aide de l'anticorps D anti-ZFP57. La bande spécifique a été analysée par spectrométrie de masse en tandem à la plate-forme de spectrométrie de masse de l'Institut Jacques Monod. Cependant, aucun peptide de ZFP57 n'a pu être détecté. Il est possible que les concentrations protéiques aient été un peu faibles. De plus, il semble que la séquence protéique de ZFP57 la rende difficile à détecter par spectrométrie de masse.
Néanmoins, l'expérience de compétition précédemment présentée nous semble suffisamment convaincante pour attester de la spécificité de l'anticorps D vis-à-vis de ZFP57 et nous avons donc poursuivi l'étude de l'expression protéique au niveau cellulaire.

3.2.3 Localisation de la protéine par fractionnement cellulaire

Deux approches ont été utilisées pour localiser plus finement la protéine : le fractionnement cellulaire suivi d'un western blot et la microscopie confocale.
Le fractionnement cellulaire a été réalisé sur les quatre lignées cellulaires COX, PGF, QBL et APD et suivi d'un western blot anti-ZFP57. Des anticorps reconnaissant spécifiquement des protéines de chaque fraction ont été inclus comme témoin de charge. Les résultats montrent une localisation de ZFP57 au niveau du noyau et de la chromatine de COX, ce qui semble également cohérent avec la fonction attendue de la protéine (**Figure 53**).

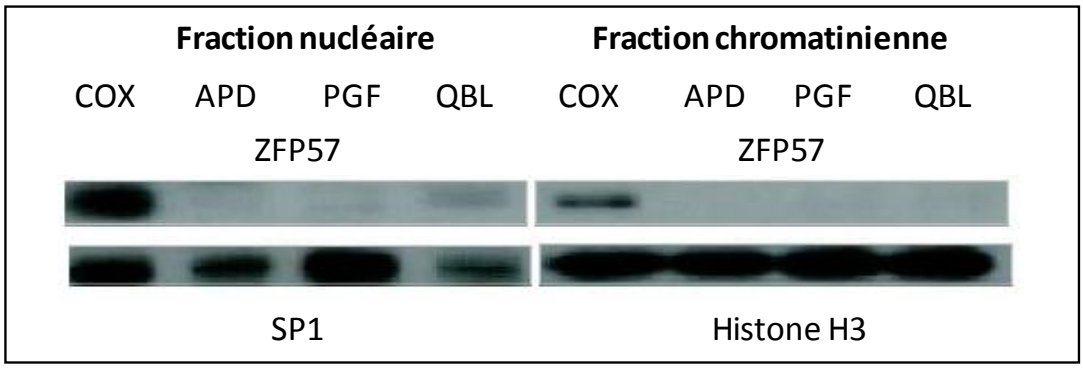


Figure 53. Western blot anti-ZFP57 dans les fractions du noyau et de la chromatine dans les 4 lignées lymphoblastoïdes. Le facteur de transcription SP1 et l'histone H3 ont été utilisées comme contrôle de charge pour la fraction du noyau et de la chromatine respectivement.

3.2.4 Localisation de la protéine par microscopie confocale

Une analyse par microscopie confocale a été également conduite sur les quatre lignées cellulaires. Après fixation et perméabilisation, un double marquage dirigé contre les lamines qui bordent l'enveloppe nucléaire du côté du nucléoplasme et contre ZFP57 a été effectué. L'analyse au microscope confocal a été effectuée par Marc Clément, dans l'unité 1148 (équipe d'Antonino Nicoletti) à l'hôpital Bichat.
Les résultats montrent une co-localisation de ZFP57 avec les lamines, ce qui indique une localisation péri-nucléaire de la ZFP57 (**Figure 54**). Ces résultats affinent donc ceux obtenus par fractionnement cellulaire et western blot.

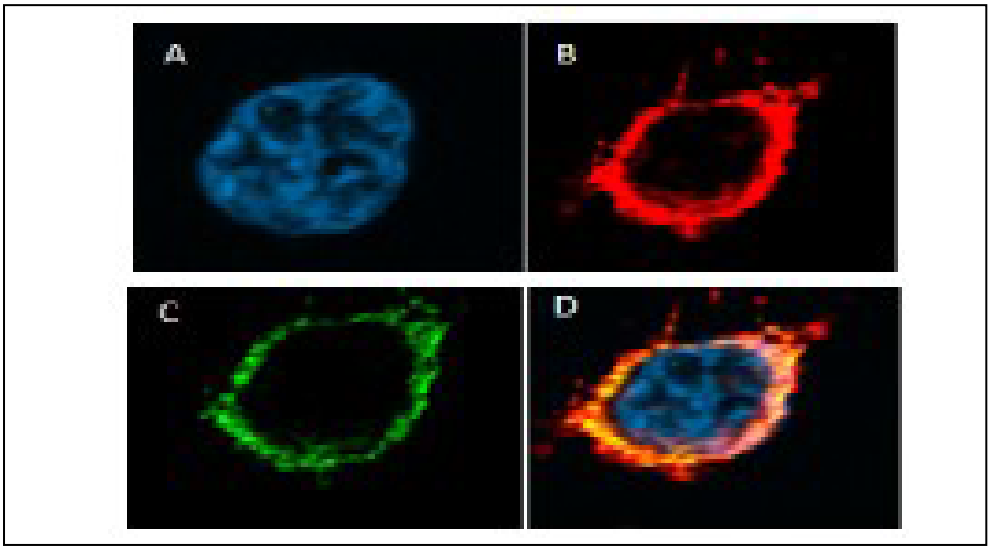


Figure 54. Microscopie confocale dans la lignée COX révélant une localisation périnucléaire de ZFP57. (A) DAPI (bleu); (B) lamines (rouge); (C) ZFP57 (vert) ; (D) superposition de B et C.

3.2.5. Mise en évidence de l'expression protéique de ZFP57 dans des cellules primaires du sang de sujets volontaires sains adultes

Tharshana Stephen avait procédé au test des différents anticorps en cytométrie en flux avec perméabilisation nucléaire sur les lignées COX et PGF. L'anticorps D s'était révélé le meilleur avec une différence d'expression maximale de ZFP57 entre COX (67,8% des cellules l'expriment) et PGF (18,6%) (**Figure 55**).

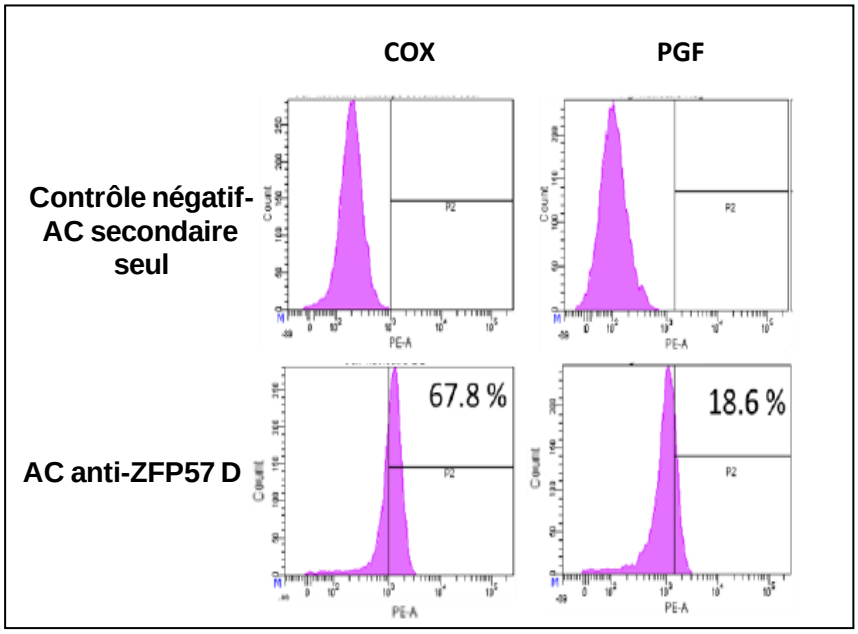


Figure 55. Cytométrie en flux réalisée sur les lignées COX (à gauche) et PGF (à droite). Chaque lignée a été fixée, perméabilisée au niveau nucléaire puis marqué à l'aide de l'anticorps D (2µg/mL).

A l'aide de cet anticorps, nous avons entrepris de caractériser les populations sanguines qui expriment ZFP57 chez des sujets sains. Le laboratoire a eu accès à du sang de volontaires sains appartenant à la cohorte CosImmGen (CIG) de la plateforme ICAREB de l'Institut Pasteur (Dr

Marie-Noëlle Ungeheuer). L'avantage de cette cohorte est qu'il est possible de rappeler les participants pour de nouveaux prélèvements. Tharshana Stephen avait génotypé les sujets de la cohorte pour les variants associés à l'expression de *ZFP57*. Nous en avons sélectionné 7 pour rappel et prélèvement des leucocytes circulants. Parmi eux, 1 seul est homozygote pour l'haplotype rare « a » associé à l'expression du transcrit de ZFP57, 2 sont homozygotes pour l'haplotype fréquent « A », et 4 sont hétérozygotes. Tous les porteurs de l'haplotype « a » ont une expression transcriptionnelle du gène (**Figure 56**).

