Université de Montréal

Mémoire de Maîtrise

Investigation génétique de NAFLD dans le diabète de type 2 via construction d'un modèle de prédiction de la maladie et par criblage du locus PNPLA3-SAMM50

Par : Redha Attaoua

Département de Biochimie et médecine moléculaire

Université de Montréal

Mémoire présenté en vue de l'obtention du grade de Maîtrise en Bio-informatique

Juillet 2019

Remerciements

Je dédie tout d'abord ce mémoire de maîtrise à mes parents, à mes frères et à ma petite sœur.

Mes remerciements les plus chaleureux vont en premier à mon directeur de maîtrise, Pr. Pavel Hamet. Je le remercie pour l'accueil au sein de son équipe de recherche et pour sa grande disponibilité et ses conseils précieux et intarissables.

Mes remerciements vont également à mon codirecteur de maîtrise, Dr. Simon Gravel. Je le remercie pour ses conseils notamment en biostatistiques et en bio-informatique, et aussi pour son soutien.

Je remercie Pr. Johanne Tremblay pour ses conseils et son soutien durant toute la période de stage. Pr. Tremblay et Pr. Hamet ont été un motivateur de taille pour moi.

Mes grands remerciements vont aussi à Dr. Mounsif Haloui, chercheur senior au sein du groupe. Mounsif reste pour moi un exemple d'abnégation dans le travail et sa culture scientifique a été d'un grand apport pour moi.

Mes remerciements vont également à François-Christophe Marois-Blanchet, à François Harvey et à Dr. Ramzan Tahir pour leur aide notamment sur le plan technique.

Je remercie aussi Madame Carole Long et Camil Hishmih.

A ceux-là et à d'autres,

Je dédie mon mémoire de maîtrise

…

Résumé

La stéatose hépatique non-alcoolique (NAFLD) est une altération hépatique fréquente dans le diabète de type 2 (DT2) et est associée à diverses complications telles que la mortalité. L'établissement d'outils de prédiction non-invasifs de NAFLD est primordial. Mon projet de maîtrise avait pour objectif d'établir des marqueurs génétiques de NAFLD dans le DT2 via deux stratégies : 1) une sélection non-ciblée des marqueurs génétiques (SNPs) via la méthode LASSO et 2) une sélection ciblée de SNPs rapportés comme liés à la maladie ou à des altérations associées. Une population de 4098 patients avec DT2 d'origine caucasienne (ADVANCE) a été utilisée. Des données statistiques sommaires d'études pangénomiques ont été exploitées pour sélectionner, via LASSO, les marqueurs génétiques (SNPs) à inclure dans le score de risque polygénique (PRS). J'ai également développé un modèle de 3210 SNPs ajusté par des covariables capable de prédire les taux élevés de ALT (AUC=0,69) et la mortalité non-cardiovasculaire (AUC=0,66). Le criblage du locus candidat PNPLA3-SAMM50 a mis en avant une diversité des associations génétiques aux différentes altérations métaboliques comme les taux de ALT (substitut du diagnostic de NAFLD) (rs2294915, $P = 1{,}83 \times 10^{-7}$), à la mortalité non-cardiovasculaire (rs2294917, $P = 3{,}9 \times 10^{-4}$) et à l'efficacité de la thérapie intensive antidiabétique chez certains patients de la population (porteurs GG de rs16991236, P=0,007). Mes travaux ont permis de mieux comprendre le fond génétique de NAFLD dans le DT2 et laissent envisager l'établissement d'outils de diagnostic et de suivi de la maladie plus adéquats.

Mots-clés. NAFLD, diabète de type 2, mortalité non-cardiovasculaire, SNP, score de risque polygénique, PNPLA3-SAMM50

Abstract

Non-alcoholic fatty liver disease (NAFLD) is a liver disorder more frequent in type 2 diabetes (T2D) and is associated with complications such as mortality. For this reason, establishing non-invasive tools for predicting NAFLD is crucial. My master's project aimed to establish genetic markers for NAFLD in T2D using two strategies: 1) a non-targeted selection of genetic markers (SNPs) by the LASSO method and 2) a targeted selection of SNPs reported as associated with the disease or its related abnormalities. A population involving 4098 patients with T2D and Caucasian ancestry was used. Summary statistics data of pangenomic studies were exploited for the selection of SNPs to be involved in the polygenic risk score (PRS). I also designed a model of 3210 SNPs adjusted by covariates and able to predict the high rates of ALT (AUC=0.69) and non-cardiovascular death (AUC=0.66). Mapping of the candidate locus PNPLA3-SAMM50 allowed the observation of diversity in terms of genetic association with the metabolic abnormalities such as ALT (surrogate of NAFLD) (rs2294915, $P = 1.83x10^{-7}$), non-cardiovascular death (rs2294917, $P = 3.9x10^{-4}$) and the efficiency of the intensive antidiabetic therapy within a subgroup in the population (individuals with GG of rs16991236, P = 0.007). My studies allowed for a better understanding of the genetic background of NAFLD in T2D and open perspectives for establishing more adequate tools for diagnosis and follow-up of the disease.

Keywords. NAFLD, type 2 diabetes, non-cardiovascular death, SNP, polygenic risk score, PNPLA3-SAMM50

Table des matières

Introduction **1**

Chapitre 1 : Maladies complexes et outils de prédiction **4**

I- Exemple de marqueurs exploités dans l'exploration génétique des maladies complexes : les SNP (*single nucleotide polymorphisms*). **4**

II- Utilisation des SNPs dans l'exploration des maladies complexes. **4**

III- La stratégie par criblage dense du génome humain. **5**

III-1- Notion de DL dans les études par GWAS. 6
III-2- Outils statistiques pour l'exploitation des données des études par GWAS. 7
III-3- Limites des études par GWAS. 10

IV- Modélisation statistique pour la prédiction et la stratification des maladies complexes. **11**

IV-1- Méthode statistique LASSO. *12*

V- Criblage des loci de susceptibilité aux maladies complexes. **14**

Chapitre 2 : Diabète de type 2 et complications **16**

I- Complications microvasculaires du DT2. ..... **16**

I-1- La rétinopathie diabétique. *16*

I-2- La néphropathie diabétique. *17*

II- Complications macrovasculaires du DT2. ..... **17**

II-1- L'accident vasculaire cérébral. 17
II-2- L'infarctus du myocarde. 17

III- Complications hépatiques dans DT2. ..... **18**

Chapitre 3 : Stéatose hépatique non-alcoolique **19**

I- Données statistiques sur la prévalence du NAFLD. **19**

II- Physiopathologie du NAFLD. **20**

III- NAFLD et DT2. ..... **20**

IV- Autres facteurs influençant NAFLD. **21**

V- Mortalité liée à NAFLD. **21**

VI- Thérapies antidiabétiques dans le traitement de NAFLD. **22**

VII- Outils de diagnostic non-invasif de NAFLD. **23**

A- L'imagerie médicale dans le diagnostic et le suivi de NAFLD. 23

B- Marqueurs biologique et scores clinico-biologiques pour le diagnostic et le suivi de NAFLD. 24

B-1- L'alanine aminotransférase (ALT): *..... 24*

B-2- L'aspartate aminotransférase (AST): *..... 24*

C- Marqueurs génétiques liés à NAFLD. 25

C-1- PNPLA3 (patatin like phospholipase domain containing 3). *..... 25*

C-2- Criblage du locus PNPLA3-SAMM50. *..... 27*

C-3- Modèles génétiques de prédiction de NAFLD rapportés dans la littérature. *..... 28*

Objectifs et hypothèse de travail 30

Matériel et méthodes. 32

I- Base de données exploitée lors de l'analyse préliminaire : 32

II- Population étudiée. 32

III- Outils statistiques et bioinformatiques exploités. 35

IV- Sélection des SNPs pour la construction des PRS. 37

Résultats 40

I- Résultats obtenus lors de l'analyse préliminaire. 40

II- Résultats obtenus lors de l'investigation au sein de la population ADVANCE. 42

A- Exploration génétique via une approche non-ciblée 44

Conception par la méthode statistique LASSO d'un modèle de prédiction de NAFLD. *..... 44*

B- Exploration génétique via une approche ciblée 47

B-1- Conception, via sélection ciblée des SNPs, d'un modèle de prédiction de NAFLD et de la mortalité dans le DT2. *..... 48*

B-2- Criblage dense du locus PNPLA3-SAMM50 *..... 57*

Discussion et conclusions 69

Bibliographie 76

Liste des abréviations

ACCORD	*Action to Control Cardiovascular Risk in Diabetes*
ADIPOQ	*Adiponectin gene*
ADVANCE	*Action in Diabetes and Vascular Disease: Preterax and Diamicron MR Controlled Evaluation (ADVANCE) trial*
ALT	alanine aminotransférase
ANOVA	analyse de la variance
AST	Aspartate amino-transférase
AUC	area under the receiver operating characteristic
AVC	accident vasculaire cérébral
CARTaGENE	*Quebec's population-based biobank for public health and personalized genomics*
CEU ancestry	Utah residents (CEPH) with Northern and Western European
DisGeNET	database of gene-disease associations
DYSF	*dysferlin gene*
FTO	*FTO alpha-ketoglutarate dependent dioxygenase gene*
GATAD2A	*GATA zinc finger domain containing 2A*
GCKR:	*glucokinase regulator gene*
GIANT	*The Genetic Investigation of ANthropometric Traits consortium*
GLM	modèle linéaire généralisé
GLP1	*glucagon-like peptide-1*
GWAS	*genome-wide association study*
HDL	*High-density lipoprotein*
IMC	indice de masse corporelle
KLF6	*Kruppel like factor 6 gene*
LASSO	*least absolute shrinkage and selection operator*
LASSOSUM	LASSO for GWAS with summary statistics

MAF	*minor-allele frequency*
MBOAT	Membrane Bound O-Acyltransferase Domain Containing 1 gene
NAFLD	Non-alcoholic fatty liver disease
NASH	stéatohépatite non-alcoolique
OPTITHERA	optimisation des approches thérapeutiques en soins de première ligne
PARVB	*parvin beta gene*
PLINK	outils pour analyses par *genome association study*
PNPLA3	*patatin like phospholipase domain containing 3 gene*
PRS	score de risque polygénique
SAMM50	*SAMM50 sorting and assembly machinery component gene*
SNiPA	outil d'annotation et de balayage des variations génétiques
SNP	*single nucleotide polymorphism*
UKBB	*UK Biobank*
UKPDS	*The UK Prospective Diabetes Study*
VADT	*Veterans Affairs Diabetes Trial*
VEP	*Variant Effect Predictor*

Liste des tableaux

Tableau 1. *Nombre des gènes associés à NAFLD ou aux différentes pathologies et phénotypes en relation avec cette maladie.*

Tableau 2. *Données phénotypiques de la population ADVANCE ventilées par A) les 4 groupes de traitement et par B) les 3 niveaux des taux plasmatiques de ALT. L'intitulé des groupes de traitement est indiqué dans le texte ci-dessus. Les patients pour lesquels le taux de ALT ne sont pas disponibles ne sont pas inclus dans les tableaux.*

Tableau 3. *Potentiel de prédiction des modèles construits par le programme LASSOSUM. Ces modèles incluent les données statistiques sommaires 1) des taux plasmatiques de ALT, 2) des taux plasmatiques de AST, 3) du tour de taille et 4) des taux plasmatiques de triglycérides. Les modèles sont éventuellement ajustés par l'âge et le sexe des patients.*

Tableau 4. *Phénotypes et groupes phénotypiques correspondant aux 3210 SNPs constitutifs du PRS.*

Tableau 5. *Potentiel de prédiction du phénotype « taux élevés de ALT » et des mortalités totale, cardiovasculaire et non-cardiovasculaire par le PRS ajusté ou pas par l'âge, le sexe et la composante anthropogénétique des patients.*

Tableau 6. *SNPs constituant le groupe phénotypique « profil stéatose hépatique » du PRS*

Tableau 7. *Liste des 112 SNPs associés à la mortalité non-cardiovasculaire lors du GWAS dans ADVANCE, leurs coordonnées dans le chromosome 22 et leur localisation exonique, intronique ou intergénique.*

Liste des figures

Figure 1. *Représentation schématique du principe d'une étude pangénomique par GWAS (genome-wide association study) ainsi que des étapes majeures de l'exploitation des données qui y sont générées*

Figure 2. *Représentation d'un modèle statistique par régression linéaire.*

Figure 3. *A) Représentation schématique de l'impact des variations génétiques (VG) sur le gène dans lequel elles se situent ou sur des gènes plus éloignés. B) Association des variations génétiques, situés dans un locus, à différents aspects de la maladie.*

Figure 4. *Représentation schématique du locus PNPLA3-SAMM50*

Figure 5. *Représentation schématique de la stratégie établie pour la mise en place d'outils de marquage génétique de NAFLD et de la mortalité dans le DT2*.

Figure 6. *Représentation schématique des patients de la population ADVANCE répartis, suivant un schéma factoriel randomisé, en 4 groupes de traitement (thérapie hypoglycémiante et thérapie anti-hypertension artérielle). Schéma adapté et modifié à partir de la figure 11 de (ADVANCE-Group 2010).*

Figure 7. *Phénotypes A) quantitatifs et B) qualitatifs explorés dans l'étude*

Figure 8 : *Représentation schématique du processus de construction d'un PRS par LASSOSUM*

Figure 9 : *Croisement des listes des gènes candidats de NAFLD et de quelques altérations cardiométaboliques qui lui sont associées. Les listes de gènes ont été extraites de la base de données Disgenet (http://www.disgenet.org/search)*

Figure 10. Représentation schématique de la stratégie de construction du PRS via sélection ciblée des SNPs.

Figure 11. *Mots-clés des phénotypes et des origines des populations utilisés pour l'extraction des intitulés des phénotypes à partir des études par GWAS et des méta-analyses répertoriées dans la base de données GWAS catalog (https://www.ebi.ac.uk/gwas/).*

Figure 12. *Représentation des différentes étapes suivies pour la construction du modèle de prédiction de NAFLD et de la mortalité dans ADVANCE, via sélection ciblée des SNPs*

Figure 13. *Stratification par le PRS (deux quintiles supérieurs vs autres quintiles des scores du groupe phénotypique « profil stéatose hépatique ») des phénotypes « taux élevés de ALT » et mortalité non-cardiovasculaire. L'ajustement des valeurs a été effectué par l'âge, le sexe et le premier vecteur (PC1) de la composante anthropogénétique des patients.*

Figure 14. *Manhattan-Plot du GWAS des taux plasmatiques de l'enzyme ALT en ligne de base dans la population ADVANCE*

Figure 15. *Manhattan-plots des phénotypes A) « mortalité totale », B) « mortalité par causes non-cardiovasculaires » et C) « mortalité d'origine cardiovasculaire » au sein de la population ADVANCE.*

Figure 16. *Cartographie de l'association de 221 SNPs du locus PNPLA3-SAMM50 aux taux de ALT et à la mortalité globale ou cardiovasculaire au sein de la population ADVANCE en présence ou pas du traitement intensif hypoglycémiant*.

Figure 17. *Profil de DL déterminé par les 112 SNPs associés à la mortalité non-cardiovasculaire dans la population ADVANCE. Les positions des 8 SNPs tags sont indiqués dans la figure.*

Figure 18. *Taux plasmatiques de ALT chez les patients porteurs de chacun des trois génotypes des 8 SNPs tags.*

Figure 19. *Fréquence de la mortalité non-cardiovasculaire chez les patients porteurs de chacun des trois génotypes des 8 SNPs tags. Les nombres de patients dans chaque groupe génotypique sont indiqués sous les graphes (les individus décédés par causes cardiovasculaires sont exclus dans ce calcul des fréquences). Les fréquences de la mortalité non-cardiovasculaire sont rapportées dans les barres.*

Figure 20. *Profils **(A)** et détails **(B)** des associations les plus pertinentes des 8 SNPs tags du locus PNPLA3-SAMM50 aux composantes du DT2.*

Figure 21. *Fréquence de la mortalité non-cardiovasculaire chez les patients de ADVANCE, soumis aux traitements hypoglycémiants standard ou intensif, porteurs des génotypes homozygotes pathogènes (homozygotes des allèles associés à la mortalité non-cardiovasculaire) de chacun des 8 SNPs tags. Le nombre des patients de chaque groupe génotypique est rapporté dans les barres.* Les différences significatives sont indiquées par l'étoile.

Figure 22. *Fréquence, en présence de la thérapie hypoglycémiante intensive et de la thérapie hypoglycémiante standard, de la mortalité non-cardiovasculaire chez les patients de ADVANCE porteurs des génotypes homozygotes pathogènes de chacun des 112 SNPs associés à ce phénotype. Les carrés rouges indiquent les différences significatives (P < 0,05) entre les deux thérapies.*

Introduction

Le domaine de la génétique des maladies complexes a réalisé des avancées importantes les deux dernières décennies grâce à la mise en place de technologies de génotypage à haut débit (Ha, Freytag et al. 2014). Ces dernières ont été utilisées pour le criblage de quantités gigantesques de polymorphismes afin de comparer leurs variations entre des individus atteints de maladies complexes et d'autres sains (études pangénomiques par GWAS) (Manolio 2010). Le but est de rechercher des loci de susceptibilité. Des centaines de régions du génome ont été ainsi associées à divers phénotypes complexes (MacArthur, Bowler et al. 2017). Cependant, les études ont rapporté, dans leur majorité, des associations génétiques relativement faibles (https://www.ebi.ac.uk/gwas/). Cela est dû au caractère complexe et multigénique des maladies étudiées qui ne peut être suffisamment décrit par des polymorphismes pris individuellement. Il est donc important de combiner les variations génétiques afin de mettre en place des outils puissants pour la prédiction et le suivi des pathologies. Diverses stratégies ont été appliquées. Certaines ont consisté en la combinaison de variations génétiques issues de différentes région du génome alors que d'autres se sont focalisées sur l'exploration détaillée de certains loci (Kitamoto, Kitamoto et al. 2014, Khera, Chaffin et al. 2018). Par ailleurs, les variations génétiques exploitées peuvent être sélectionnées de manière ciblée vu leurs associations connues avec les pathologies (Domingue, Belsky et al. 2014) ou de façon non-ciblée en utilisant des outils statistiques notamment par intelligence artificielle (Mavaddat, Michailidou et al. 2019).

La stéatose non-alcoolique du foie (NAFLD) est une pathologie complexe caractérisée par l'accumulation de tissu adipeux, notamment de triglycérides et d'acides gras, au niveau du tissu hépatique (EASL-EASD-EASO. 2016, Townsend and Newsome 2016). Elle est associée à diverses altérations métaboliques et à la mortalité (Targher, Bertolini et al. 2006, Ekstedt, Hagstrom et al. 2015). Cette pathologie est plus fréquente dans le diabète de type 2 (DT2) et est associée à des profils plus délétères dans la maladie (Targher, Bertolini et al. 2006, Younossi, Koenig et al. 2016, Amiri Dash Atan, Koushki et al. 2017). NAFLD constitue une des causes de décès vu les complications qu'elle engendre, comme l'hépatocarcinome du foie (Younossi, Otgonsuren et al. 2015). La maladie survient

généralement en absence de signes cliniques et son diagnostic de référence est basé sur la mise en évidence de lésions histopathologiques au niveau du foie (Loguercio, De Simone et al. 2004, Friedman, Neuschwander-Tetri et al. 2018). Il est donc important de mettre au point des outils de détection puissants, précis et moins invasifs. A ce niveau, la génétique peut être d'un grand apport.

Dans ce contexte, les travaux réalisés au cours de ma maîtrise ont focalisé sur la mise au point de marqueurs génétiques pour la prédiction de NAFLD et des complications associées, notamment la mortalité, dans le DT2 ainsi que de l'efficacité des thérapies associées.

Mes études ont été réalisées au sein d'une population de patients avec DT2 (ADVANCE), recrutée dans le cadre d'un essai clinique établi sur 5 années (Patel, MacMahon et al. 2008). En absence de diagnostic histopathologique de NAFLD, les taux plasmatiques de l'enzyme hépatique alanine aminotransférase (ALT) ont été utilisés comme substitut. Bien que peu corrélés avec le diagnostic de la stéatose hépatique non-alcoolique (Browning, Szczepaniak et al. 2004), les niveaux plasmatiques de cette enzyme sont intéressants à étudier vu qu'ils ont été associés à diverses altérations métaboliques et à la mortalité (Ekstedt, Hagstrom et al. 2015, Martin-Rodriguez, Gonzalez-Cantero et al. 2017). Cela peut aider à mieux comprendre certaines complications du DT2 notamment la mortalité liée à l'atteinte hépatique (Wild, Walker et al. 2018).

Deux stratégies de criblage génétique ont été appliquées: la sélection non-ciblée et la sélection ciblée des marqueurs génétiques. Le but a été de mettre au point des marqueurs de NAFLD, des complications du DT2 tels que la mortalité, et de l'efficacité des thérapies associées. Des outils statistiques et bioinformatiques ainsi que des bases de données pangénomiques ont été exploités à cet effet.

Ces travaux pourraient permettre la mise en évidence de variations génétiques décrivant les diverses composantes de la stéatose hépatique non-alcoolique et du DT2. Cela peut ainsi mieux expliquer la relation entre les altérations hépatiques, cardiovasculaires et métaboliques chez les sujets diabétiques. L'étude pourrait également permettre la stratification de la population de patients afin de détecter ceux qui sont les plus à risque,

surtout de mortalité mais aussi des individus bénéficiant plus des thérapies associées au DT2.

Chapitre 1 : Maladies complexes et outils de prédiction

Contrairement aux maladies Mendéliennes qui sont causées par des mutations et des variations génétiques pénétrantes et agrégeant au sein de familles (Winsor 1988) les maladies complexes sont des altérations polygéniques et multifactorielles. En effet, la susceptibilité à des pathologies, telles que Alzheimer et les maladies cardiovasculaires est due à des variations au niveau de gènes interagissant avec des facteurs environnementaux et socioéconomiques (Migliore and Coppede 2009, Sacerdote, Ricceri et al. 2012, Patel, Chen et al. 2013, Simon, Sylvestre et al. 2016, Cooper 2018)

I- Exemple de marqueurs exploités dans l'exploration génétique des maladies complexes : les SNP (*single nucleotide polymorphisms*).

Il s'agit de variations génétiques mono-nucléotidiques, biallélique et assez bien réparties dans le génome (https://www.ncbi.nlm.nih.gov/snp/). A l'heure actuelle, 162 millions de variations mono-nucléotidiques sont répertoriées (http://db.systemsbiology.net/kaviar/). Les SNPs peuvent s'avérer fonctionnels en 1) modifiant les séquences des protéines (Bottini, Musumeci et al. 2004), 2) en changeant les niveaux d'expression génique (Yasuda, Takeshita et al. 2000) ou en modifiant l'épissage des gènes (Kralovicova, Gaunt et al. 2006).

Les SNPs ont été les variations génétiques les plus exploitées parmi les marqueurs du génome vu leur nombre important et la bonne résolution génomique qu'ils apportent mais aussi la facilité de leur génotypage (McGuigan and Ralston 2002, Hoffmann, Kvale et al. 2011).

II- Utilisation des SNPs dans l'exploration des maladies complexes.

Différentes approches ont été appliquées dans la recherche des gènes de susceptibilité aux maladies complexes. Les plus utilisées sont : 1) la stratégie du gène candidat, lorsque le polymorphisme est connu pour son implication dans le processus physiopathologique de la maladie (Bell, Horita et al. 1984) (Santoro, Cirillo et al. 2006), 2) la stratégie par liaison génétique, qui consiste en la recherche d'une agrégation entre le gène et la pathologie au

sein des familles (Duggirala, Blangero et al. 1999) (Bertram, Blacker et al. 2000) et 3) le criblage dense du génome humain (Manolio 2010).

III- La stratégie par criblage dense du génome humain.

Dite par GWAS (*genome-wide association study*), elle consiste en le criblage de l'ensemble du génome via le génotypage d'un nombre important de marqueurs génétiques (allant jusqu'à des millions de polymorphismes) dans le cadre d'études de type cas-contrôle afin de rechercher des loci dont le profil de variation est statistiquement différent entre les individus atteints de la pathologie et ou du phénotype et ceux qui y sont sains (Manolio 2010) (figure 1).

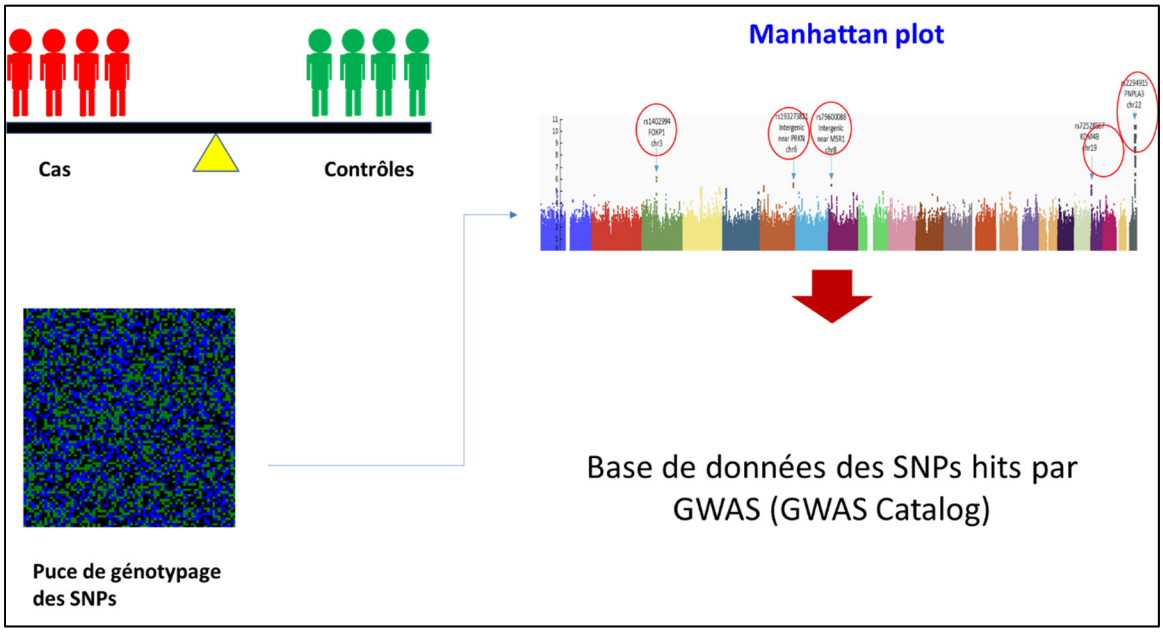

Figure 1. *Représentation schématique du principe d'une étude pangénomique par GWAS (genome-wide association study) ainsi que des étapes majeures de l'exploitation des données qui y sont générées*

Cela a été rendu possible grâce aux outils de génotypage pangénomiques tels que les plateformes Affymetrix (https://www.thermofisher.com/ca/en/home/life-science/microarray-analysis/human-genotyping-pharmacogenomic-microbiome-

solutions-microarrays.html) et Illumina (https://www.illumina.com/). Parmi les puces développées dans le cadre de la plateforme Affymetrix, la puce Affymetrix 5.0. Celle-ci permet le génotypage de 1 million de variations génétiques environ dont plus de 500000 SNPs et plus de 400000 autres polymorphismes (http://tools.thermofisher.com/content/sfs/brochures/genomewide_snp5_datasheet.pdf). La puce 6.0, quant à elle, permet un criblage plus dense du génome (1,8 millions de polymorphismes dont la moitié environ sont des SNPs) (https://www.thermofisher.com/order/catalog/product/901182). Enfin, la puce *UK-BioBANK*, comporte plus de 800000 SNPs dont une partie consiste en des polymorphismes utiles en pathologie ou en anthropogénétique (http://www.ukbiobank.ac.uk/wp-content/uploads/2014/04/UK-Biobank-Axiom-Array-Content-Summary-2014.pdf).

Bien que le nombre de SNPs génotypés via ces trois puces ne représente qu'une partie des polymorphismes actuellement connus, elles permettent une couverture génomique assez intéressante. Cela est obtenu grâce aux profils de déséquilibre de liaison (DL) existant entre les SNPs (Lewontin and Kojima 1960) mais aussi à l'inférence via *imputation* de génotypes de SNPs non inclus dans les puces (Schafer 1999) (Zeggini, Scott et al. 2008).

III-1- Notion de DL dans les études par GWAS.

Le DL consiste en la transmission simultanée des allèles des SNPs avec une fréquence plus importante que celle due au hasard (Lewontin and Kojima 1960) (Hedrick 1987). Ce phénomène est à l'origine de la diversité génétique entre les populations humaines due à la variabilité des tailles de blocs DL entre elles (Daly, Rioux et al. 2001, HapMap 2005). Divers paramètres ont été mis en avant afin d'estimer le DL. Il s'agit notamment de D' et r^2 (Lewontin 1964, Hill and Robertson 1968). Tandis que D' permet la mesure des taux de DL en prenant en considération les oscillations éventuelles des fréquences des variations génétiques dans la population (Lewontin 1964), r^2 représente le carré de la corrélation entre les deux allèles et est influencé notamment par le nombre d'individus au sein de la population explorée et par la fréquence des variations génétiques qui s'y trouvent (Hill and Robertson 1968). Des valeurs élevées de DL sont attestés lorsque D' et r^2 sont proches de 1 alors que des valeurs proches ou égales à 0 sont synonymes d'un niveau très bas de DL (Abecasis, Noguchi et al. 2001, Mueller 2004, Eberle, Rieder et al. 2006).