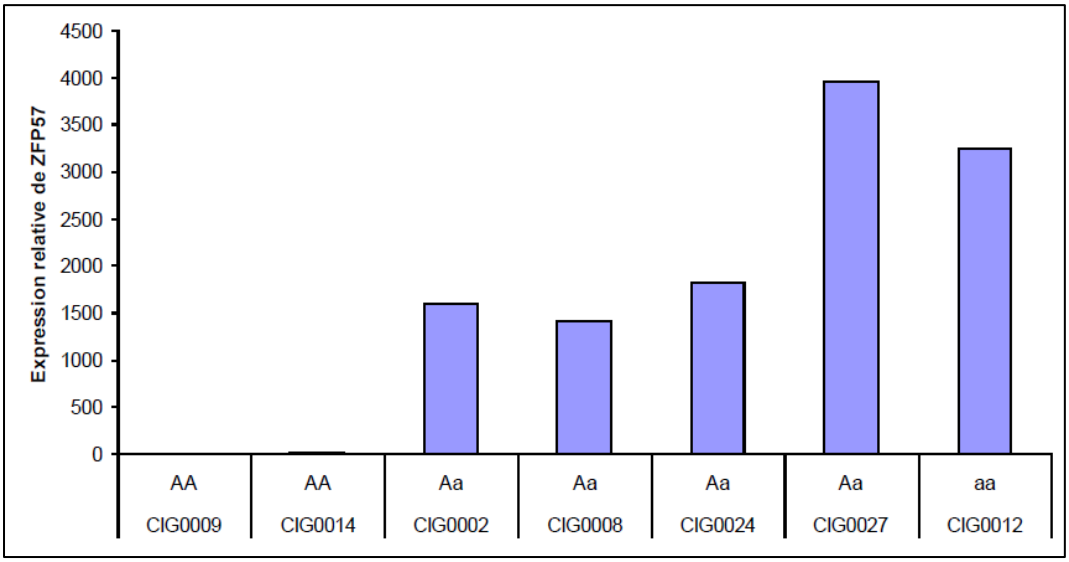


Figure 56. Expression relative des transcrits de *ZFP57* dans les leucocytes des 7 volontaires CIG inclus dans l'étude de cytométrie en flux. Les sujets (identifiés par leur code CIG) sont classés en fonction de leur génotype pour les variants associés à l'expression des messagers de *ZFP57*.

La stratégie a tout d'abord consisté à identifier les différents groupes cellulaires selon leur taille et granulosité, et ainsi exclure les débris cellulaire et les doublets de cellule (**Figure 57**). Puis, dans chaque groupe cellulaire, une fenêtre électronique a été définie pour chaque type cellulaire exprimant le marqueur de surface lui correspondant : lymphocytes T CD8+, lymphocytes T CD4+, lymphocytes B CD19+, et monocytes CD14+. Pour ces derniers, en plus de la population classique monocytaire qui exprime fortement CD14, une seconde population de cellules moyennes (CD14 moyen) a été identifiée chez certains individus. Cette population minoritaire de monocytes, généralement CD16+, est considérée comme à fort potentiel inflammatoire. Enfin, on identifie le groupe des granulocytes pour lequel nous n'avons pas de marqueur de surface.

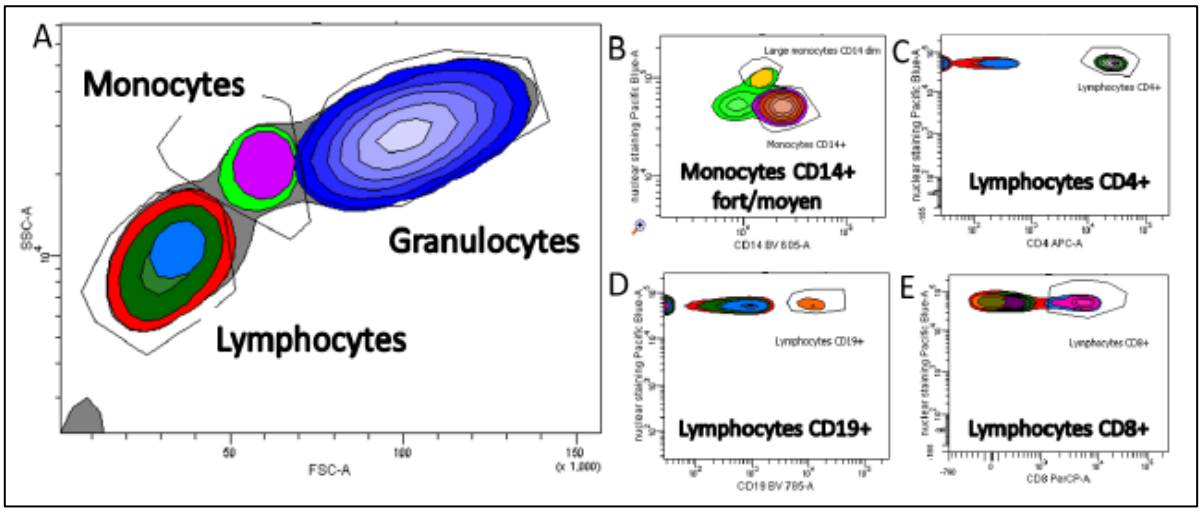


Figure 57. Cytométrie en flux, stratégie d'isolement des groupes cellulaires en fonction de la taille et de la granulométrie (A), puis de marquages avec des anticorps spécifiques (B-E).

Les résultats ont été analysés par Marc Clément. On observe une expression de ZFP57 dans tous les types cellulaires présents, que l'on considère le pourcentage de cellules positives ou les moyennes des intensités de fluorescence (**Figure 58**). Ces deux mesures quantitatives de l'expression protéique de ZFP57 sont d'ailleurs bien corrélées (coefficient de corrélation de Spearman = 0,65, $p=2,8x10^{-6}$).

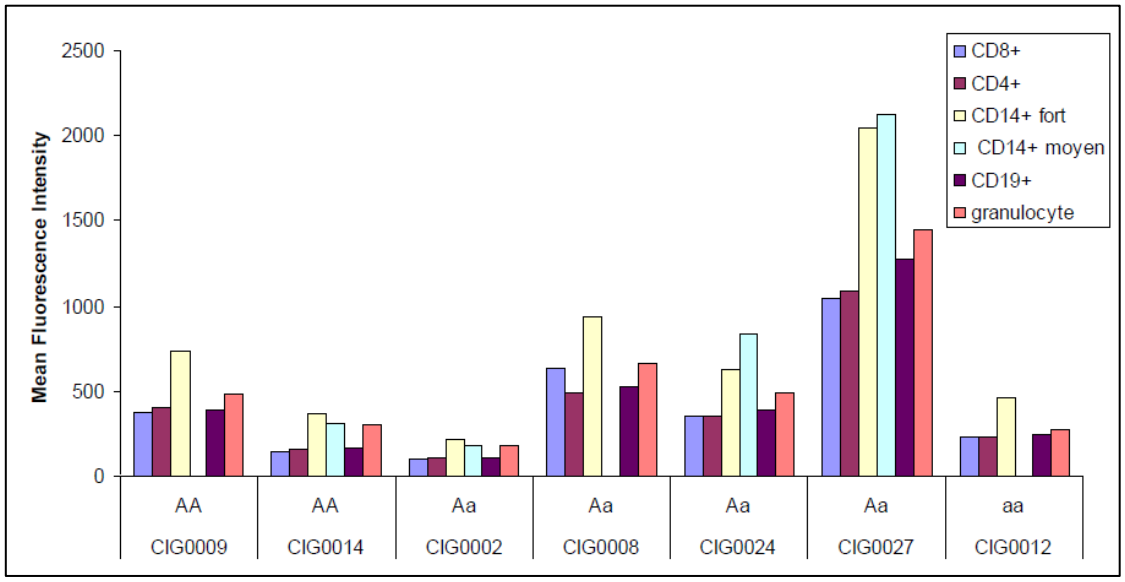


Figure 58. Moyenne de l'intensité de fluorescence par type cellulaire et par individu. Les sujets sont classés en fonction de leur génotype pour les allèles associés à l'expression du messager.