Le DL représente une notion importante en génétique vu qu'il permet de focaliser sur des polymorphismes représentatifs d'autres variations génétiques (Carlson, Eberle et al. 2004) et réduire ainsi les efforts en matière de temps et de conception des outils d'analyse. Il facilite par conséquence la recherche des loci impliqués dans les pathologies.

Diverses bases de données ont été établies dans le but de décrire les profils de DL au sein de différentes populations mondiales. Il s'agit notamment de 1) la base de données *HapMap International*, actuellement obsolète (HapMap 2003, HapMap 2005) et 2) la base de données *1000 Genomes* (Auton, Brooks et al. 2015). Cette dernière regroupe les profils de DL de 26 populations issues de diverses parties du monde, dont la population d'origine caucasienne CEU (recrutée aux États-Unis) (Auton, Brooks et al. 2015).

Par ailleurs, des logiciels, comme Haploview (Barrett, Fry et al. 2005), ont été mis en place afin d'exploiter les données de DL disponibles au sein de *HapMap International* ou dans *1000 Genomes*. Cet outil présente une interface interactive et permet la représentation graphique des profils de DL de la population analysée. Il peut être aussi exploité pour marquer (taguer) une région génomique en se basant sur les niveaux de DL entre les polymorphismes (Barrett, Fry et al. 2005).

Effectuer une bonne étude pangénomique impose l'utilisation d'outils statistiques de qualité. Cela assure un maximum d'efficacité dans la détection des loci d'intérêt tout en évitant les artéfacts.

III-2- Outils statistiques pour l'exploitation des données des études par GWAS.

Divers outils statistiques ont été appliqués pour analyser les données d'association génétique dans le cadre des études par GWAS. Parmi elles :

- **La régression linéaire.**

  Cette analyse concerne les phénotypes quantitatifs, tels que l'indice de masse corporelle (IMC) ou la glycémie. Elle consiste à rechercher une linéarité en termes de relation entre le phénotype quantitatif et la variation du gène au sein de la population (Schneider, Hommel et al. 2010, Buzkova 2013) (figure 2). La régression linéaire établit un modèle via l'équation :

$$y = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \beta_0 + \varepsilon$$

où $\beta_1 \ldots \beta_n$ représentent les impacts des variables indépendantes (variables génétiques par exemple), $X_1 \ldots X_n$ les variations indépendantes dans le modèle tandis que l'erreur de prédiction du phénotype par le modèle est représentée par ε (Schneider, Hommel et al. 2010, Yu, Yao et al. 2014).

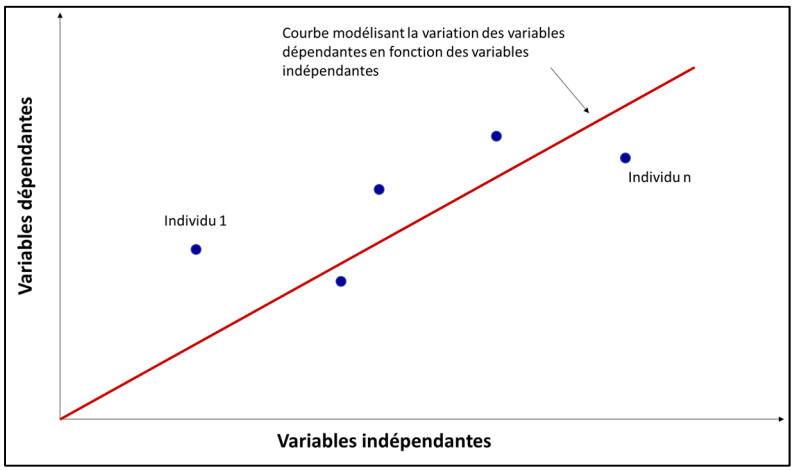

Figure 2. *Représentation d'un modèle statistique par régression linéaire.*

L'application de la régression linéaire au sein de populations de faibles tailles nécessite généralement l'utilisation d'un phénotype ayant une distribution gaussienne au sein de la population (Altman and Bland 1995, Dobson 2001). Divers phénotypes, notamment en biomédecine n'ont pas de distribution normale et le deviennent après transformation logarithmique (Limpert, Stahel et al. 2001). Celle-ci peut être à l'origine d'une diminution de la variance du phénotype au sein de la population ce qui permet l'augmentation de la puissance des analyses statistiques (Norton and Strube 2001, Warner 2013). Par ailleurs, la distribution normale permet, surtout lorsque les populations sont de faibles tailles, l'exploitation de tests paramétriques très fréquemment utilisés en statistiques, tels que t-test et ANOVA (Nayak and Hazra 2011).

- **La régression logistique.**

Permet de calculer l'association génétique à un phénotype binaire comme l'atteinte par une maladie ou pas (Talmud, Cooper et al. 2015). Cette analyse permet de déterminer, via calcul de l'*odds-ratio* (OR), l'augmentation ou la diminution du risque d'avènement de la pathologie en présence d'un allèle, d'un génotype ou d'un score donné (Reddy, Wang et al. 2011, Talmud, Cooper et al. 2015).

- **Modèle linéaire généralisé (GLM).**

Ce dernier peut être utilisé pour réaliser diverses types de modélisation, telles que linéaires ou logistiques (Nelder and Wedderburn 1972, Muller 2004). Il permet de rechercher le modèle de régression décrivant au mieux le profil de variation d'une variable dépendante en établissant une fonction de lien *f*:

***Profil de variation* = *f*(modèle de régression)**

Il est ainsi possible de décrire le profil de variation d'une variable dépendante binomiale en appliquant la fonction logarithme (log) alors que la fonction identité reste la plus appropriée pour modéliser une variable dépendante quantitative normalement distribuée (Nelder and Wedderburn 1972, Muller 2004).

En génétique, GLM est généralement exploité afin d'inclure plusieurs variations génétiques (Nelder and Wedderburn 1972, Anche, Bijma et al. 2015).

GLM, notamment sous certaines versions, comme le modèle linéaire généralisé mixte, représente un outil statistique de choix dans les études pangénomiques. En effet, cette analyse est capable de prendre en considération de multiples variations génétiques mais aussi des variations confondantes, telles que l'âge ou la composante anthropogénétique (Hoffman 2013, Hunter, Robinson et al. 2016, Moses 2016). A noter que les phénotypes binaires étudiés par GLM ont généralement des profils de distribution binomiale ou surviennent de manière répétée suivant un modèle de poisson (Goupil, Brachemi et al. 2013, Costa-Urrutia, Abud et al. 2017) alors que les phénotypes quantitatifs sont en général

normalement distribués ou rendus ainsi en rapportant leur distribution à une autre échelle, logarithmique par exemple (Kathiresan, Manning et al. 2007, Ostchega, Porter et al. 2011, Zhou, Liang et al. 2012).

Bien que la régression statistique soit un outil d'analyse performant, il demeure qu'effectuer une multitude de tests statistiques en même temps puisse être à l'origine d'erreurs statistiques de type I dites également associations faussement positives (Chen, Feng et al. 2017). Pour y pallier, on réalise des corrections des niveaux d'association par le nombre de tests effectués via des tests comme celui de Bonferroni (NCBI 2015).

III-3- Limites des études par GWAS.

Les investigations par GWAS ont permis des avancées considérables dans la compréhension de la composante génétique des phénotypes et des pathologies complexes vu qu'elles ont mis en avant des associations génétiques déjà connues (Rung, Cauchi et al. 2009) ou pas (Frayling, Timpson et al. 2007). Les résultats de ces études ont été regroupés dans la base de données *GWAS Catalog* (MacArthur, Bowler et al. 2017) (https://www.ebi.ac.uk/gwas/) afin d'être exploités par la communauté scientifique.

Malgré cela, ces études ont des limites. En effet, les associations génétiques qui y sont obtenues concernent des SNPs fréquents pris séparément et restent souvent faibles (ampleurs d'effets ou OR généralement bas) (https://www.ebi.ac.uk/gwas/). En outre, l'effet des variations génétiques peu fréquentes (MAF < 1%) est peu étudié dans les maladies complexes vu la taille relativement faible des populations et cohortes utilisées (Zuk, Schaffner et al. 2014). Pour y pallier, des consortiums ont été établis afin de mutualiser les données des investigations et effectuer des méta-analyses (Fuchsberger, Flannick et al. 2016, Li, Li et al. 2017). Des démarches de recrutement massif d'échantillons ont été également entamées, à l'image du projet *UK Biobank* (Bycroft, Freeman et al. 2018) (https://www.ukbiobank.ac.uk/). Le but a été de constituer des populations de très grandes tailles dans lesquelles les données de génotypage de millions de SNPs et de caractérisation de milliers de phénotypes sont incluses. Par ailleurs, des stratégies et des outils statistiques ont été mis en place ou sont en cours de développement afin d'optimiser la capture de l'héritabilité des maladies (Vilhjalmsson, Yang et al. 2015, Mak, Porsch et al. 2017, Choi, Mak et al. 2018). Ces derniers permettent, entre autres, la

combinaison des données d'association génétique afin d'expliquer les mécanismes physiopathologiques qui y sont à l'origine, prédire la maladie et stratifier les patients (Mega, Stitziel et al. 2015, Khera, Chaffin et al. 2018).

IV- Modélisation statistique pour la prédiction et la stratification des maladies complexes.

Diverses stratégies peuvent être utilisées pour sélectionner les variations génétiques destinées à la construction des modèles de risque. Il est possible d'inclure un maximum voire la totalité des variations génétiques issues des études par GWAS en admettant le caractère hautement polygénique de la pathologie (Boyle, Li et al. 2017, Khera, Chaffin et al. 2018) comme il est concevable de se focaliser sur certains polymorphismes et construire un score de risque polygénique (PRS) (Pilling, Kuo et al. 2017).

Différents processus sont utilisés pour la construction des PRS. La méthode la plus classique consiste en 3 phases (Chatterjee, Shi et al. 2016). La première permet la détermination du risque relatif de la maladie via la combinaison de facteurs de risque génétiques et d'autres environnementaux. La seconde étape consiste en la combinaison du risque relatif ainsi calculé avec le risque d'atteinte par la maladie au sein de la population générale, ce qui permet l'obtention du modèle. La dernière étape a pour but de valider le modèle (Chatterjee, Shi et al. 2016).

Construire un modèle de prédiction de risque de bonne qualité impose l'absence de toute surexploitation des données de la population analysée (Choi, Mak et al. 2018). Cette surestimation (*overfitting*) peut rendre le modèle peu efficace voire inefficace dans la caractérisation de la pathologie sur d'autres populations (Nunez, Steyerberg et al. 2011). Il est donc important que les populations utilisées pour la construction et le test du modèle soient distinctes (Nunez, Steyerberg et al. 2011, Choi, Mak et al. 2018).

Le potentiel d'un modèle de risque est également apprécié en termes de sa capacité à stratifier les individus en fonction du risque qu'ils ont à développer la pathologie. Cet aspect peut s'avérer crucial dans la prise en charge médicale et la rationalisation des moyens thérapeutiques. En effet, l'investigation réalisée dans trois populations par Mega et ses collègues a permis la stratification des individus en utilisant un PRS de 27 SNPs (Mega, Stitziel et al. 2015). L'étude a constaté un bénéfice plus marqué de la thérapie par

statines chez les patients atteints de la maladie coronarienne ayant les scores du modèle les plus élevés.

L'estimation du potentiel de prédiction peut être faite via le calcul de la surface sous la courbe (AUC : *area under the receiver operating characteristic*), appelée aussi *c-statistics* (Khera, Chaffin et al. 2018). Cela permet d'apprécier la sensibilité et la spécificité du modèle (Fawcett 2003). Des potentiels de prédiction intéressants sont attestés en présence de valeurs AUC s'éloignant de 0,5 et se rapprochant plus de 1 (Wang, Hu et al. 2016). A noter que le potentiel de prédiction d'une maladie, en utilisant AUC par exemple, dépend de la capacité des facteurs de susceptibilité à décrire cette maladie mais aussi de la prévalence de la pathologie au sein de la population (Lewis, Whitwell et al. 2007, Wray, Yang et al. 2010). L'estimation de la variance de la maladie peut être également obtenue via le coefficient de corrélation R^2 (Barreira, Broyles et al. 2014, Alexander, Tropsha et al. 2015). Ce dernier varie souvent entre 0 et 1 (Kvalseth 1985). Une valeur R^2 égale à 0 est synonyme d'une absence de prédiction des individus atteints du phénotype étudié alors que le modèle est attesté comme capable de détecter l'ensemble des individus porteurs du phénotype lorsque $R^2 = 1$ (Kvalseth 1985, Taylor 1990). A noter enfin que la correction du potentiel du PRS par le nombre de tests multiples effectués peut s'avérer intéressante à réaliser (Warrier and Baron-Cohen 2018). Cela demeure toutefois non obligatoire dans la conception des modèles de risque (Choi, Mak et al. 2018, Khera, Chaffin et al. 2019).

La sélection des SNPs candidats dans le modèle peut être manuelle et consiste en la recherche de ceux qui sont le plus en association avec le phénotype ou ses composantes (Belsky and Israel 2014). Il est également possible de sélectionner les SNPs en absence de toute hypothèse en utilisant des outils statistiques adéquats tels que l'analyse LASSO (Tibshirani 1996).

IV-1- Méthode statistique LASSO.

La méthode LASSO (*least absolute shrinkage and selection operator*) est une approche par régression statistique (Tibshirani 1996). Elle permet la conception de modèles de prédiction à partir des données pangénomiques en des temps relativement faibles en comparaison avec d'autres méthodes (Hepp, Schmid et al. 2016). De plus, LASSO est assez efficace pour la prédiction des phénotypes les moins fréquents au sein de la population

(Ambler, Seaman et al. 2012) et pour la recherche des facteurs de susceptibilité à la pathologie (Lu, Zhou et al. 2017).

Cette méthode permet la pénalisation du vecteur des ampleurs d'effets des variables afin de ramener une partie des coefficients de régression à une valeur nulle, diminuant ainsi les dimensions de la matrice des variables indépendantes (Tibshirani 1996). Cela est réalisé suivant l'équation :

$$\widehat{\boldsymbol{\beta}} = \underset{\beta_1,\ldots,\beta_p}{argmin} \sum_{i=1}^{N} (y_i - \sum_j \beta_j \, x_{ij})^2 \;\; sous\ la\ contrainte \sum_{j=1}^{P} |\beta_j| \leq t$$

où p représente le nombre de variables prédictives incluses dans l'analyse et N le nombre d'individus étudiés. Par ailleurs, x_{ij} sont les variables de prédiction (variations génétiques par exemple) et Y_i représente la présence ou non du phénotype chez le $i^{ème}$ individu. A noter que $\widehat{\boldsymbol{\beta}}$ est le vecteur des ampleurs d'effet après pénalisation tandis que $\sum_{i=1}^{N} (y_i - \sum_j \beta_j x_{ij})^2$ représente la somme des carrés des résidus de la régression statistique et **t** est le seuil établi dans la pénalisation des ampleurs d'effet (Tibshirani 1996).

La méthode sous sa version de base reste relativement limitée en termes de nombre de polymorphismes et de biomarqueurs sélectionnés (Zou and Hastie 2005). Cela peut limiter la puissance du modèle de prédiction. Pour y pallier, des analyses statistiques par pénalisation comme *Elastic Net* ont été développées (Zou and Hastie 2005). De plus, des programmes d'exploitation, comme LASSOSUM, dédiés au criblage des données statistiques sommaires issues des études pangénomiques ont été mis en place (Mak, Porsch et al. 2017).

- **Le programme LASSOSUM.**

LASSOSUM a été développé par un groupe de recherche à Hong Kong en modifiant l'équation de base de LASSO (Mak, Porsch et al. 2017). Cette modification a permis d'exploiter les ampleurs d'effet (bêtas ou OR) rapportés dans la littérature ainsi que le profil de DL d'une population de référence pour la construction de modèles de prédiction en utilisant des populations dont disposent les chercheurs (Mak, Porsch et al. 2017).

Les profils de DL exploités dans l'analyse sont issus de populations d'origine européenne, africaine ou asiatique, répertoriées dans la base de données *1000 Genomes* (Berisa and Pickrell 2016). Il est également possible d'exploiter les profils de DL d'autres populations. Par ailleurs, le programme permet la validation du modèle de prédiction même s'il n'y a pas de population dédiée, grâce à un processus dit « pseudo-validation » (Mak, Porsch et al. 2017).

Bien que la combinaison de marqueurs génétiques issus de diverses régions génomiques puisse s'avérer très efficace dans la prédiction et la stratification des maladies complexes, l'analyse génétique détaillée de certains loci, notamment ceux ayant un fort impact dans la pathologie, peut être intéressante.

V- Criblage des loci de susceptibilité aux maladies complexes.

Les variations génétiques rapportées en association à la maladie peuvent avoir des effets au sein des gènes dans lesquels elles se situent mais également sur des gènes plus éloignés (figure 3 (A)) (Stratigopoulos, Padilla et al. 2008, Jowett, Curran et al. 2010, Claussnitzer, Dankel et al. 2015). Par ailleurs, des groupes de polymorphismes au sein d'un locus de susceptibilité peuvent révéler divers aspects de la maladie, tels que son avènement et son profil (figure 3 (B)) (Kitamoto, Kitamoto et al. 2014).

Les avancées statistiques réalisées jusque-là ont permis de mettre à disposition des outils intéressants pour la conception de modèles de risque (Vilhjalmsson, Yang et al. 2015, Mak, Porsch et al. 2017). L'exploitation de ces outils pour prédire les phénotypes complexes, notamment ceux encore peu explorés comme les complications du diabète de type 2 (DT2), est une opportunité pour les chercheurs et une lueur d'espoir pour les patients.

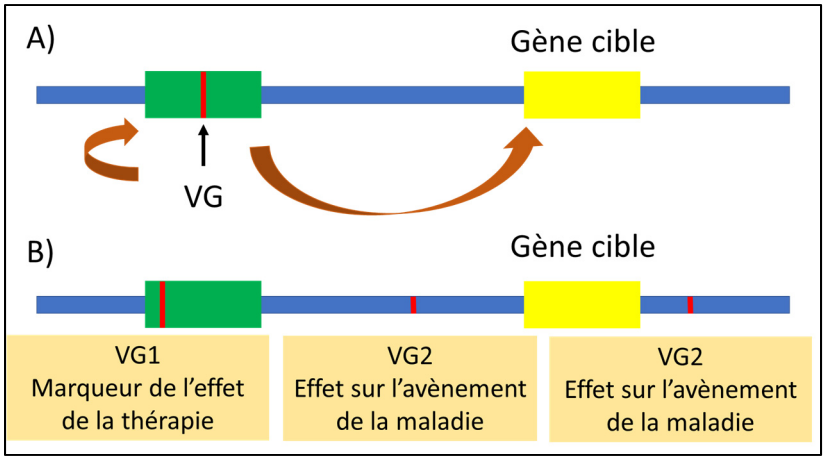

Figure 3. *A) Représentation schématique de l'impact des variations génétiques (VG) sur le gène dans lequel elles se situent ou sur des gènes plus éloignés. B) Association des variations génétiques, situés dans un locus, à différents aspects de la maladie.*

Chapitre 2 : Diabète de type 2 et complications

Le DT2 est une maladie complexe caractérisée par la diminution de l'action de l'hormone hypoglycémiante « insuline » sur les organes cibles (résistance à l'insuline) (Perley and Kipnis 1966). Cette maladie représente la partie majeure des diabètes (https://idf.org/52-about-diabetes.html).

De plus en plus d'individus sont atteints de DT2 à travers le monde. En effet, la prévalence de la maladie, selon l'Organisation Mondiale de la Santé, a été de 4,7% en 1980 et a presque doublé pour atteindre 9% environ un quart de siècle plus tard (https://www.who.int/news-room/fact-sheets/detail/diabetes). Ce constat alarmant serait dû à l'épidémie mondiale d'obésité et de surcharge pondérale (https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight) accompagnant un terrain évident de susceptibilité génétique (Medici, Hawa et al. 1999, Grant, Thorleifsson et al. 2006).

Le DT2 constitue une problématique majeure de santé publique vu les complications qu'il engendre. Celles-ci peuvent être d'ordre microvasculaire, macrovasculaire ou hépatique.

I- Complications microvasculaires du DT2.

Les atteintes microvasculaires dans le DT2 sont dues en grande partie à l'élévation chronique de la glycémie. En effet, l'hyperglycémie est à l'origine de l'activation de processus moléculaires engendrant, entre autres, un état de stress oxydant et un dépôt de résidus glucidiques, à l'origine d'altérations des tissus (Hammes, Martin et al. 1991, Lee and Chung 1999, Sato, Sato et al. 2005, Ishibashi, Yamagishi et al. 2012). Les altérations microvasculaires ont été rapportées comme plus fréquentes en présence de l'hypertension artérielle de même que la thérapie antihypertensive a permis de réduire leur présence et a été à l'origine d'un pronostic plus favorable chez les patients diabétiques (Teuscher, Schnell et al. 1988, Brenner, Cooper et al. 2001). Parmi les complications microvasculaires du DT2 :

I-1- La rétinopathie diabétique.

La rétinopathie diabétique constitue une des complications invalidantes chez les sujets diabétiques vu qu'elle est à l'origine de cécité (Leasher, Bourne et al. 2016). A des phases

précoces de la maladie, la rétine est alimentée par un système vasculaire normal. Elle est dite non-proliférative. Plus tardivement, une structure vasculaire anormale apparait au niveau du tissu rétinien. A ce stade la rétinopathie est appelée proliférative (Shah and Chen 2011).

I-2- La néphropathie diabétique.

Il s'agit d'une altération rénale dont les taux dans le DT2 dépassent 40% (Afkarian, Sachs et al. 2013). Elle comporte 5 étapes en fonction du degré de détérioration de la fonction rénale (Haneda, Utsunomiya et al. 2015). Le suivi de la maladie est réalisé via dosage des taux urinaires de l'albumine et via estimation des niveaux de la filtration glomérulaire (Haneda, Utsunomiya et al. 2015). La néphropathie diabétique constitue un enjeu de santé publique vu son association à des taux de mortalité plus élevés (Afkarian, Sachs et al. 2013) mais aussi à cause des coûts faramineux alloués à la prise en charge médicale notamment en termes de dialyse et de transplantation rénale (Gordois, Scuffham et al. 2004, Nichols, Vupputuri et al. 2011, Jarl, Desatnik et al. 2018).

II- Complications macrovasculaires du DT2.

Parmi elles :

II-1- L'accident vasculaire cérébral.

L'accident vasculaire cérébral (AVC) est présent à environ 8% dans le DT2 (Einarson, Acs et al. 2018). Une plus grande susceptibilité à l'AVC a été rapportée chez les patients diabétiques (Cui, Iso et al. 2011, Khoury, Kleindorfer et al. 2013). Cette relation entre les diabètes et l'AVC aurait diverses causes parmi lesquelles la résistance à l'insuline (Rundek, Gardener et al. 2010) et le stress oxydant (Chan 1996, Baynes and Thorpe 1999). Par ailleurs, une susceptibilité plus importante aux AVC dans les diabètes serait due à la présence plus marquée d'individus atteints d'hypertension artérielle parmi les sujets diabétiques (MacMahon, Peto et al. 1990, Kabakov, Norymberg et al. 2006, Mills, Bundy et al. 2016).

II-2- L'infarctus du myocarde.

Les sujets atteints de DT2 représentent une population à plus grand risque d'infarctus du myocarde. Cela a été rapporté dans une étude réalisée au sein d'une population du

Royaume-Uni (Liang, Vallarino et al. 2014). Celle-ci a mis en avant une incidence d'infarctus de myocarde plus prononcée chez les patients diabétiques de type 2.

III- Complications hépatiques dans DT2.

Plusieurs pathologies du foie ont été associées au DT2. Alors que certaines d'entre elles sont d'origine infectieuse (à titre d'exemple, l'hépatite C coexiste plus avec le DT2 chez les personnes d'une certaine tranche d'âge) (Mehta, Brancati et al. 2000), d'autres sont plutôt métaboliques, telles que la stéatose hépatique non-alcoolique (Younossi, Koenig et al. 2016, Amiri Dash Atan, Koushki et al. 2017, Cusi, Sanyal et al. 2017).

Chapitre 3 : Stéatose hépatique non-alcoolique

La stéatose hépatique non-alcoolique, dite également la maladie du foie gras non-alcoolique (NAFLD), est une anomalie atteignant le foie et est attestée par l'agglomération du tissu adipeux sur plus de 5% de la surface hépatique chez des individus non-alcooliques (EASL-EASD-EASO. 2016). Elle atteint un individu sur 4 en moyenne (Younossi, Koenig et al. 2016). NAFLD comporte différents niveaux allant des stéatoses les plus banales jusqu'à des stades graves caractérisés par la cirrhose du foie (Chalasani, Younossi et al. 2018). La maladie peut également évoluer vers le carcinome hépatique (Younossi, Otgonsuren et al. 2015).

La complication du NAFLD vers le carcinome hépatique reste peu fréquente (Younossi, Otgonsuren et al. 2015) et des évolutions moins morbides peuvent être observées. Il peut s'agir d'un état inflammatoire également appelé hépato-stéatite non alcoolique (NASH) (EASL-EASD-EASO. 2016). Les hépatocytes changent de morphologie à ce stade et un état de fibrose hépatique peut s'y installer (Chalasani, Younossi et al. 2018). Cet état de fibrose serait lié au décès des patients (Ekstedt, Hagstrom et al. 2015).

I- Données statistiques sur la prévalence du NAFLD.

Bien que NAFLD soit une pathologie très présente chez les individus ayant une surcharge pondérale (Li, Liu et al. 2016), elle a été également mise en évidence chez des sujets de poids normal (Younossi, Stepanova et al. 2012). Des taux élevés de la maladie ont été mis en avant dans différentes régions de la planète. En effet, 13,5% de la population africaine serait atteinte de NAFLD, de même que les taux avoisinent 24% et 30% environ au sein des populations d'Amérique du Nord et d'Amérique du Sud, respectivement (Younossi, Koenig et al. 2016). Cette disparité des taux de la pathologie entre les populations a été d'ailleurs corroborée dans une récente investigation (Kim, Kim et al. 2019).

II- Physiopathologie du NAFLD.

La résistance à l'insuline est un désordre métabolique très présent dans NAFLD (Sanyal, Campbell-Sargent et al. 2001). Cette relation est médiée par des adipokines insulino-sensiblisantes comme l'adiponectine ou des cytokines impliquées dans l'inflammation (Bugianesi, Pagotto et al. 2005, De Taeye, Novitskaya et al. 2007). L'installation de la résistance à l'insuline est accompagnée par l'augmentation de la lipolyse au niveau du tissu adipeux sous-cutané et par l'élévation de la néolipogenèse au niveau hépatique (Eissing, Scherer et al. 2013, Verboven, Wouters et al. 2018). La récupération des triglycérides et des acides gras par le foie ainsi que l'excès de production des lipides au niveau hépatique engendrent l'accumulation lipidique dans les hépatocytes et causent la stéatose.

La stéatose hépatique non-alcoolique est associée à diverses altérations et pathologies métaboliques, ainsi :

III- NAFLD et DT2.

NAFLD est une pathologie qui atteint un individu sur 4 et sa prévalence s'avère plus importante dans le DT2 (Younossi, Koenig et al. 2016, Amiri Dash Atan, Koushki et al. 2017, Cusi, Sanyal et al. 2017). Cela serait lié au contexte de résistance à l'insuline accompagnant le foie stéatosé non-alcoolique (Bugianesi, Gastaldelli et al. 2005). L'insulinorésistance est également à l'origine d'effets toxiques sur le foie via l'hyperglycémie et l'hypertriglycéridémie chroniques qu'elle engendre (Olefsky, Farquhar et al. 1974, Ota, Takamura et al. 2007, Shibata, Ichikawa et al. 2008). NAFLD est par ailleurs responsable de processus malins au niveau du foie, notamment l'hépatocarcinome (Kanwal, Kramer et al. 2018). La mortalité par les cancers du foie a été d'ailleurs rapportée comme étant un peu plus fréquente chez les patients diabétiques par rapport à ceux non-atteints de diabète (Campbell, Newton et al. 2012). A noter aussi que les niveaux de mortalité chez les diabétique de type 2 deviennent plus marqués lorsque ceux-ci ont un foie atteint de stéatose non-alcoolique (Wild, Walker et al. 2018).

III- NAFLD et altérations cardiométaboliques.

La présence plus marquée des altérations cardiovasculaires et de NAFLD chez les patients avec DT2 peut expliquer la relation rapportée entre la stéatose hépatique non-alcoolique et

la pathologie cardiovasculaire (Kabakov, Norymberg et al. 2006, Younossi, Koenig et al. 2016, Cusi, Sanyal et al. 2017, Motamed, Rabiee et al. 2017). Des travaux ont mis en évidence la modulation de certaines structures artérielles chez les patients avec NAFLD et syndrome métabolique et attestent de l'association de la maladie avec l'athérosclérose (Kim, Kim et al. 2009). Cependant, un résultat opposé a été constaté en présence de DT2 (Loffroy, Terriat et al. 2015).