Parmi les 7 individus étudiés, le sujet CIG0027 est celui qui exprime le plus largement ZFP57 au niveau protéique quel que soit le type cellulaire (6,05% des cellules versus <0,6 pour tous les autres individus) mais plus largement dans ses monocytes CD14+ (47,1%), ses CD19+ (42,6%) et ses CD8+ (40,7%). C'était aussi celui qui exprimait le plus le messager du gène. Cependant, on ne met pas en évidence de corrélation entre les niveaux d'expression des transcrits et ceux des protéines (tests de corrélation Spearman dans chaque type cellulaire non significatifs). Ainsi, nous mettons pour la première fois en évidence l'expression protéique de ce gène dans plusieurs populations cellulaires du système immunitaire de l'adulte sain.

3.3. Discussion et perspectives

Ce travail renforce l'intérêt de *ZFP57* comme nouveau gène candidat dans la prédisposition au DT1 au sein du CMH.

Mise en évidence d'une association génétique de polymorphismes de *ZFP57* avec le DT1

En premier lieu, nous mettons en évidence une association génétique au DT1 de variants du gène *ZFP57*. Nous avions déjà détecté une telle association sur les échantillons de la cohorte EUR du T1DGC pour deux SNPs rs3870968 et rs3129054. Ces deux associations n'avaient pu être répliquées dans la seule cohorte Lyon 1. En revanche, en effectuant cette fois l'analyse d'association familiale sur la cohorte combinée incluant les cohortes Lyon 1 et Lyon 2, nous répliquons ces associations ($p=6,5x10^{-3}$ et $1,34x10^{-2}$). La taille de notre échantillon au regard de celle du T1DGC explique le manque de puissance rencontré pour atteindre le seuil de significativité pangénomique de 5.10^{-8}. Au seuil 5%, nous identifions également l'association au DT1 de deux nouveaux SNPs de *ZFP57*. Ces deux SNPs, rs396660 et rs445150 étaient jusqu'à cette étude les deux meilleurs SNPs associés à la variation de l'expression de *ZFP57* dans les leucocytes de la cohorte Lyon 1. C'est la première fois que nous parvenons à associer un même variant de *ZFP57* au DT1 et à un eQTL (expression quantitative trait locus) de *ZFP57*.

Le CMH étant caractérisé par un déséquilibre de liaison très marqué et étendu, il était impératif de s'assurer que les associations détectées n'étaient pas la résultante d'un déséquilibre de liaison entre ces variants et les facteurs de risque majeurs du CMH, *HLA-DR3*, *HLA-DR4* et l'allèle C du SNP rs9273363 que nous avons identifié dans notre projet « French-GWAS » comme étant le mieux associé de l'ensemble du CMH. Or, pour aucun des variants nous ne détectons de déséquilibre de liaison avec ces facteurs de risque, puisque la valeur r^2 estimant le DL demeure toujours inférieure à 1% (r^2 entre 0,0089 et 0,0000026). Ainsi, ces associations sont potentiellement indépendantes des DR3, DR4 et rs9273363.

Plus encore, après analyse conditionnelle sur l'allèle DR3, ces associations persistent et gagnent en signification jusqu'à $p=4,4x10^{-3}$ pour rs396660. Avec cette analyse, nous répliquons aussi l'association d'un autre SNP précédemment associé dans la cohorte Lyon 1. Nous identifions également l'association de six SNPs supplémentaires précédemment rapportés comme influençant l'expression de ZFP57 dans différentes cohortes (lignées lymphoblastoïdes du T1DGC, cellules mononuclées du sang périphérique de volontaires sains à Oxford, leucocytes de la cohorte Lyon 1) ou dans des bases de données d'eQTLs comme GTEx (gtexportal.org). En conditionnant sur l'allèle C du SNP rs9273363 cette fois, nous détectons même une association au DT1 pour la totalité des 9 eQTLs situés dans ZFP57 ou voisins de ZFP57. En parallèle, et grâce au génotypage par sondes Taqman de la cohorte Lyon 1 pour 4 des eQTLs absents de l'immunochip, nous retrouvons la totalité de ces eQTLs dans la cohorte Lyon 1 de manière très significative (p-value de $1,96x10^{-11}$ à $3,74x10^{-17}$), le mieux associé étant le SNP rs2747431.

En revanche, en conditionnant sur l'allèle DR4, aucune association n'est significative. Pour autant, cela ne signifie pas que ces associations sont le reflet de l'association de DR4, puisque aucun déséquilibre de liaison n'est observé avec DR4. Les variants associés n'étant pas non plus en déséquilibre de liaison avec DR3, les allèles associés peuvent donc être présents à la fois sur les haplotypes porteurs DR3 et sur les haplotypes porteurs de DR4. Ainsi, le fait que les associations soient uniquement visibles lorsque l'analyse est conditionnée sur DR3 indique que l'effet pourrait n'être présent que sur les haplotypes $DR3^+$. Ceci suggère fortement un mécanisme épistasique en

cis. Pour la première fois, ce travail permet donc de co-localiser des variants de *ZFP57* associés au DT1 et à l'expression de *ZFP57* et suggère qu'il s'agit d'associations en cis sur les haplotypes porteurs de DR3.
La difficulté à présent est d'identifier le variant causal. Dans ce but, nous espérons gagner en clarté après l'analyse complète des échantillons du T1DGC. J'ai complété pour cette cohorte la quantification de l'expression de *ZFP57* chez des sujets européens que nous n'avions pas encore testés et surtout chez tous les sujets nord-américains du projet « T1DGC-Express ». L'analyse de leur expression combinée à celle du récent génotypage de la cohorte par la puce immunochip est en cours. Nous espérons préciser les meilleures associations de *ZFP57* au DT1 et à son expression.
Toutefois nous avons constaté que selon la cohorte testée, ce n'est pas le même SNP qui présente la meilleure association à l'expression de *ZFP57*. Cela peut dépendre des SNPs réellement génotypés, comme nous l'avons vu pour la cohorte de Lyon dans laquelle le nouveau génotypage de 4 SNPs a permis d'identifier trois SNPs mieux associés que précédemment. Les fluctuations d'échantillonnage peuvent aussi être invoquées, de même que le type d'analyse paramétrique ou non paramétrique. Mais il est aussi essentiel de rappeler que dans des analyses d'eQTL ou de transcriptome, le contexte cellulaire et dynamique est essentiel. Un même variant peut être un eQTL dans un type cellulaire donné et avoir un effet nul voire opposé dans un autre type cellulaire. Concernant *ZFP57*, bien que des eQTLs aient été trouvés dans plusieurs types cellulaires à ce jour, aucune évidence d'eQTL spécifique d'une cellule n'a été relevée. Il n'est pas exclu que plusieurs variants puissent avoir un rôle indépendant ou en épistasie sur l'expression de *ZFP57*, selon qu'ils modifient par exemple un site consensus de liaison pour des facteurs de transcription de *ZFP57*. Plusieurs sites d'hypersensibilité à la DNAse ont été identifiés par le projet ENCODE ou nous-mêmes (données non publiées) dans le 1er intron de *ZFP57* où se concentrent ces eQTLs.
Un autre point que nous devons comprendre est que selon le variant associé au DT1, l'allèle de risque peut-être tantôt associé à une diminution de l'expression de *ZFP57*, tantôt à une augmentation. Une analyse haplotypique, c'est-à-dire prenant en compte l'ensemble des SNPs dans leur contexte chromosomique, devrait aider à mieux comprendre ces fluctuations.
Ainsi l'ensemble de ces résultats connecte pour la première fois la variation génétique de *ZFP57*, celle de son expression et la susceptibilité au DT1.

Mise en évidence d'une expression protéique de ZFP57 dans des cellules du système immunitaire de l'adulte sain

Compte-tenu du rôle connu de la protéine ZFP57 dans le contrôle épigénétique de l'expression génique au sein des cellules embryonnaires, il devenait indispensable de démontrer l'existence de la protéine dans le système immunitaire adulte pour étayer le rôle causal de *ZFP57* dans la prédisposition au DT1. En effet, un mécanisme purement épigénétique dès les premiers stades du développement embryonnaire pourrait certes se manifester après la naissance, mais cela parait peu probable.
Dans cette perspective, un anticorps polyclonal spécifique a été produit chez des lapins par immunisation avec des peptides dérivés de la séquence protéique de ZFP57. Il est intéressant de noter que les épitopes associés à l'extrémité C-terminale sont apparus comme les plus immunogènes, ce qui suggère une relative accessibilité de cette extrémité comparée aux autres régions testées. C'est sur cette base que nous essayons à présent de développer un anticorps monoclonal.