Par ailleurs, l'hypertension artérielle a été rapportée comme plus présente chez les patients atteints des formes avancées de la stéatose hépatique non-alcoolique (Ryoo, Suh et al. 2014, Younossi, Koenig et al. 2016). D'un autre côté, NAFLD a été lié aux stades les plus avancés de l'AVC (Li, Hu et al. 2018).

IV- Autres facteurs influençant NAFLD.

Parmi ces facteurs, l'âge des patients. En effet, des travaux ont mis en évidence la différence de la prévalence de la maladie en fonction de l'âge de la population (Younossi, Stepanova et al. 2012). Cette disparité serait liée au vieillissement des hépatocytes au sein desquelles la stéatose s'installe plus facilement (Ogrodnik, Miwa et al. 2017). NAFLD serait par ailleurs influencé par le sexe des patients (Caballeria, Pera et al. 2010, Eguchi, Hyogo et al. 2012) mais aussi par leur composante ethnique (Rich, Oji et al. 2018, Kim, Kim et al. 2019).

V- Mortalité liée à NAFLD

La stéatose hépatique est une maladie survenant généralement en absence de signes cliniques et n'est de manière générale diagnostiquée que par hasard dans le cadre d'analyses médicales de routine (Loguercio, De Simone et al. 2004, Friedman, Neuschwander-Tetri et al. 2018). A cause de ça, le diagnostic de la maladie n'est établi que tardivement lorsque cette dernière arrive à des stades avancés, ce qui rend sa prise en charge difficile. NAFLD est associé à un risque plus élevé de mortalité. Cela a été attesté par exemple par Nseir et ses collègues, qui ont rapporté des taux de décès plus importants chez les personnes ayant un foie stéatosé et atteintes de pneumonie (Nseir, Mograbi et al. 2019). La fibrose hépatique constitue un état fortement lié au décès des patients avec NAFLD. En effet, l'investigation effectuée dans une population de Suède a mis en avant un taux de décès plus important (HR = 1,29) en présence de NAFLD et qui le devient d'avantage chez

les patients avec NAFLD ayant une fibrose hépatique avancée (HR = 3,3) (Ekstedt, Hagstrom et al. 2015). Par ailleurs, l'origine cardiovasculaire du décès des patients atteints de NAFLD a été mise en avant dans l'étude réalisée par Haflidadottir et ses collègues. Cette dernière atteste qu'environ la moitié des individus décédés ont des altérations cardiovasculaires comme origine et 7% des décès sont la conséquence d'anomalies au niveau du foie (Haflidadottir, Jonasson et al. 2014). Le taux faible de mortalité attribué aux problèmes hépatiques dans cette étude serait sous-estimé dans le sens où les atteintes hépatiques telles que NAFLD sont largement liées aux altérations cardiovasculaires, comme je l'ai précédemment mentionné (Motamed, Rabiee et al. 2017).

Le chevauchement des mécanismes physiopathologiques à l'origine de l'atteinte hépatique et des anomalies cardiométaboliques permet d'envisager une exploitation de thérapies dédiées au traitement de maladies comme le DT2 ou à l'hypertension dans la prise en charge de NAFLD.

VI- Thérapies antidiabétiques dans le traitement de NAFLD.

Diverses classes thérapeutiques utilisées dans la prise en charge du DT2 ont fait l'objet d'études afin de tester leur impact sur NAFLD. Parmi elles, les thiazolidinediones (Mayerson, Hundal et al. 2002, Musso, Cassader et al. 2017), les analogues de GLP1 (Ding, Saxena et al. 2006, Petit, Cercueil et al. 2017), les biguanides et notamment la metformine (Sofer, Boaz et al. 2011) et les sulfonylurées tels que gliclazide (Feng, Gao et al. 2017).

- **Gliclazide.**

Il s'agit d'une molécule appartenant à la classe des sulfonylurées. Cette classe thérapeutique est exploitée pour la régulation de la glycémie chez les patients diabétiques (Gotfredsen 1976, Kolterman, Gray et al. 1984). Ces médicaments peuvent également agir sur des processus métaboliques comme la lyse lipidique (Shi, Moustaid-Moussa et al. 1999). Lorsqu'elles sont appliquées de manière intensive, notamment en association avec d'autres molécules, les sulfonylurées améliorent le pronostic du DT2 via l'amélioration du profil glucidique et des niveaux de la tension artérielle (Patel, MacMahon et al. 2008, Zoungas, Chalmers et al. 2014). Dans ce contexte, l'utilisation des sulfonylurées dans la thérapie du foie stéatosé chez les diabétiques de type 2 peut s'avérer intéressante.

Récemment, une investigation effectuée dans une population de Chine avec DT2 a permis de constater une régression considérable de la stéatose hépatique chez des patients atteints de NAFLD suite à la prise de gliclazide (Feng, Gao et al. 2017). La thérapie croisée du NAFLD et du DT2 apporte ainsi une preuve supplémentaire de la relation existant entre les deux maladies.

VII- Outils de diagnostic non-invasif de NAFLD.

Les complications associées à NAFLD précédemment rapportées mettent en avant la nécessité d'établissement d'outils de diagnostic performants afin de détecter la maladie notamment à des phases précoces. Le diagnostic de référence reste à l'heure actuelle l'analyse histopathologique des tissus hépatiques (Adams, Sanderson et al. 2005). Cette démarche est cependant invasive. Afin d'y pallier, des outils tels que l'imagerie ont été développés dans le but de détecter NAFLD et suivre son évolution (Lee and Park 2014).

A- L'imagerie médicale dans le diagnostic et le suivi de NAFLD.

Diverses technologies d'imagerie ont été appliquées pour diagnostiquer la stéatose hépatique non-alcoolique et déterminer ses divers stades. Tandis que certaines, comme l'ultrasonographie, utilisent des sons à haute fréquence pour la création d'images (Hassani 1974, LaBrecque, Abbas et al. 2014), d'autres mettent à profit des rayons X, comme la tomodensitométrie (Goldman 2007, Wells, Li et al. 2016), ou les propriétés des molécules sous l'effet d'un champ magnétique, telle que l'imagerie par résonnance magnétique (IRM), pour la caractérisation de la maladie (Grover, Tognarelli et al. 2015, Caussy, Alquiraish et al. 2018). Enfin des technologies dites par élastographie permettent de mesurer la modification des niveaux de rigidité de l'organe due à une éventuelle accumulation de fibres au niveau du tissu hépatique (Jiang, Huang et al. 2018).

Les performances de ces technologies restent plus ou moins importantes à divers stades de la maladie. En effet, l'ultrasonographie et l'IRM sont plus adaptées à des stéatoses hépatiques plus prononcées alors que l'imagerie par élastographie est plutôt intéressante à utiliser pour le diagnostic et le suivi de la fibrose au sein du foie (LaBrecque, Abbas et al. 2014, Imajo, Kessoku et al. 2016). L'ensemble de ces données met en évidence l'intérêt de combiner diverses technologies d'imagerie afin de bien caractériser NAFLD.

B- Marqueurs biologique et scores clinico-biologiques pour le diagnostic et le suivi de NAFLD.

Diverses enzymes hépatiques voient leurs taux plasmatiques modifiés en présence de NAFLD (Mathiesen, Franzen et al. 1999). Parmi elles :

B-1- L'alanine aminotransférase (ALT):

Bien que l'élévation des taux de ALT dans le plasma puisse parfois révéler l'existence de NAFLD, cette élévation ne corrèle que peu avec le diagnostic histopathologique du foie stéatosé non-hépatique. En effet, l'étude réalisée par Browning et ses collègues a démontré que 20 individus uniquement sur 100 atteints de stéatose hépatique ont une augmentation des taux sériques de ALT (Browning, Szczepaniak et al. 2004). Cela met en évidence, et à première vue, la capacité limitée qu'a cette enzyme à marquer la maladie. Cependant, les taux de ALT ont été rapportés comme liés à des désordres métaboliques tels que la résistance à l'insuline, la dyslipidémie et la surcharge pondérale (Yoo, Lee et al. 2008). Ils ont également été liés à la mortalité par causes cardiovasculaires et par diabètes (Yun, Shin et al. 2009) et à la mortalité associée au foie infectieux (Shim, Kim et al. 2018). L'association entre ALT et la mortalité reste cependant très variable et apparait plus prononcée chez les individus appartenant aux tranches d'âges les plus élevées (Liu, Ning et al. 2014, Schmilovitz-Weiss, Gingold-Belfer et al. 2018).

Par ailleurs, la sélection des individus ayant les taux les plus élevés de ALT (60 UI/L) permet de mieux focaliser sur des formes compliquées de la maladie, tels que NASH (Fedchuk, Nascimbeni et al. 2014). L'ensemble de ces données démontre l'intérêt de l'utilisation de biomarqueurs comme ALT dans le suivi de NAFLD en absence de diagnostic histopathologique ou par imagerie médicale.

B-2- L'aspartate aminotransférase (AST):

Comme l'ALT, l'AST est une enzyme dont les taux varient dans la pathologie hépatique (Sookoian, Castano et al. 2016). Les taux de AST combinés à ceux de ALT permettent la distinction entre la stéatose hépatique non-alcoolique et celle due à la surconsommation d'alcool. En effet, un rapport AST : ALT relativement bas atteste plutôt du foie stéatosé non-alcoolique inflammatoire alors qu'un ratio plus élevé révèle une altération hépatique due à la prise excessive d'alcool (Sorbi, Boynton et al. 1999).

Des scores clinico-biologiques de prédiction de NAFLD ont été établis. Ainsi, l'étude de Poynard et ses collègues a permis la conception d'un modèle (*SteatoTest*) incluant des biomarqueurs sanguins, l'IMC, le cholestérol sérique et la glycémie, capable de prédire la stéatose du foie avec un AUC avoisinant 0,80 (Poynard, Ratziu et al. 2005). D'un autre côté, un modèle de prédiction, incluant des biomarqueurs parmi lesquels le taux plasmatique de ALT ainsi que des paramètres anthropométriques comme l'IMC a été mis en place (Bazick, Donithan et al. 2015). Ce score est doté d'un potentiel élevé de prédiction de la stéatohépatite non-alcoolique (AUC = 0,80). Enfin, les scores FIB-4 et *NAFLD fibrosis score* ont été établis afin de diagnostiquer et suivre la fibrose dans NAFLD (Angulo, Hui et al. 2007, Shah, Lydecker et al. 2009).

Améliorer les outils de diagnostic de la stéatose hépatique non-alcoolique et mieux comprendre les mécanismes qui sont associés à la pathologie est un enjeu majeur. La génétique constitue un outil de choix dans cette démarche.

C- Marqueurs génétiques liés à NAFLD.

Divers travaux au sein de familles ou chez des jumeaux ont mis en avant la composante génétique de la stéatose hépatique non-alcoolique (Schwimmer, Celedon et al. 2009, Loomba, Schork et al. 2015). A l'heure actuelle, 333 gènes sont liés à NAFLD (http://www.disgenet.org/browser/0/1/1/C0400966/). Parmi eux TM6SF2 (*transmembrane 6 superfamily member 2*) (Kozlitina, Smagris et al. 2014), GCKR (*glucokinase regulator*) (Tan, Zain et al. 2014) et PNPLA3 (Romeo, Kozlitina et al. 2008).

C-1- PNPLA3 (*patatin like phospholipase domain containing 3*).

Le gène PNPLA3 est situé sur le bras long du chromosome 22, dans la bande cytogénétique 22q13.31. Il est étalé sur 24 kb environ (https://www.ncbi.nlm.nih.gov/gene/80339) (figure 4) et code une protéine d'environ 53 kDa dite adiponutrine (https://www.uniprot.org/uniprot/Q9NST1). Les niveaux d'expression du gène ont été rapportés comme augmentés au sein du tissu adipeux après la prise alimentaire ou en présence de surcharge pondérale (Baulande, Lasnier et al. 2001).

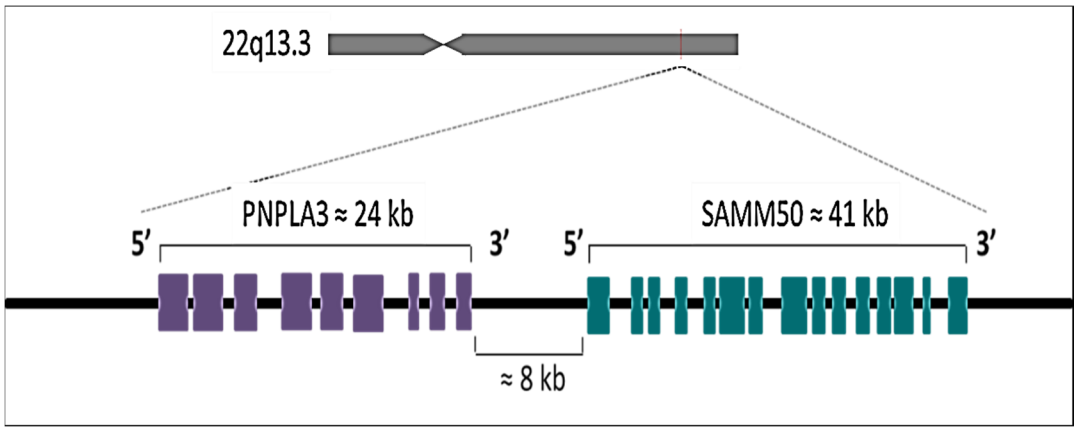

Figure 4. *Représentation schématique du locus PNPLA3-SAMM50*

Une étude effectuée en 2004 a permis de mettre en évidence un rôle de l'adiponutrine dans le processus lipolytique (Jenkins, Mancuso et al. 2004). Des investigations ont également rapporté un lien entre la variation génétique codante (I148M : SNP rs738409) et NAFLD (Romeo, Kozlitina et al. 2008). Ce polymorphisme a été lié à l'état d'avancement de la maladie mais aussi aux taux plasmatiques de diverses enzymes hépatiques comme ALT (Sookoian, Castano et al. 2009, Chambers, Zhang et al. 2011).

La variation génétique rs738409 est à l'origine d'une protéine PNPLA3 ayant une plus faible activité (Huang, Cohen et al. 2011). Cela peut favoriser la surcharge des hépatocytes en substances grasses et participer à l'avènement de la stéatose hépatique (Smagris, BasuRay et al. 2015). Le gène PNPLA3 peut s'avérer très intéressant pour marquer la gravité de la pathologie et a été même lié aux phases les plus délétères comme le carcinome hépatocellulaire (Hassan, Kaseb et al. 2013, Carpino, Pastori et al. 2017).

PNPLA3 a été par ailleurs lié aux niveaux plasmatiques de ALT. En effet, l'étude par GWAS effectuée par Chambers et ses collègues a permis de constater l'association du gène à des taux de ALT plus importants ($p = 1,2 \times 10^{-45}$) (Chambers, Zhang et al. 2011). Cela a été également rapporté par d'autres études au sein de populations de diverses origines ethniques (Hotta, Yoneda et al. 2010, Li, Qu et al. 2012). De manière intéressante, l'étude réalisée par Meffert et ses collègues a mis en avant une mortalité liée aux altérations du foie plus importante chez les individus mâles porteurs de l'allèle pathogène du gène alors que cette variante génétique a été plutôt décrite comme étant liée à un effet de protection

contre la mortalité globale chez les femmes dont le foie n'est pas stéatosé (Meffert, Repp et al. 2018). Dans une autre investigation, la relation de PNPLA3 avec le décès a été attestée chez des sujets ayant une surcharge lipidique au niveau du foie (Mandorfer, Scheiner et al. 2018).

Bien que le SNP rs738409 soit localisé au niveau du gène PNPLA3, des polymorphismes en grand DL avec cette variation génétique sont également présents dans le gène voisin SAMM50 (*SAMM50 sorting and assembly machinery component*). Ce dernier s'étend sur plus de 41 kb et renferme 15 exons (https://www.ncbi.nlm.nih.gov/gene/25813). Il code une protéine de 52 kDa environ (https://www.uniprot.org/uniprot/Q9Y512). Il est donc primordial de prendre en considération l'ensemble du locus PNPLA3-SAMM50 lors des analyses d'association à la stéatose hépatique non-alcoolique.

C-2- Criblage du locus PNPLA3-SAMM50.

La majorité des études génétiques du locus PNPLA3 s'est focalisée sur l'exploration de la variation génétique I148M (rs738409) (Romeo, Kozlitina et al. 2008, Kotronen, Johansson et al. 2009). Cet intérêt vient de l'aspect hautement mutagène du polymorphisme (Pingitore, Pirazzi et al. 2014) qui engendre un changement en termes de l'activité de la protéine après substitution du résidu isoleucine par une méthionine. Toutefois, des polymorphismes autres situés au niveau du locus peuvent s'avérer importants dans le marquage et la physiopathologie de la maladie. L'étude par criblage dense du locus englobant les gènes PNPLA3, SAMM50 et PARVB (*parvin beta*) au sein d'une population du Japon atteinte de NAFLD a permis de constater l'association de polymorphismes répartis sur l'ensemble du locus à diverses composantes de la pathologie, notamment son avènement et évolution vers l'état fibrotique (Kitamoto, Kitamoto et al. 2014). De même, l'étude effectuée par Liu et ses collègues a mis en avant une éventuelle fonctionnalité du SNP rs139051 probablement via modification des niveaux d'expression du gène PNPLA3 (Liu, Anstee et al. 2016). Ce marqueur est situé au niveau de l'intron 2 du gène et se trouve en très faible DL avec rs738409 ($r2 = 0,20$) (Liu, Anstee et al. 2016).

Ces données mettent en avant le rôle important du locus PNPLA3-SAMM50 dans la susceptibilité à la maladie. Se focaliser sur cette région du génome reste cependant

insuffisant. Etablir des modèles de prédiction et de suivi de NAFLD via combinaison de polymorphismes issus de diverses régions du génome peut s'avérer primordial.

C-3- Modèles génétiques de prédiction de NAFLD rapportés dans la littérature.

Divers travaux se sont focalisés sur l'établissement de modèles de prédiction de NAFLD. L'étude récemment réalisée par Di Costanzo et ses collègues au sein d'une population d'Italie a permis la construction d'un modèle constitué de 4 polymorphismes situés dans les gènes PNPLA3, GCKR, TM6SF2 et MBOAT7 (*membrane bound O-acyltransferase domain containing 7*), respectivement et ajusté par l'IMC, la résistance à l'insuline et les taux de triglycérides. Le modèle a montré un excellent potentiel (OR = 4,97 pour le tertile le plus élevé) en termes d'anticipation de l'avènement de NAFLD (Di Costanzo, Belardinilli et al. 2018).

Dans une autre investigation, des chercheurs ont construit, au sein d'une population du Japon, un score de 4 SNPs situés dans les gènes PNPLA3, GCKR, GATAD2A (*GATA zinc finger domain containing 2A*) et DYSF (*dysferlin*), respectivement, et ajusté par le sexe (Kawaguchi, Shima et al. 2018). Ce modèle a montré une bonne capacité de prédiction de NAFLD (AUC = 0,65).

Par ailleurs, des SNPs situés dans le locus PNPLA3-SAMM50 ainsi que dans les gènes ADIPOQ (*Adiponectin*) et COL13A1 (*collagen type XIII alpha 1 chain*) ont été exploités pour la construction d'un modèle génétique prédisant les taux plasmatiques de ALT dans la population mexicaine (Larrieta-Carrasco, Flores et al. 2018). Le score génétique constitué de 9 à 12 allèles de risque a permis d'englober 45% environ des individus ayant des taux élevés de ALT.

Bien que les résultats concernant les 3 scores génétiques ci-dessus mentionnés aient été intéressants, ils doivent être pris avec précaution. En effet, ces modèles ont été construits au sein de populations ayant servi à la sélection des SNPs qui les constituent ce qui suggère une surestimation de leurs potentiels de prédiction.

L'étude effectuée par Vespasiani-Gentilucci et ses collègues dans une population d'Italie a permis la mise en place d'un modèle anticipant l'apparition de NAFLD et ses diverses

complications (Vespasiani-Gentilucci, Dell'Unto et al. 2018). Le score polygénique établi dans cette investigation, constitué par les SNPs rs738409, rs58542926 et rs3750861 situés dans les gènes PNPLA3, TM6SF2 et KLF6 (*Kruppel like factor 6*), respectivement, est ajusté par l'âge, l'IMC, le sexe, et le diabète. Ce score a été obtenu par régression logistique après pondération par les ampleurs d'effets des polymorphismes (Vespasiani-Gentilucci, Dell'Unto et al. 2018). Le modèle a été capable de prédire NAFLD (OR = 23) mais aussi la cirrhose hépatique chez des patients avec NASH (OR = 88).

Malgré ces avancées, des améliorations sont encore nécessaires. La disponibilité de bases de données pangénomiques comme *GWAS Catalog* (MacArthur, Bowler et al. 2017) ou *UK-BioBank* (Bycroft, Freeman et al. 2018) et leur accessibilité à l'ensemble de la communauté scientifique constitue une opportunité de taille. Il devient ainsi plus facile de cribler des millions de marqueurs génétiques via diverses stratégies comme par apprentissage machine (Vilhjalmsson, Yang et al. 2015, Mak, Porsch et al. 2017) afin de sélectionner un maximum de variations génétiques liées à NAFLD et décrire au mieux la pathologie.

Objectifs et hypothèse de travail

NAFLD est une altération associée à diverses anomalies cardiométaboliques. Cette relation serait due comme précédemment mentionné à des interactions entre divers processus physiopathologiques ayant comme composante commune la résistance à l'insuline et la toxicité induite par la dérégulation des métabolismes glucidique et lipidique.

Les fonds génétiques de ces altérations métaboliques sont complexes et sont déterminés par des centaines voire des milliers de gènes de susceptibilité. L'hypothèse est que les composantes génétiques de NAFLD et de la mortalité dans le DT2 sont déterminées par des gènes liés à ces anomalies cardiométaboliques. Ainsi, le marquage, la prédiction et le suivi de la stéatose hépatique non-alcoolique et de la mortalité associée nécessitent la combinaison de centaines voire de milliers de variations génétiques associées à ces anomalies cardiométaboliques. Cela permettrait d'englober le maximum des processus causaux de NAFLD.

Dans ce contexte, mon projet de maitrise a pour objectif l'exploration de la composante génétique de NAFLD et de la mortalité dans le DT2. Le but principal est d'établir des outils de marquage et de prédiction des patients diabétiques à risque. En absence de données phénotypiques en termes de diagnostic de NAFLD par histopathologie ou par imagerie, les taux plasmatiques de l'enzyme ALT sont utilisés comme marqueur de la stéatose hépatique non-alcoolique. Bien que peu corrélés au diagnostic de NAFLD (Browning, Szczepaniak et al. 2004), les taux de cette enzyme sont intéressants dans l'exploration des diverses composantes cardiométaboliques liées à la stéatose hépatique non-alcoolique (Martin-Rodriguez, Gonzalez-Cantero et al. 2017). L'étude du fond génétique associé aux variations des niveaux de ALT peut permettre donc une meilleure compréhension des complications du DT2 d'origine hépatique.

Plus spécifiquement, deux approches sont appliquées dans le projet:

1) Une approche non-ciblée basée sur l'exploitation de la technologie LASSO pour la sélection de SNPs afin de construire un modèle de prédiction de la pathologie (figure 5).

2) Une approche ciblée dans laquelle des régions du génome connues pour leur relation avec NAFLD et ses diverses composantes sont explorées. Deux volets du travail sont effectués ici. Le premier volet aura pour but la combinaison de diverses variations génétiques rapportées dans la littérature comme associés à NAFLD ou à ses composantes afin de construire un modèle prédisant la maladie et stratifier la population en fonction du risque de complication de NAFLD et de la mortalité. Le second consiste en le criblage du locus englobant les gènes PNPLA3 et SAMM50 afin de rechercher une diversité éventuelle en termes de marquage de la maladie et de l'effet des thérapies associées (figure 5).

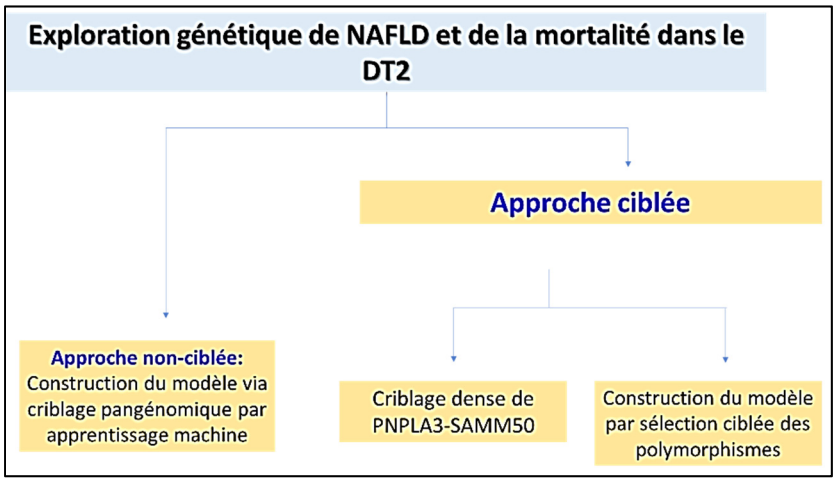

Figure 5. *Représentation schématique de la stratégie établie pour la mise en place d'outils de marquage génétique de NAFLD et de la mortalité dans le DT2.*

A noter qu'une analyse préliminaire est réalisée dans un premier temps. Celle-ci consiste en le croisement de listes de gènes candidats rapportés dans la littérature (http://www.disgenet.org/search) afin d'estimer la composante génétique commune entre NAFLD et différentes altérations métaboliques retrouvées dans le DT2.

Matériel et méthodes.

I- Base de données exploitée lors de l'analyse préliminaire :

La base de données *Disgenet* (Pinero, Bravo et al. 2017) (http://www.disgenet.org/search) a été exploitée pour l'extraction des listes des gènes candidats de NAFLD et des différentes altérations métaboliques qui lui sont associées. Sous sa version actuelle, *Disgenet* regroupe des données centrées sur les gènes obtenues après curation à partir de bases de données génétiques et pharmacogénétiques humaines ou murines. Il s'agit de: *Comparative Toxicogenomics Database (CTD)* (http://ctdbase.org/), *UniProt* (https://www.uniprot.org/), *ClinVar* (https://www.ncbi.nlm.nih.gov/clinvar/), *Orphanet* (https://www.orpha.net/consor/cgi-bin/index.php), *GWAS catalog* (https://www.ebi.ac.uk/gwas/), *Rat Genome Database* (https://rgd.mcw.edu/), *Mouse Genome Database* (http://www.informatics.jax.org/), *Genetic Association Database* (https://geneticassociationdb.nih.gov/), *Literature Human Gene Derived Network* (http://www.dbs.ifi.lmu.de/~bundschu/LHGDN.html), *PsyGeNET* (http://www.psygenet.org/), *Human Phenotype Ontology* (https://hpo.jax.org/app/), CGI (https://www.cancergenetics.com/), *ClinGen* (https://clinicalgenome.org/), *Genomics England panel app* (https://panelapp.genomicsengland.co.uk/) et GWAS db (http://jjwanglab.org/gwasdb).

II- Population étudiée.

La population étudiée au cours de la maîtrise a été recrutée dans le cadre de l'essai clinique ADVANCE (*The Action in Diabetes and Vascular Disease: Preterax and Diamicron Modified Release Controlled Evaluation*). Il s'agit d'une étude réalisée dans le cadre d'une collaboration internationale dans laquelle divers groupes de recherche à travers quatre continents ont pris part (Patel, MacMahon et al. 2008). Cela a permis le recrutement de 11140 individus atteints de DT2 afin d'y tester, pendant 5 années, l'effet de la combinaison de thérapies intensives hypoglycémiantes (gliclazide à libération prolongée) et normotensives (indapamide et perindopril) sur l'amélioration des profils microvasculaire, marcrovasculaire et de régulation glucidique des patients (Patel, MacMahon et al. 2008). A noter qu'un suivi de 5 années supplémentaires a été réalisé dans le cadre de l'essai clinique ADVANCE-ON (Zoungas, Chalmers et al. 2014).

A- Protocole de recrutement et de suivi des patients.

Les patients recrutés dans l'investigation sont âgés de 55 ans et plus (Patel, MacMahon et al. 2008). Ces derniers ont été répartis, suivant un schéma factoriel et randomisé, en 4 groupes : 1) Un premier comportant 2783 patients bénéficiant des thérapies hypoglycémiante intensive et hypotensive, 2) un second groupe de 2788 individus soumis à la thérapie intensive hypoglycémiante et à un placébo, 3) un autre groupe constitué de 2786 patients bénéficiant d'un traitement hypoglycémiant standard et d'une thérapie normotensive et 4) un dernier groupe comportant 2783 individus soumis à la thérapie hypoglycémiante standard et à un placébo (ADVANCE-Group 2010) (figure 6).

Les patients ont été suivis de manière périodique durant 5 ans afin de récupérer les données cliniques. Ces dernières ont concerné, entre autres, le diagnostic d'évènements microvasculaires, macrovasculaires, la mortalité totale et celle d'origine cardiovasculaire. De même, des mesures de paramètres biologiques comme les taux HbA1c, la glycémie et la triglycéridémie ont été réalisées (Patel, MacMahon et al. 2007, Patel, MacMahon et al. 2008, ADVANCE-Group 2010).