Cet anticorps polyclonal a pu être utilisé aussi bien en western blot qu'en cytométrie en flux ou de microscopie confocale. Ces différentes méthodes nous ont permis de mettre en évidence pour la première fois l'expression de la protéine dans des cellules humaines. Dans des lignées lymphoblastoïdes, le produit détecté présente un poids moléculaire correspondant à l'isoforme principale attendue à 52 KDa. Par cytométrie en flux, nous avons montré que ZFP57 était exprimé dans les principales populations leucocytaires du sang circulant. Une variation inter-individuelle a pu être notée mais elle n'est pas corrélée avec celle connue pour les messagers et qui dépend étroitement du polymorphisme génétique. La petite taille de notre échantillon ne permet pas de conclure de façon définitive à cette absence de corrélation. Toutefois, comme nous l'avons déjà évoqué pour le projet « T1DGC-Express », seule une petite fraction des eQTLs expliquent aussi des variations d'expression protéique car il n'existe que rarement une corrélation entre les niveaux de transcrits et ceux des protéines, compte-tenu des demi-vies différentes de ces espèces moléculaires, des modifications post-traductionnelles, ou des partenaires de ces protéines tels que KAP1 pour ZFP57. Ces différentes raisons peuvent aussi expliquer pourquoi nous n'avons pu détecter la protéine ZFP57 dans la lignée lymphoblastoïde APD bien qu'elle exprime *ZFP57* comme COX au niveau transcriptionnel. Si COX et APD partagent les mêmes génotypes pour les SNPs eQTL connus de *ZFP57*, leurs haplotypes du CMH sont toutefois différents (*HLA-A*01,-*08,-DRB1*03* pour COX et *HLA-A*01,-B*40,-DRB1*13* pour APD) et il n'est pas exclu que des allèles particuliers de ces haplotypes puissent être associés à une variation d'expression protéique.
La localisation nucléaire de la protéine ZFP57 a été authentifiée par fractionnement cellulaire suivi de western-blot et par microspcopie confocale. Le fractionnement cellulaire indique de plus que la protéine peut être associée à la chromatine, tandis que la microscopie confocale pointe sur une localisation péri-nucléaire correspondant généralement à la chromatine transcriptionnellement active. Ceci suggère que ZFP57 puisse jouer un rôle régulateur dans l'expression génique comme cela est établi dans l'embryon chez la souris.

Perspectives

En conclusion, ces résultats nous encouragent à poursuivre l'étude du gène *ZFP57* dans la prédisposition au DT1 et à rechercher les mécanismes fonctionnels sous-jacents. Dans cette perspective, le re-séquençage du gène chez des patients permettra de caractériser complètement le polymorphsime du gène et d'éclairer les analyses fonctionnelles qui pourraient être entreprises. Parmi celles-ci, il nous semble important de caractériser les modalités de la régulation de l'expression de *ZFP57*, en fonction du fond génétique, du type cellulaire et d'un éventuel contexte inflammatoire ou pharmacologique. L'anticorps monoclonal en cours de développement devrait être un outil précieux. Il nous aidera aussi à identifier les partenaires de ZFP57 au niveau de la chromatine des cellules du système immunitaire, mais aussi ses cibles grâce à des expériences de ChIP-Seq. Les partenaires comme les cibles pourraient alors constituer de nouveaux gènes candidats au DT1. On peut espérer que leur étude contribuera à réduire la part manquante de l'hérédité de cette maladie complexe.

Conclusion générale

A l'époque où j'ai rejoint le laboratoire, une cinquantaine de régions génomiques avaient été associées dans la prédisposition au DT1, en particulier grâce à l'étude de cohortes internationales. Cependant, ces associations n'expliquaient qu'imparfaitement la composante génétique du DT1. Le travail réalisé dans le cadre de ce mémoire avait pour objectif de détecter de nouveaux gènes. Il m'a permis d'aborder l'étude des facteurs génétiques de la maladie en combinant différentes méthodes, aussi bien pangénomiques que ciblées sur des gènes candidats.

Ma première approche était purement génétique. L'objectif était de répliquer les résultats d'associations génétiques au DT1 obtenus dans une cohorte française de « découverte » au sein d'une cohorte de « réplication » plus grande. Par génotypage à l'aide de sondes TaqMan, j'ai répliqué pour la première fois dans une cohorte française les associations majeures connues du DT1 dans le CMH et au niveau du gène de l'insuline. Mon travail a aussi identifié une nouvelle association suggestive pour le variant rs17638639 en 2q31.2, dont l'effet est comparable à celui de l'insuline. Ce travail montre l'intérêt de mener des études génétiques sur des populations de différentes origines.

La seconde approche reposait sur les résultats d'un transcriptome différentiel entre patients et leurs contrôles familiaux de la cohorte internationale du T1DGC sélectionnés en fonction de leur génotype pour le CMH. Mon travail portant sur la quantification des transcrits et des protéines des gènes d'intérêt a confirmé une diminution de l'expression chez les patients des transcrits de quatre gènes de la voie de l'interferon-γ (*IFNγ*, *STAT1*, *IRF1* et *IL27*) et de la protéine IFNγ elle-même, ainsi que des transcrits d'un gène candidat du CMH, *UBD*. De plus, j'ai pu relier UBD à la voie de l'IFNγ en révélant la co-expression de ces gènes. Ces résultats confortent l'intérêt du gène *UBD* dans la prédisposition au DT1 et valident la démarche d'un transcriptome différentiel sélectif pour identifier de nouveaux gènes.

Enfin, la dernière partie de mon travail était plus ciblée sur un gène candidat *ZFP57*, identifié initialement par un transcriptome différentiel. Ma démarche a été cette fois-ci à la fois génétique et fonctionnelle. Grâce à la combinaison du génotypage par sondes Taqman et à la quantification des transcrits de ce gène par qPCR, j'ai pu relier pour la première fois des variants de ce gène régulant son expression et les associer génétiquement à la prédisposition au DT1. Sur le plan fonctionnel, en mettant au point un anticorps spécifique polyclonal, j'ai démontré son expression protéique dans des cellules sanguines immunitaires humaines après la naissance, confirmant la pertinence de ce gène comme candidat au DT1.

Bibliographie

Aly, Theresa A., Erin E. Baschal, Mohamed M. Jahromi, Maria S. Fernando, Sunanda R. Babu, Tasha E. Fingerlin, Adam Kretowski, et al. 2008. « Analysis of Single Nucleotide Polymorphisms Identifies Major Type 1A Diabetes Locus Telomeric of the Major Histocompatibility Complex ». *Diabetes* 57 (3): 770‑76. doi:10.2337/db07-0900.

Aly, Theresa A., Akane Ide, Mohamed M. Jahromi, Jennifer M. Barker, Maria S. Fernando, Sunanda R. Babu, Liping Yu, et al. 2006. « Extreme Genetic Risk for Type 1A Diabetes ». *Proceedings of the National Academy of Sciences* 103 (38): 14074‑79. doi:10.1073/pnas.0606349103.

Barrett, Jeffrey C., David G. Clayton, Patrick Concannon, Beena Akolkar, Jason D. Cooper, Henry A. Erlich, Cécile Julier, et al. 2009. « Genome-Wide Association Study and Meta-Analysis Find That over 40 Loci Affect Risk of Type 1 Diabetes ». *Nature Genetics* 41 (6): 703‑7. doi:10.1038/ng.381.

Baschal, Erin E., Suparna A. Sarkar, Theresa A. Boyle, Janet C. Siebert, Jean M. Jasinski, Katharine R. Grabek, Taylor K. Armstrong, et al. 2011. « Replication and Further Characterization of a Type 1 Diabetes-Associated Locus at the Telomeric End of the Major Histocompatibility Complex ». *Journal of Diabetes* 3 (3): 238‑47. doi:10.1111/j.1753-0407.2011.00131.x.

Blankenhorn, Elizabeth P., Laura Cort, Dale L. Greiner, Dennis L. Guberski, et John P. Mordes. 2009. « Virus-Induced Autoimmune Diabetes in the LEW.1WR1 Rat Requires Iddm14 and a Genetic Locus Proximal to the Major Histocompatibility Complex ». *Diabetes* 58 (12): 2930‑38. doi:10.2337/db09-0387.

Bronson, P. G., P. P. Ramsay, G. Thomson, L. F. Barcellos, et Diabetes Genetics Consortium. 2009. « Analysis of Maternal-Offspring HLA Compatibility, Parent-of-Origin and Non-Inherited Maternal Effects for the Classical HLA Loci in Type 1 Diabetes ». *Diabetes, Obesity & Metabolism* 11 Suppl 1 (février): 74‑83. doi:10.1111/j.1463-1326.2008.01006.x.

Brozzi, Flora, Sarah Gerlo, Fabio Arturo Grieco, Matilda Juusola, Alexander Balhuizen, Sam Lievens, Conny Gysemans, et al. 2016. « Ubiquitin D Regulates IRE1α/C-Jun N-Terminal Kinase (JNK) Protein-Dependent Apoptosis in Pancreatic Beta Cells ». *The Journal of Biological Chemistry* 291 (23): 12040‑56. doi:10.1074/jbc.M115.704619.

Bustin, Stephen A., Vladimir Benes, Jeremy A. Garson, Jan Hellemans, Jim Huggett, Mikael Kubista, Reinhold Mueller, et al. 2009. « The MIQE Guidelines: Minimum Information for Publication of Quantitative Real-Time PCR Experiments ». *Clinical Chemistry* 55 (4): 611‑22. doi:10.1373/clinchem.2008.112797.

Cerami, Ethan G., Benjamin E. Gross, Emek Demir, Igor Rodchenkov, Ozgün Babur, Nadia Anwar, Nikolaus Schultz, Gary D. Bader, et Chris Sander. 2011. « Pathway Commons, a Web Resource for Biological Pathway Data ». *Nucleic Acids Research* 39 (Database issue): D685-690. doi:10.1093/nar/gkq1039.

Cirulli, Elizabeth T., et David B. Goldstein. 2010. « Uncovering the Roles of Rare Variants in Common Disease through Whole-Genome Sequencing ». *Nature Reviews. Genetics* 11 (6): 415‑25. doi:10.1038/nrg2779.

Cooper, Jason D., Deborah J. Smyth, Adam M. Smiles, Vincent Plagnol, Neil M. Walker, James E. Allen, Kate Downes, et al. 2008. « Meta-Analysis of Genome-Wide Association Study Data Identifies Additional Type 1 Diabetes Risk Loci ». *Nature Genetics* 40 (12): 1399‑1401. doi:10.1038/ng.249.

de Bakker, Paul I. W., et Soumya Raychaudhuri. 2012. « Interrogating the Major Histocompatibility Complex with High-Throughput Genomics ». *Human Molecular Genetics* 21 (R1): R29-36. doi:10.1093/hmg/dds384.

DeWan, Andrew T., Elizabeth W. Triche, Xuming Xu, Ling-I. Hsu, Connie Zhao, Kathleen Belanger, Karen Hellenbrand, et al. 2010. « PDE11A Associations with Asthma: Results of a Genome-Wide Association Scan ». *The Journal of Allergy and Clinical Immunology* 126 (4): 871‑873.e9. doi:10.1016/j.jaci.2010.06.051.

Dilthey, Alexander, Stephen Leslie, Loukas Moutsianas, Judong Shen, Charles Cox, Matthew R. Nelson, et Gil McVean. 2013. « Multi-Population Classical HLA Type Imputation ». *PLoS Computational Biology* 9 (2): e1002877. doi:10.1371/journal.pcbi.1002877.

Dixon, Anna L., Liming Liang, Miriam F. Moffatt, Wei Chen, Simon Heath, Kenny C. C. Wong, Jenny Taylor, et al. 2007. « A Genome-Wide Association Study of Global Gene Expression ». *Nature Genetics* 39 (10): 1202‑7. doi:10.1038/ng2109.

Dudbridge, Frank. 2008. « Likelihood-Based Association Analysis for Nuclear Families and Unrelated Subjects with Missing Genotype Data ». *Human Heredity* 66 (2): 87‑98. doi:10.1159/000119108.

Ehehalt, S., A. Neu, D. Michaelis, P. Heinke, A. M. Willasch, K. Dietz, et DIARY-Group Baden-Wuerttemberg. 2012. « Incidence of Type 1 Diabetes in Childhood before and after the Reunification of Germany--an Analysis of Epidemiological Data, 1960-2006 ». *Experimental and Clinical Endocrinology & Diabetes: Official Journal, German Society of Endocrinology [and] German Diabetes Association* 120 (8): 441‑44. doi:10.1055/s-0032-1309045.

Fert, Ingrid, Nicolas Cagnard, Simon Glatigny, Franck Letourneur, Sébastien Jacques, Judith A. Smith, Robert A. Colbert, et al. 2014. « Reverse Interferon Signature Is Characteristic of Antigen-Presenting Cells in Human and Rat Spondyloarthritis ». *Arthritis & Rheumatology (Hoboken, N.J.)* 66 (4): 841‑51. doi:10.1002/art.38318.

Groop, Leif, et Flemming Pociot. 2014. « Genetics of Diabetes--Are We Missing the Genes or the Disease? » *Molecular and Cellular Endocrinology* 382 (1): 726‑39. doi:10.1016/j.mce.2013.04.002.

Hall, Aisling O'Hara, Daniel P. Beiting, Cristina Tato, Beena John, Guillaume Oldenhove, Claudia Gonzalez Lombana, Gretchen Harms Pritchard, et al. 2012. « The Cytokines Interleukin 27 and Interferon-γ Promote Distinct Treg Cell Populations Required to Limit Infection-Induced Pathology ». *Immunity* 37 (3): 511‑23. doi:10.1016/j.immuni.2012.06.014.

Hilner, Joan E., Letitia H. Perdue, Elizabeth G. Sides, June J. Pierce, Ana M. Wägner, Alan Aldrich, Amanda Loth, et al. 2010. « Designing and Implementing Sample and Data Collection for an International Genetics Study: The Type 1 Diabetes Genetics Consortium (T1DGC) ». *Clinical Trials (London, England)* 7 (1 Suppl): S5‑32. doi:10.1177/1740774510373497.

Horton, Roger, Richard Gibson, Penny Coggill, Marcos Miretti, Richard J. Allcock, Jeff Almeida, Simon Forbes, et al. 2008. « Variation Analysis and Gene Annotation of Eight MHC Haplotypes: The MHC Haplotype Project ». *Immunogenetics* 60 (1): 1‑18. doi:10.1007/s00251-007-0262-2.

Horton, Roger, Laurens Wilming, Vikki Rand, Ruth C. Lovering, Elspeth A. Bruford, Varsha K. Khodiyar, Michael J. Lush, et al. 2004. « Gene Map of the Extended Human MHC ». *Nature Reviews. Genetics* 5 (12): 889‑99. doi:10.1038/nrg1489.

Hu, Xinli, Aaron J. Deutsch, Tobias L. Lenz, Suna Onengut-Gumuscu, Buhm Han, Wei-Min Chen, Joanna M. M. Howson, et al. 2015. « Additive and Interaction Effects at Three Amino Acid Positions in HLA-DQ and HLA-DR Molecules Drive Type 1 Diabetes Risk ». *Nature Genetics* 47 (8): 898‑905. doi:10.1038/ng.3353.

Hunt, Karen A., Vanisha Mistry, Nicholas A. Bockett, Tariq Ahmad, Maria Ban, Jonathan N. Barker, Jeffrey C. Barrett, et al. 2013. « Negligible Impact of Rare Autoimmune-Locus Coding-Region Variants on Missing Heritability ». *Nature* 498 (7453): 232‑35. doi:10.1038/nature12170.

Jia, Xiaoming, Buhm Han, Suna Onengut-Gumuscu, Wei-Min Chen, Patrick J. Concannon, Stephen S. Rich, Soumya Raychaudhuri, et Paul I. W. de Bakker. 2013. « Imputing Amino Acid Polymorphisms in Human Leukocyte Antigens ». *PloS One* 8 (6): e64683. doi:10.1371/journal.pone.0064683.

Kelchtermans, Hilde, Alfons Billiau, et Patrick Matthys. 2008. « How Interferon-Gamma Keeps Autoimmune Diseases in Check ». *Trends in Immunology* 29 (10): 479‑86. doi:10.1016/j.it.2008.07.002.

Kim, Seongho. 2015. « Ppcor: An R Package for a Fast Calculation to Semi-Partial Correlation Coefficients ». *Communications for Statistical Applications and Methods* 22 (6): 665‑74. doi:10.5351/CSAM.2015.22.6.665.

Lee, Caroline G. L., Jianwei Ren, Ian S. Y. Cheong, Kenneth H. K. Ban, London L. P. J. Ooi, Soo Yong Tan, Alison Kan, et al. 2003. « Expression of the FAT10 Gene Is Highly Upregulated in Hepatocellular Carcinoma and Other Gastrointestinal and Gynecological Cancers ». *Oncogene* 22 (17): 2592‑2603. doi:10.1038/sj.onc.1206337.

Lenz, Tobias L., Aaron J. Deutsch, Buhm Han, Xinli Hu, Yukinori Okada, Stephen Eyre, Michael Knapp, et al. 2015. « Widespread Non-Additive and Interaction Effects within HLA Loci Modulate the Risk of Autoimmune Diseases ». *Nature Genetics* 47 (9): 1085‑90. doi:10.1038/ng.3379.

Li, Xiajun, Mitsuteru Ito, Fen Zhou, Neil Youngson, Xiaopan Zuo, Philip Leder, et Anne C. Ferguson-Smith. 2008. « A Maternal-Zygotic Effect Gene, Zfp57, Maintains Both Maternal and Paternal Imprints ». *Developmental Cell* 15 (4): 547‑57. doi:10.1016/j.devcel.2008.08.014.

Mackay, Deborah J. G., Jonathan L. A. Callaway, Sophie M. Marks, Helen E. White, Carlo L. Acerini, Susanne E. Boonen, Pinar Dayanikli, et al. 2008. « Hypomethylation of Multiple Imprinted Loci in Individuals with Transient Neonatal Diabetes Is Associated with Mutations in ZFP57 ». *Nature Genetics* 40 (8): 949‑51. doi:10.1038/ng.187.