Cet essai a mis en avant une amélioration du profil glucidique des patients diabétiques traduit par une diminution des taux de HbA1c jusqu'à 6,5% (Patel, MacMahon et al. 2008). Par ailleurs, l'étude a permis d'observer une régression des anomalies microvasculaires notamment celles affectant le rein après thérapie intensive (Patel, MacMahon et al. 2008).

B- Données génétiques dans ADVANCE.

L'essai clinique ADVANCE a été l'occasion de réaliser un volet génétique du projet supervisé par les Prs. Pavel Hamet et Johanne Tremblay. Plus de 4000 individus ayant une origine caucasienne ont ainsi été génotypés sur des puces. Des millions de génotypes ont été obtenus après imputation. Diverses investigations génétiques de grande qualité ont été effectuées dans le cadre d'explorations individuelles mais aussi dans le cadre de consortiums (Gorski, Tin et al. 2015, Hamet, Haloui et al. 2017, Wuttke, Li et al. 2019). Ces données génétiques sont regroupées au sein de la base de données OPTITHERA au Centre de recherche de l'Hôpital de Montréal (Responsables : Prs. Pavel Hamet et Johanne Tremblay).

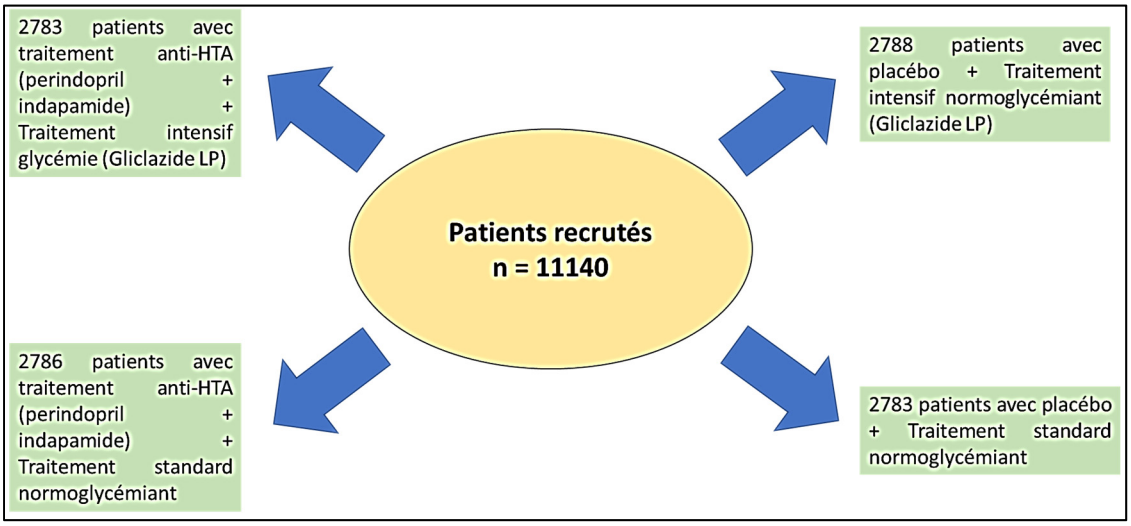

Figure 6. *Représentation schématique des patients de la population ADVANCE répartis, suivant un schéma factoriel randomisé, en 4 groupes de traitement (thérapie hypoglycémiante et thérapie anti-hypertension artérielle). Schéma adapté et modifié à partir de la figure 11 de (ADVANCE-Group 2010) et à partir de données issues de (Zoungas, de Galan et al. 2009).*

Le nombre de patients mis à contribution dans l'investigation est de 4098 patients. Il s'agit d'individus d'origine caucasienne dont les données de génotypage sont disponibles. Dans notre laboratoire, une partie de la population (2228 échantillons) a été génotypée via la puce *Affymetrix 6.0*, une seconde partie (1092 échantillons) via la puce *UKBiobank* et le reste des patients (778 patients) a été génotypé par la puce *Affymetrix 5.0* (Hamet, Haloui et al. 2017).

Comme précédemment mentionné, en absence de données de diagnostic de NAFLD par histopathologie ou par imagerie, les taux plasmatique de ALT seront exploités comme substitut.

Parmi les 4098 patients, 4051 ont les taux plasmatiques de ALT disponibles. Dix millions de SNPs environ ont des génotypes disponibles, dont plus de 99,9% sont en équilibre de Hardy-Weinberg (seuil = $5,7 \times 10^{-7}$). Par ailleurs, 1008 (25%) parmi les 4098 patients ont

bénéficié aussi bien de la thérapie hypoglycémiante intensive que du traitement antihypertensif, 1026 (25%) ont été soumis uniquement au traitement hypoglycémiant intensif et à un placébo, 1042 (25%) ont bénéficié de la thérapie antihypertensive et d'un traitement hypoglycémiant standard et 1022 (25%) ont été traités par thérapie standard hypoglycémiante et par placébo.

La population a été stratifiée en 3 groupes d'individus en fonction de leurs taux plasmatiques de ALT en ligne de base. Un premier groupe de 3138 patients ayant des taux ≤ 35 UI/L (relativement bas), un second groupe constitué de 470 malades dont les taux varient entre 35 et 45 UI/L (moyens) et un dernier groupe de 443 patients ayant des niveaux plasmatiques de ALT > 45 UI/L (élevés).

Trente-huit phénotypes ont été pris en considération. Ces derniers décrivent les aspects cardiovasculaires, rénaux et d'homéostasie glucidique : 21 phénotypes quantitatifs et 17 phénotypes qualitatifs ont été explorés (figure 7).

III- Outils statistiques et bioinformatiques exploités.

Deux langages de programmation ont été mis à contribution dans les analyses : les langages Python et R. Python a été utilisé pour l'organisation des données et leur configuration afin de les rendre exploitables dans les calculs.

A- Outils exploités lors de l'analyse préliminaire.

Comme déjà mentionné, l'analyse a consisté en le croisement de listes de gènes candidats issues de la base de données *Disgenet* (http://www.disgenet.org). J'ai croisé la liste des gènes candidats de chacune des altérations cardiométaboliques d'intérêt avec la liste des gènes de NAFLD. Afin d'attester de la validité des listes des gènes communs lors de chaque croisement, le test de Fisher a été exploité en utilisant la fonction « fisher.test » via langage R. Nommons le nombre de gènes communs (n_1). J'ai créé en parallèle une liste de gènes de même taille que celle des gènes de l'altération cardiométabolique étudiée. Cette liste est obtenue en prenant au hasard des gènes parmi ceux répertoriés dans *Disgenet* (http://www.disgenet.org/). Elle est ensuite croisée avec la liste des gènes candidats de NAFLD. Pour chaque croisement, nommons le nombre de gènes communs (n_0). L'hypothèse nulle (H_0) stipule que la liste de gènes communs lors de chaque croisement

est due au hasard alors que l'hypothèse alternative (H_1) (P du test de Fisher < 0,05) atteste que cette liste de gènes communs n'est pas due au hasard.

B- Outils bioinformatiques et statistiques exploités lors de l'investigation au sein de la population ADVANCE.

Si ce n'est pas indiqué, les variables d'ajustement prises en considération dans les analyses sont l'âge, le sexe et l'abus de prise d'alcool par les patients.

Les régressions statistiques linéaire et logistique ont été exploitées dans divers calculs. L'ajustement des moyennes des phénotypes quantitatifs et des fréquences des phénotypes qualitatifs, par l'âge, le sexe et l'excès de prise d'alcool a été effectué en utilisant le programme lsmeans (https://cran.r-project.org/web/packages/lsmeans/index.html. Cela est réalisé suivant la ligne de commande :

Phenotype ~ groupe_patients + age + sexe + abus_d'alcool

où groupe_patients correspond aux catégories des patients en fonction de leurs taux plasmatiques de ALT ou de leur génotype. Les niveaux de signification des associations sont indiqués en P de contraste.

Par ailleurs, le logiciel Haploview a été exploité afin d'établir le profil de DL du locus PNPLA3-SAMM50 au sein de la population ADVANCE et taguer les SNPs qui s'y trouvent. L'extraction des SNPs tags a été réalisée via la méthode *pairwise-tagging* non-agressive (de Bakker, Burtt et al. 2006). Cette méthode permet la sélection des polymorphismes en fonction des niveaux de DL (r^2) entre les paires de SNPs en absence de toute spécification des SNPs tags désirés. Un seuil r^2 est établi à cet effet.

La puissance statistique des résultats les plus pertinents sera calculée via le module pwr (https://cran.r-project.org/web/packages/pwr/index.html). La puissance du test (comparaison entre deux moyennes ou fréquences) est établie par la ligne de commande :

pwr.t2n.test(n1, n2, d)

où **n1** et **n2** représentent le nombre de patients dans les deux groupes comparés et **d** est le coefficient de régression de Cohen (Cohen 1988). Ce dernier a été calculé à partir des moyennes ajustées et de l'erreur standard des phénotypes.

Par ailleurs, la correction des analyses multivariées par le test de Bonferroni a été effectuée en utilisant le module podkat (https://bioconductor.org/packages/release/bioc/html/podkat.html). Les résultats ayant été corrigés par ce test sont indiqués dans le texte.

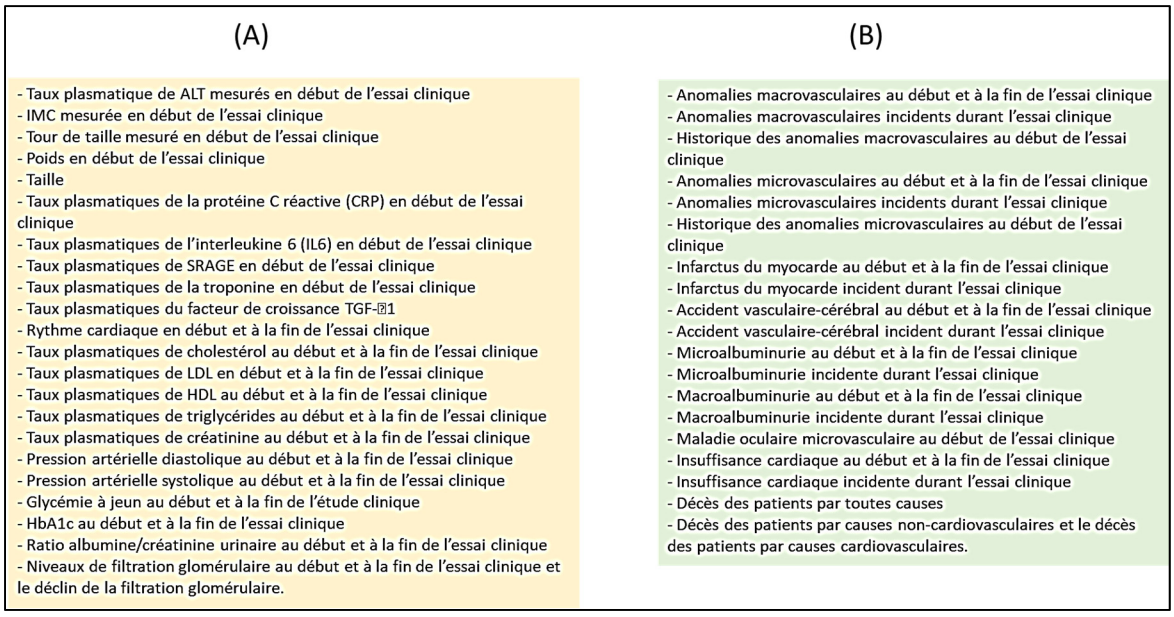

Figure 7. *Phénotypes A) quantitatifs et B) qualitatifs explorés dans l'étude*

IV- Sélection des SNPs pour la construction des PRS.

Cette sélection a été réalisée via deux stratégies : par apprentissage machine (programme LASSOSUM) ou par sélection ciblée.

- **Construction du PRS par LASSOSUM.**

Les données d'environ 9 millions de SNPs issus du génotypage de la population ADVANCE dans la puce *Affymetrix 6.0* et de plus de 6 millions de SNPs provenant du génotypage de la population dans la puce *UKB-chip* ont été exploitées sous forme de fichiers .fam, .bim, .bed. La population de référence utilisée, est celle génotypée dans la puce *Affymetrix UK BioBank*. Elle est constituée de 1092 patients (1079 individus ont les taux de ALT disponibles). La population test, génotypée dans la puce *Affymetrix 6.0*,

comporte 2228 individus (2220 patients ont les données de ALT renseignés). Les fichiers des données statistiques sommaires utilisés, ont été extraits de la base de données *GWAS catalog* (https://www.ebi.ac.uk/gwas/downloads/summary-statistics) alors que les fichiers rapportant les profils de DL sont ceux de la population *1000 Genomes d'origine caucasienne* (Berisa and Pickrell 2016) (figure 8).

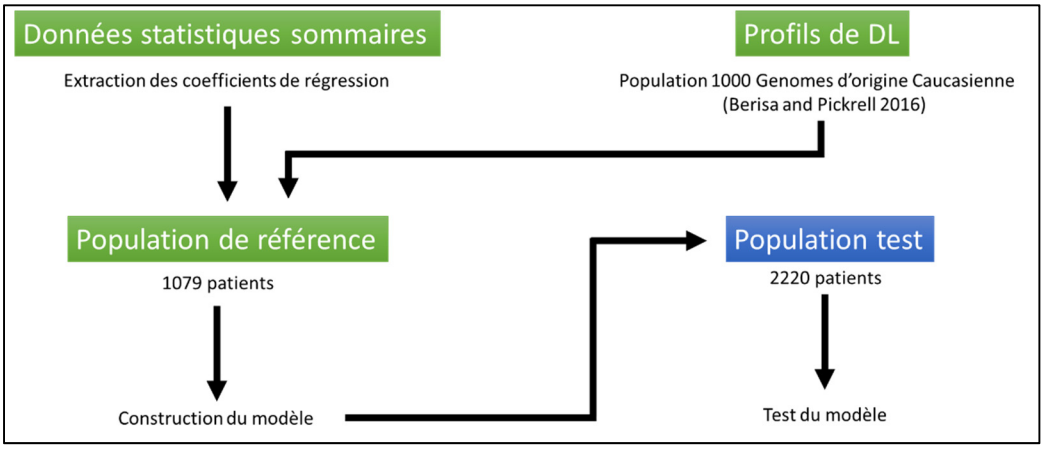

Figure 8 : *Représentation schématique du processus de construction d'un PRS par LASSOSUM*

- **Construction des PRS via sélection ciblée des SNPs.**

Cette stratégie est basée sur l'extraction, à partir des fichiers statistiques sommaires, de SNPs ayant été fortement associés dans la littérature à des phénotypes en relation avec NAFLD. Les données statistiques sommaires issues des bases de données *GWAS catalog* (https://www.ebi.ac.uk/gwas/summary-statistics), *GeneATLAS* (http://geneatlas.roslin.ed.ac.uk/), *Giant Consortium DataBase* (https://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files#WHRadjBMI_.28download_GZIP.29), *Global Lipids Genetics Consortium* Database (http://csg.sph.umich.edu/willer/public/lipids2013/) et *Magic Consortium* DataBase (https://www.magicinvestigators.org/downloads/) ont été exploitées à cet effet.