McCarthy, Mark I., Gonçalo R. Abecasis, Lon R. Cardon, David B. Goldstein, Julian Little, John P. A. Ioannidis, et Joel N. Hirschhorn. 2008. « Genome-Wide Association Studies for Complex Traits: Consensus, Uncertainty and Challenges ». *Nature Reviews. Genetics* 9 (5): 356‑69. doi:10.1038/nrg2344.

Meka, Rakeshchandra R., Shivaprasad H. Venkatesha, Steven Dudics, Bodhraj Acharya, et Kamal D. Moudgil. 2015. « IL-27-Induced Modulation of Autoimmunity and Its Therapeutic Potential ». *Autoimmunity Reviews* 14 (12): 1131‑41. doi:10.1016/j.autrev.2015.08.001.

Miao, Feng, Irene Gaw Gonzalo, Linda Lanting, et Rama Natarajan. 2004. « In Vivo Chromatin Remodeling Events Leading to Inflammatory Gene Transcription under Diabetic Conditions ». *The Journal of Biological Chemistry* 279 (17): 18091‑97. doi:10.1074/jbc.M311786200.

Mühl, Heiko, et Josef Pfeilschifter. 2003. « Anti-Inflammatory Properties of pro-Inflammatory Interferon-Gamma ». *International Immunopharmacology* 3 (9): 1247‑55. doi:10.1016/S1567-5769(03)00131-0.

Nejentsev, Sergey, Joanna M. M. Howson, Neil M. Walker, Jeffrey Szeszko, Sarah F. Field, Helen E. Stevens, Pamela Reynolds, et al. 2007. « Localization of Type 1 Diabetes Susceptibility to the MHC Class I Genes HLA-B and HLA-A ». *Nature* 450 (7171): 887‑92. doi:10.1038/nature06406.

Nejentsev, Sergey, Neil Walker, David Riches, Michael Egholm, et John A. Todd. 2009. « Rare Variants of IFIH1, a Gene Implicated in Antiviral Responses, Protect against Type 1 Diabetes ». *Science (New York, N.Y.)* 324 (5925): 387‑89. doi:10.1126/science.1167728.

Nguyen, Cao, Michael D. Varney, Leonard C. Harrison, et Grant Morahan. 2013. « Definition of High-Risk Type 1 Diabetes HLA-DR and HLA-DQ Types Using Only Three Single Nucleotide Polymorphisms ». *Diabetes* 62 (6): 2135‑40. doi:10.2337/db12-1398.

Okazaki, S., S. Tanase, B. K. Choudhury, K. Setoyama, R. Miura, M. Ogawa, et C. Setoyama. 1994. « A Novel Nuclear Protein with Zinc Fingers down-Regulated during Early Mammalian Cell Differentiation ». *The Journal of Biological Chemistry* 269 (9): 6900‑6907.

Oliva, Joan, Fawzia Bardag-Gorce, Andrew Lin, Barbara A. French, et Samuel W. French. 2010. « The Role of Cytokines in UbD Promoter Regulation and Mallory-Denk Body-like Aggresomes ». *Experimental and Molecular Pathology* 89 (1): 1‑8. doi:10.1016/j.yexmp.2010.04.001.

Onengut-Gumuscu, Suna, Wei-Min Chen, Oliver Burren, Nick J. Cooper, Aaron R. Quinlan, Josyf C. Mychaleckyj, Emily Farber, et al. 2015. « Fine Mapping of Type 1 Diabetes Susceptibility Loci and Evidence for Colocalization of Causal Variants with Lymphoid Gene Enhancers ». *Nature Genetics* 47 (4): 381‑86. doi:10.1038/ng.3245.

Parkes, Miles, Adrian Cortes, David A. van Heel, et Matthew A. Brown. 2013. « Genetic Insights into Common Pathways and Complex Relationships among Immune-Mediated Diseases ». *Nature Reviews. Genetics* 14 (9): 661‑73. doi:10.1038/nrg3502.

Patel, Tejas, Vasu Patel, Rajvir Singh, et Sundararajan Jayaraman. 2011. « Chromatin Remodeling Resets the Immune System to Protect against Autoimmune Diabetes in Mice ». *Immunology and Cell Biology* 89 (5): 640‑49. doi:10.1038/icb.2010.144.

Plant, Katharine, Benjamin P. Fairfax, Seiko Makino, Claire Vandiedonck, Jayachandran Radhakrishnan, et Julian C. Knight. 2014. « Fine Mapping Genetic Determinants of the Highly Variably Expressed MHC Gene ZFP57 ». *European Journal of Human Genetics: EJHG* 22 (4): 568‑71. doi:10.1038/ejhg.2013.244.

Pruim, Randall J., Ryan P. Welch, Serena Sanna, Tanya M. Teslovich, Peter S. Chines, Terry P. Gliedt, Michael Boehnke, Gonçalo R. Abecasis, et Cristen J. Willer. 2010. « LocusZoom: Regional Visualization of Genome-Wide Association Scan Results ». *Bioinformatics (Oxford, England)* 26 (18): 2336‑37. doi:10.1093/bioinformatics/btq419.

Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, et al. 2007. « PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses ». *American Journal of Human Genetics* 81 (3): 559‑75. doi:10.1086/519795.

Quenneville, Simon, Gaetano Verde, Andrea Corsinotti, Adamandia Kapopoulou, Johan Jakobsson, Sandra Offner, Ilaria Baglivo, et al. 2011. « In Embryonic Stem Cells, ZFP57/KAP1 Recognize a Methylated Hexanucleotide to Affect Chromatin and DNA Methylation of Imprinting Control Regions ». *Molecular Cell* 44 (3): 361‑72. doi:10.1016/j.molcel.2011.08.032.

Rakyan, Vardhman K., Huriya Beyan, Thomas A. Down, Mohammed I. Hawa, Siarhei Maslau, Deeqo Aden, Antoine Daunay, et al. 2011. « Identification of Type 1 Diabetes-Associated DNA Methylation Variable Positions That Precede Disease Diagnosis ». *PLoS Genetics* 7 (9): e1002300. doi:10.1371/journal.pgen.1002300.

Ram, Ramesh, Munish Mehta, Quang T. Nguyen, Irma Larma, Bernhard O. Boehm, Flemming Pociot, Patrick Concannon, et Grant Morahan. 2016. « Systematic Evaluation of Genes and Genetic Variants Associated with Type 1 Diabetes Susceptibility ». *Journal of Immunology (Baltimore, Md.: 1950)* 196 (7): 3043‑53. doi:10.4049/jimmunol.1502056.

Riggs AD, Martienssen RA, Russo VEA 1996. Introduction. In Epigenetic mechanisms of gene regulation (ed. Russo VEA, et al.), pp. 1–4 Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY

Rjasanowski, I., P. Heinke, D. Michaelis, et T. L. Kurajewa. 1990. « The Higher Frequency of Type I (Insulin-Dependent) Diabetes in Fathers than in Mothers of Type I-Diabetic Children ». *Experimental and Clinical Endocrinology* 95 (1): 91‑96. doi:10.1055/s-0029-1210939.

Smith, Judith A., Michael D. Barnes, Dihua Hong, Monica L. DeLay, Robert D. Inman, et Robert A. Colbert. 2008. « Gene Expression Analysis of Macrophages Derived from Ankylosing Spondylitis Patients Reveals Interferon-Gamma Dysregulation ». *Arthritis and Rheumatism* 58 (6): 1640‑49. doi:10.1002/art.23512.

Stefan, Mihaela, Weijia Zhang, Erlinda Concepcion, Zhengzi Yi, et Yaron Tomer. 2014. « DNA Methylation Profiles in Type 1 Diabetes Twins Point to Strong Epigenetic Effects on Etiology ». *Journal of Autoimmunity* 50 (mai): 33‑37. doi:10.1016/j.jaut.2013.10.001.

Stewart, C. Andrew, Roger Horton, Richard J. N. Allcock, Jennifer L. Ashurst, Alexey M. Atrazhev, Penny Coggill, Ian Dunham, et al. 2004. « Complete MHC Haplotype Sequencing for Common Disease Gene Mapping ». *Genome Research* 14 (6): 1176‑87. doi:10.1101/gr.2188104.

Tada, Y., Y. Yamaguchi, T. Kinjo, X. Song, T. Akagi, H. Takamura, T. Ohta, T. Yokota, et H. Koide. 2015. « The Stem Cell Transcription Factor ZFP57 Induces IGF2 Expression to Promote Anchorage-Independent Growth in Cancer Cells ». *Oncogene* 34 (6): 752‑60. doi:10.1038/onc.2013.599.

Tang, Walfred W. C., Sabine Dietmann, Naoko Irie, Harry G. Leitch, Vasileios I. Floros, Charles R. Bradshaw, Jamie A. Hackett, Patrick F. Chinnery, et M. Azim Surani. 2015. « A Unique Gene Regulatory Network Resets the Human Germline Epigenome for Development ». *Cell* 161 (6): 1453‑67. doi:10.1016/j.cell.2015.04.053.

Trynka, Gosia, Karen A. Hunt, Nicholas A. Bockett, Jihane Romanos, Vanisha Mistry, Agata Szperl, Sjoerd F. Bakker, et al. 2011. « Dense Genotyping Identifies and Localizes Multiple Common and Rare Variant Association Signals in Celiac Disease ». *Nature Genetics* 43 (12): 1193‑1201. doi:10.1038/ng.998.

UK10K Consortium, Klaudia Walter, Josine L. Min, Jie Huang, Lucy Crooks, Yasin Memari, Shane McCarthy, et al. 2015. « The UK10K Project Identifies Rare Variants in Health and Disease ». *Nature* 526 (7571): 82‑90. doi:10.1038/nature14962.

Vandiedonck, Claire, Martin S. Taylor, Helen E. Lockstone, Katharine Plant, Jennifer M. Taylor, Caroline Durrant, John Broxholme, Benjamin P. Fairfax, et Julian C. Knight. 2011 « Pervasive Haplotypic Variation in the Spliceo-Transcriptome of the Human Major Histocompatibility Complex ». *Genome Research* 21 (7): 1042‑54. doi:10.1101/gr.116681.110.

Wallace, Chris, Deborah J. Smyth, Meeta Maisuria-Armer, Neil M. Walker, John A. Todd, et David G. Clayton. 2010. « The Imprinted DLK1-MEG3 Gene Region on Chromosome 14q32.2 Alters Susceptibility to Type 1 Diabetes ». *Nature Genetics* 42 (1): 68‑71. doi:10.1038/ng.493.

Wandstrat, A., et E. Wakeland. 2001. « The Genetics of Complex Autoimmune Diseases: Non-MHC Susceptibility Genes ». *Nature Immunology* 2 (9): 802‑9. doi:10.1038/ni0901-802.

Warram, J. H., A. S. Krolewski, M. S. Gottlieb, et C. R. Kahn. 1984. « Differences in Risk of Insulin-Dependent Diabetes in Offspring of Diabetic Mothers and Diabetic Fathers ». *The New England Journal of Medicine* 311 (3): 149‑52. doi:10.1056/NEJM198407193110304.

Wellcome Trust Case Control Consortium, Nick Craddock, Matthew E. Hurles, Niall Cardin, Richard D. Pearson, Vincent Plagnol, Samuel Robson, et al. 2010. « Genome-Wide Association Study of CNVs in 16,000 Cases of Eight Common Diseases and 3,000 Shared Controls ». *Nature* 464 (7289): 713‑20. doi:10.1038/nature08979.

Westra, Harm-Jan, et Lude Franke. 2014. « From Genome to Function by Studying eQTLs ». *Biochimica Et Biophysica Acta* 1842 (10): 1896‑1902. doi:10.1016/j.bbadis.2014.04.024.

Westra, Harm-Jan, Marjolein J. Peters, Tõnu Esko, Hanieh Yaghootkar, Claudia Schurmann, Johannes Kettunen, Mark W. Christiansen, et al. 2013. « Systematic Identification of Trans eQTLs as Putative Drivers of Known Disease Associations ». *Nature Genetics* 45 (10): 1238‑43. doi:10.1038/ng.2756.

Wigginton, Janis E., et Gonçalo R. Abecasis. 2005. « PEDSTATS: Descriptive Statistics, Graphics and Quality Assessment for Gene Mapping Data ». *Bioinformatics (Oxford, England)* 21 (16): 3445‑47. doi:10.1093/bioinformatics/bti529.

Wu, Linfeng, Sophie I. Candille, Yoonha Choi, Dan Xie, Lihua Jiang, Jennifer Li-Pook-Than, Hua Tang, et Michael Snyder. 2013. « Variation and Genetic Control of Protein Abundance in Humans ». *Nature* 499 (7456): 79‑82. doi:10.1038/nature12223.

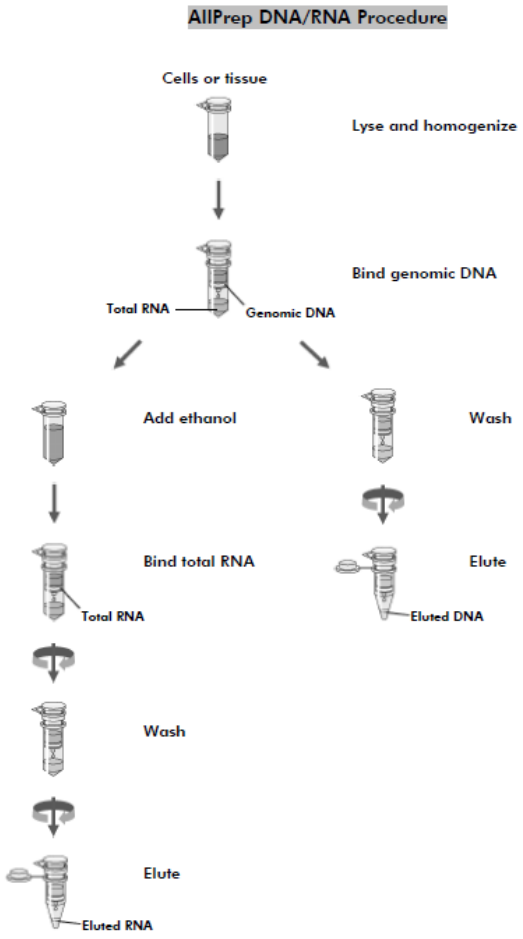
Yan, D.-W., D.-W. Li, Y.-X. Yang, J. Xia, X.-L. Wang, C.-Z. Zhou, J.-W. Fan, et al. 2010. « Ubiquitin D Is Correlated with Colon Cancer Progression and Predicts Recurrence for Stage II-III Disease after Curative Surgery ». *British Journal of Cancer* 103 (7): 961‑69. doi:10.1038/sj.bjc.6605870.

Yang, Yisheng, et Lawrence Chan. 2016. « Monogenic Diabetes: What It Teaches Us on the Common Forms of Type 1 and Type 2 Diabetes ». *Endocrine Reviews* 37 (3): 190‑222. doi:10.1210/er.2015-1116.

Zamani, Fatemeh, Shohreh Almasi, Tohid Kazemi, Rana Jahanban Esfahlan, et Mohammad Reza Aliparasti. 2015. « New Approaches to the Immunotherapy of Type 1 Diabetes Mellitus Using Interleukin-27 ». *Advanced Pharmaceutical Bulletin* 5 (Suppl 1): 599‑603. doi:10.15171/apb.2015.081.

Annexe

Annexe 1. Schéma des étapes du protocole d'extraction d'ADN et d'ARN sur colonne Allprep DNA/RNA (Qiagen).



Annexe 2. Séquences des primers utilisés en qPCR

Gène cible	Séquence	Efficacité (%)	Taille amplicon (pb)
UBD	Forward : CTTCCTGCCTCTGTGTGCAT Reverse : CCCAGCAAAAGAACCTGGTC	91,4	145
IFNγ	Forward : AGATCCCATGGGTTGTGTGT Reverse : AAAGCACTGGCTCAGATTGC	91,9	118
IL27	Forward : GCCAGGAGTGAACCTGTACCT Reverse : GAAGCGTGGTGGAGATGAAG	84,7	122
IRF1	Forward : CCATTCACACAGGCCGATAC Reverse : TCCTGCTCTGGTCTTTCACC	91,8	123
STAT1	Forward : CGATGGGCTCAGCTTTCA Reverse : GCTGGCGTTAGGACCAAGA	95,4	117
ZFP57	Forward : GCCTTTCAGAAGGCAAGAAG Reverse : CCCCTCATCTCTCAGACTGG	60	62
GAPDH	Forward : CCTCAACGACCACTTTGTCA Reverse : GAGGGTCTCTCTCTTCCTCTTGT	78,9	156
B-actine	Forward : GGACTTCGAGCAAGAGATGG Reverse : AGGAAGGAAGGCTGGAAGAG	89,2	138
RPL30	Forward : CCTAAGGCAGGAAGATGGTG Reverse : AATGACCAATTTCGCTTTGC	93,4	155

Annexe 3. Recommandations essentielles de MIQE suivies pour chaque qPCR dans le laboratoire

Design expérimental
Définition des groupes contrôles et expérimentaux
Nombre dans chaque groupe
Echantillon
Description
Volume de l'échantillon
Comment l'échantillon a été congelé
Condition de stockage et durée
Extraction des acides nucléiques
Procédure
Nom du kit utilisé
Traitement DNase ou RNase
Evaluation de la contamination (ADN ou ARN)
Quantification des acides nucléiques
Instrument et méthode
Pureté (A260/A280)
Rendement
Intégrité de l'ARN : méthode/instrument
RIN
Electrophorèse
Test inhibition par courbe de dilutions
Reverse transcription
Conditions complètes des réactions
Quantité d'ARN et volume réactionnel
Cq sans et avec transcriptase reverse
Conditions de stockage de l'ADNc
Information sur la cible de qPCR
Localisation de l'amplicon
Longueur de l'amplicon
PCR in silico
Localisation de chaque primer (exon/intron)
Variants ciblés
Oligonucléotides
Séquence des primers
Protocole de qPCR
Conditions complètes de réaction
Volume réactionnel et quantité d'ADNc
Polymérase utilisée et concentration
Instrument de qPCR utilisé
Validation de la qPCR
Efficacité de la qPCR
Spécificité (courbe et température de fusion)
Gamme dynamique linéaire (évaluée par dilution en série)
Etude des Cq des puits sans matrice
Analyse des data
Logiciel d'analyse des données brutes de qPCR
Justification du nombre et du choix des gènes de référence
Description de la méthode de normalisation
Nombre et concordance des réplicats
Variation intra-plaque
Reproductibilité (coefficient de variation)
Méthode statistique d'analyse d'expression différentielle
Logiciel d'analyse statistique (R)

Annexe 4. Liste des anticorps utilisés en western blot

Cible	Référence	Séquence cible	Espèce hôte	Espèce immunogène	Poids moléculaire attendu (kDa)	Localisation cellulaire
HSP90	ab13495	Non renseignée	Lapin	Humain	90	Cytoplasme
SP1	ab13370	Entre les acides aminés 750 et 785 (C-terminal)	Lapin	Humain	90	Nucléaire
HDAC2	ab16032	Autour des résidus 450 (C-terminal)	Lapin	Humain	55,3	Nucléaire
KAP1	ab10483	Proche du 1er exon	Lapin	Humain	100	Nucléaire
PCNA	ab18197	Autour des résidus 200 (C-terminal)	Lapin	Humain	29	Nucléaire
Histone H3	ab62642	Non renseignée	Lapin	Humain	15	Chromatine
Vimentine	ab71144	Non renseignée	Lapin	Humain	54	Cytosquelette
Ctyokératine 18	ab52948	C-terminal	Lapin	Humain	48	Cytosquelette
Tubuline	ab15246	Entre les acides aminés 426 et 450 (C-terminal)	Lapin	Humain	50	Cytosquelette
ZFP57-A1	ab50944	N-ter : MFEQLKPIEPRDCWREARVKKKPVTFEDVAVNFTQEEWDCLDASQRVLYQ	Lapin	Humain	61,4 ; 59,2 et 51,2	Attendue : Chromatine et nucléaire
ZFP57-A2	ab173866	ZF4-5 : PVTRTQAPITGTLCQDARSNSHPVKPSR	Lapin	Humain	61,4 ; 59,2 et 51,3	Attendue : Chromatine et nucléaire
ZFP57-B	Proteogenix	ZF4-5 : cys-QNQEPVDGNQ & cys-EPIFRTEGPMAQN	Lapin	Humain	61,4 ; 59,2 et 51,3	Attendue : Chromatine et nucléaire
ZFP57-C	Proteogenix	ZF4-5 : cys-QNQEPVDGNQ & cys-EPIFRTEGPMAQN	Lapin	Humain	61,4 ; 59,2 et 51,4	Attendue : Chromatine et nucléaire
ZFP57-D	Proteogenix	C-ter : cys-HGGDQSPPRIHTPR &CKGDKTKEAVSILKHK	Lapin	Humain	61,4 ; 59,2 et 51,5	Attendue : Chromatine et nucléaire
ZFP57-E	Proteogenix	C-ter : cys-HGGDQSPPRIHTPR &CKGDKTKEAVSILKHK	Lapin	Humain	61,4 ; 59,2 et 51,6	Attendue : Chromatine et nucléaire
β-actine	ab8227	Entre les acides aminés 1 et 100	Lapin	Humain	42	Cytoplasme Cytosquelette

Annexe 5. Liste des anticorps utilisés en cytométrie en flux

Cible	Référence	Dilution
ZFP57-A1	Abcam-ab50944	1/1000
ZFP57-B	Proteogenix	1/1000
ZFP57-C	Proteogenix	1/1000
ZFP57-D	Proteogenix	1/25
ZFP57-E	Proteogenix	1/1000
BD PerCP CD8	BD-345774	10 µL par tube
BD APC CD4	BD-555349	10 µL par tube
Biolegend brilliant violet 785 CD19	BD-302239	5 µL par tube
eBioscience APC-eFluor CD31	BD- 47031942	5 µL par tube
Biolegend brilliant violet 605 D14	BD-301833	5 µL par tube
Vybrant Dye Cycle Violet stain	V35003	5 µL par échantillon
anti-lapin couplé PE	Abcam-ab7007	1/250

ÉCOLE PRATIQUE DES HAUTES ÉTUDES
Sciences de la Vie et de la Terre

Approches génétiques et génomiques pour l'identification de gènes prédisposant à une maladie multifactorielle : le diabète de type 1

Morgane Bourmaud

09 décembre 2016

RESUME

Le **diabète de type 1** (DT1) est une maladie multifactorielle dans laquelle plus de 50 régions génétiques ont été impliquées. A l'instar des autres maladies auto-immunes, le **Complexe Majeur d'Histocompatibilité** (CMH) domine largement, impliquant les gènes de classe II *HLA-DRB1* et *-DQB1*. Toutefois, la composante génétique n'est pas encore totalement expliquée**.** D'autres gènes du CMH et hors du CMH doivent intervenir. Pour les identifier, trois approches ont été menées : une étude d'association pangénomique réalisée pour la première fois dans une cohorte française, un transcriptome différentiel entre des patients et leurs contrôles apparentés et l'étude génétique fonctionnelle de *ZFP57*, un gène candidat du CMH.

Mon premier objectif était de répliquer les **associations génétiques** obtenues dans une cohorte française de « découverte »**.** Les résultats obtenus avec la cohorte de « réplication » confirment l'association majeure du CMH avec les SNPs tagguant les allèles de risque *HLA-DR3* et *DR4* ($p=4,2x10^{-31}$ et $9,2x10^{-27}$). Ils répliquent aussi pour la première fois l'association au DT1 dans le CMH de classe II de rs9273363 qui est le variant le mieux associé (OR=7,18, $p=8,5x10^{-59}$). Le gène de l'insuline est le second locus associé significativement au DT1 ($p=2,3x10^{-8}$, OR=2,1). Surtout, nous rapportons pour la première fois une nouvelle association suggestive pour rs17638639 en 2q31.2 ($p=3,32x10^{-6}$, OR=2,16) avec un effet comparable à celui de l'*INS*. Elle devance celle connue pour le gène *PTPN22* qui n'atteint pas le seuil de suggestivité dans cette cohorte.

Le second objectif était de valider par qPCR des gènes différentiellement exprimés entre patients et leurs contrôles familiaux identifiés par un transcriptome dans la cohorte internationale T1DGC. Il s'agit de 4 gènes de la voie de l'**interferon-γ** (*IFNγ*, *STAT1*, *IRF1* et *IL27*) et d'***UBD***, un gène candidat du CMH. Dans la cohorte européenne, on retrouve une diminution significative de l'expression de ces 5 transcrits chez les patients, ainsi que de la protéine IFNγ dont l'expression est corrélée à celle des transcrits ($1,7x10^{-34}$). De plus, ces gènes sont quantitativement corrélés entre eux, avec notamment une corrélation partielle significative entre *UBD* et *IFNγ* ($r^2=0,183$, $p=3,52x10^{-15}$), ce qui permet de rattacher *UBD* à la voie de l'IFNγ.

Le troisième objectif était de tester des polymorphismes de ***ZFP57***, un gène du CMH codant une protéine à doigts de Zinc impliquée dans les mécanismes épigénétiques embryonnaires et muté dans des diabètes néonataux transitoires. Pour la première fois, nous montrons que les mêmes variants de *ZFP57* peuvent en contrôler l'expression et être associés au DT1. Cette association semble impliquer une interaction en *cis* avec l'allèle DR3. Nous avons également fait produire et mis au point un anticorps polyclonal spécifique chez l'homme qui a permis de localiser la protéine dans le noyau au niveau de la chromatine et de démontrer son expression dans les leucocytes circulants de l'adulte pour la première fois. Ce travail renforce l'intérêt de *ZFP57* comme nouveau gène candidat dans la prédisposition au DT1 au sein du CMH.

Mots clés : diabète de type 1, CMH, association génétique, transcriptome, interferon-γ, *UBD*, *ZFP57*