Le logiciel *PLINK* (http://zzz.bwh.harvard.edu/plink/) a été mis à contribution pour la gestion des données génotypiques et pour la sélection des SNPs représentatifs en fonction du niveau de DL (index r^2) au sein de la population ADVANCE. Cette sélection a été également effectuée en utilisant la base de données *snipa* (Arnold, Raffler et al. 2015) (https://snipa.helmholtz-muenchen.de/snipa3/index.php?task=pairwise_ld).

La construction du PRS est réalisée en calculant la somme des modèles additifs pondérés par les coefficients de régression (bêtas) issus de la littérature. Le PRS est ensuite testé via GLM afin de déterminer son potentiel de prédiction. Ce dernier est estimé par AUC en utilisant le module pROC (https://cran.r-project.org/web/packages/pROC/index.html), suivant les lignes de commande :

Formule <- as.formula(Phénotype ~ variable$_1$ + variable$_2$ + … + variable$_n$)

modele <- glm(Formule, data=matrice_données, family="binomial")

Ce module a été également exploité dans la comparaison des potentiels des modèles de prédiction via la fonction « roc.test ».

Enfin, le potentiel de stratification par le modèle est établi via le calcul des quintiles des scores de risque et la détermination de la fréquence du phénotype d'intérêt dans chaque strate de la population.

Résultats

I- Résultats obtenus lors de l'analyse préliminaire.

L'exploration de la base de données *Disgenet* (http://www.disgenet.org/) a permis de mettre en évidence 333 gènes candidats de NAFLD, 1671 gènes de DT2 et 11 gènes liés aux taux plasmatiques de ALT (tableau 1).

Phénotype	Nombre de gènes candidats
Diabète non-insulino dépendant	1671
Infarctus du myocarde	965
Insuffisance cardiaque	814
Néphropathie diabétique	560
L'accident vasculaire cérébral ischémique	393
NAFLD	333
Hypertension essentielle	277
Stéatohépatite non-alcoolique	208
Hypertriglycéridémie	156
Résistance à l'insuline	80
Obésité viscérale	44
Insuffisance rénale chronique	28
Taux de ALT sérique	11
Taux plasmatiques élevés de HbA1c	2

Tableau 1. *Nombre des gènes associés à NAFLD ou aux différentes pathologies et phénotypes en relation avec cette maladie (données extraites à partir de la base de données Disgenet http://www.disgenet.org/).*

L'analyse a permis de constater une fraction importante (58%) des gènes candidats de NAFLD communs avec le DT2 (P test_fisher = 9,5 x 10^{-37}, $n_1 / n_0 = 4,8$) de même que 11,5% des gènes du DT2 sont communs à NAFLD (P test_fisher = 1,3 x 10^{-26}) (figure 9). Cela démontre la relation entre les deux phénotypes et peut expliquer la fréquence élevée du foie stéatosé non-alcoolique chez les sujets diabétiques de type 2 rapportée dans la littérature (Younossi, Koenig et al. 2016, Amiri Dash Atan, Koushki et al. 2017, Cusi, Sanyal et al. 2017).

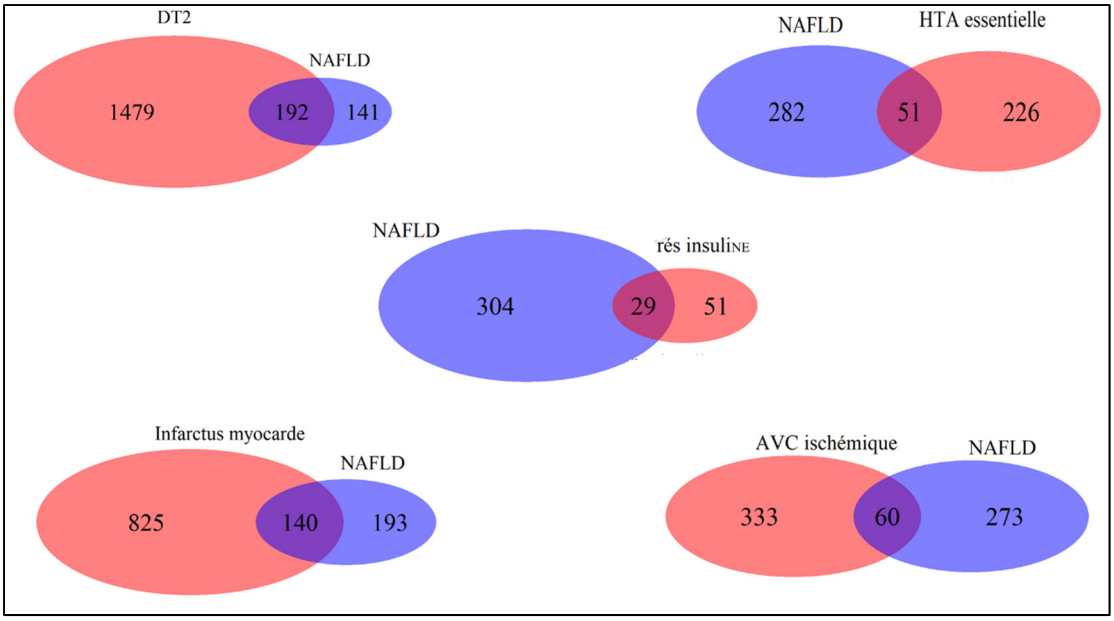

Figure 9 : *Croisement des listes des gènes candidats de NAFLD et de quelques altérations cardiométaboliques qui lui sont associées. Les listes de gènes ont été extraites de la base de données Disgenet (http://www.disgenet.org/search)*

Par ailleurs, plus de 36% des gènes de la résistance à l'insuline sont communs avec la stéatose hépatique non-alcoolique (P test_fisher = 3,0 x 10^{-8}, f_1 / f_0 = 14,5) et un tiers environ des gènes associés à l'hypertriglycéridémie sont des gènes candidats du NAFLD (P test_fisher = 7,4 x 10^{-12}, f_1 / f_0 = 11,75). De manière intéressante, 18% uniquement des gènes liés aux variations des taux plasmatiques de l'enzyme ALT (2 gènes sur 9, parmi lesquels PNPLA3) sont communs avec NAFLD. Ceci confirme les données de la littérature attestant de la capacité limitée de ALT à marquer le phénotype stéatose hépatique (Browning, Szczepaniak et al. 2004). L'analyse a également mis en avant la présence de 18% environ des gènes de l'hypertension essentielle (P test_fisher = 7,7 x 10^{-11}, n_1 / n_0 = 8,5), de 14% des gènes de l'infarctus du myocarde (P test_fisher = 3,0 x 10^{-29}, n_1 / n_0 = 10,0), de 15% des gènes de l'AVC ischémique (P test_fisher = 6,2 x 10^{-14}, n_1 / n_0 = 12,0) et de 13% des gènes de l'insuffisance cardiaque (P test_fisher = 1,3 x 10^{-19}, n_1 / n_0 = 7,5) dans NAFLD (figure 9).

Ces résultats suggèrent encore une fois un fond génétique partagé entre NAFLD, le DT2 et les altérations cardiométaboliques.

II- Résultats obtenus lors de l'investigation au sein de la population ADVANCE.

L'analyse des données phénotypiques au sein des groupes de patients ayant bénéficié des thérapies hypoglycémiante standard et un placébo (groupe 1), des thérapie intensive hypoglycémiante et un placébo (groupe 2), de traitement hypoglycémiant standard et une thérapie hypotensive (groupe 3) ou de traitements hypoglycémiant intensif et hypotensif (groupe 4) a permis de constater des taux de ALT en ligne de base plus élevés chez les patients du groupe 4 (29,16 ± 0,59 UI/L, P = 0,04) et du groupe 3 (29,25 ± 0,58 UI/L, P = 0,03) en comparaison avec ceux du groupe 1 (27,60 ± 0,58 UI/L). Les complications microvasculaires ont été également plus présentes en ligne de base chez les diabétiques du groupe 2 (0,48 ± 0,02) par rapport à celles du groupe 3 (0,44 ± 0,02; P = 0,04). Le reste des données phénotypiques n'a pas été significativement différent entre les 4 groupes de patients (Tableau 2A).

Comme déjà mentionné, la population ADVANCE (4051 patients ayant les taux de ALT disponibles) a été répartie en trois groupes d'individus : des patients ayant des taux de ALT élevés (ALT > 45 UI/L) → groupe 1, ceux avec des niveaux médians (35 UI/L < ALT ≤ 45 UI/L) → groupe 2 et des patients avec des taux plasmatiques bas (ALT ≤ 35 UI/L) → groupe 3. Cela a permis de constater un profil métabolique plus défavorable accompagnant l'élévation des taux de l'enzyme notamment en termes du tour de taille des patients ($P = 2,3 \times 10^{-12}$), de leurs taux plasmatiques de HDL ($P = 5,4 \times 10^{-6}$), des triglycérides (P = 0,04) ainsi que des taux de HbA1c ($P = 3,7 \times 10^{-6}$) (tableau 2B).

L'analyse n'a toutefois pas mis en avant de différence significative des taux de mortalité entre les groupes d'individus, qu'il s'agisse de mortalité globale (P = 0,82), d'origine cardiovasculaire (P = 0,71) ou d'origine non-cardiovasculaire (P = 0,48) (tableau 2B). De même, la régression logistique a démontré l'absence de corrélation entre les mortalités non-cardiovasculaire ou cardiovasculaires et les taux de ALT (P = 0,31 et P = 0,54 ; respectivement).

A)

Phénotype	Groupe 1 n = 1013	SE	Groupe 2 n = 1013	SE	Groupe 3 n = 1029	SE	Groupe 4 n = 996	SE	1 vs 3	1 vs 4
ALT (UI/L)	27,60	0,58	28,21	0,59	29,25	0,58	29,16	0,59	**0,03**	**0,04**
HbA1c (%)	7,27	0,05	7,26	0,05	7,33	0,05	7,31	0,05	0,32	0,55
Cholestérole (mmol/L)	5,20	0,04	5,17	0,04	5,22	0,04	5,19	0,04	0,56	0,91
Nombre de médicaments pris contre le diabète de type 2	1,30	0,03	1,38	0,03	1,35	0,03	1,36	0,03	0,25	0,14
LDL (mmol/L)	3,10	0,03	3,05	0,03	3,10	0,03	3,06	0,03	0,98	0,39
HDL (mmol/L)	1,28	0,01	1,28	0,01	1,27	0,01	1,28	0,01	0,33	0,80
Triglycérides (mmol/L)	1,97	0,04	1,95	0,04	2,00	0,04	1,93	0,04	0,62	0,41
Rythme cardiaque (battements/min)	72,40	0,41	72,84	0,42	72,64	0,41	72,99	0,42	0,65	0,27
Pression artérielle systolique (mmHg)	147,13	0,71	147,20	0,71	146,79	0,70	147,47	0,71	0,72	0,71
Tour de taille (cm)	103,15	0,41	103,52	0,41	103,76	0,41	103,63	0,42	0,26	0,38
Fréquence de décès par toutes causes	0,15	0,01	0,13	0,01	0,11	0,01	0,11	0,01	**0,01**	**0,03**
Fréquence de décès par causes cardiovasculaires	0,09	0,01	0,07	0,01	0,05	0,01	0,05	0,01	**5,02E-04**	**0,002**
Fréquence de décès par causes non-cardiovasculaires	0,06	0,01	0,07	0,01	0,06	0,01	0,06	0,01	0,89	0,84

B)

Phenotype quantitatif	Groupe1 ALT <= 35 UI/L n = 3138	SE	Groupe2 35 < ALT <= 45 UI/L n = 470	SE	Groupe3 ALT > 45 UI/L n = 443	SE	P_value(contrast) Grp2 vs Grp 1	P_value (contrast) Grp 3 vs Grp 1
ALT (IU/L)	22,07	0,24	39,49	0,52	64,22	0,53	NA	NA
HbA1c (%)	7,24	0,03	7,43	0,07	7,56	0,07	**0,004**	**3,68E-06**
Cholestérole (mmol/L)	5,21	0,02	5,15	0,05	5,11	0,05	0,24	0,06
Nombre de médicaments pris contre le DT2	1,34	0,02	1,35	0,04	1,41	0,04	0,72	0,08
LDL (mmol/L)	3,10	0,02	3,00	0,05	3,02	0,05	0,06	0,15
HDL (mmol/L)	1,29	0,01	1,25	0,02	1,21	0,02	**0,03**	**5,35E-06**
Triglycérides (mmol/L)	1,93	0,03	2,11	0,06	2,06	0,06	**0,003**	**0,04**
Rythme cardiaque (battements/min)	72,52	0,27	73,11	0,58	73,70	0,61	0,33	0,06
Pression artérielle systolique (mmHg)	146,82	0,46	148,57	1,00	147,97	1,04	0,09	0,28
Tour de taille (cm)	102,74	0,27	105,44	0,58	107,08	0,60	**6,22E-06**	**2,32E-12**
Fréquence de décès par toutes causes	0,13	0,01	0,12	0,02	0,13	0,02	0,93	0,82
Fréquence de décès par causes CV	0,06	0,01	0,08	0,01	0,06	0,01	0,29	0,71
Fréquence de décès par causes non CV	0,07	0,01	0,05	0,01	0,07	0,01	0,26	0,48

Tableau 2. *Données phénotypiques de la population ADVANCE ventilées par A) les 4 groupes de traitement et par B) les 3 niveaux des taux plasmatiques de ALT. L'intitulé des groupes de traitement est indiqué dans le texte ci-dessus. Les patients pour lesquels le taux de ALT ne sont pas disponibles ne sont pas inclus dans les tableaux.*

Ce résultat met en avant la difficulté d'attester la relation entre les variations des taux de ALT et les niveaux de décès déjà rapportée dans la littérature (Liu, Ning et al. 2014). Dans ce contexte, l'analyse génétique peut s'avérer intéressante afin de mieux comprendre

certains mécanismes moléculaires liant la mortalité et les variations des taux de l'enzyme. L'analyse génétique a été abordée via deux approches : une ciblée et une autre non-ciblée.

A- Exploration génétique via une approche non-ciblée

Cette exploration tend à concevoir des modèles de prédiction. Afin d'être le plus exhaustif possible, j'ai sélectionné des SNPs en absence de toute hypothèse physiopathologique via la méthode statistique LASSO (Tibshirani 1996).

Conception par la méthode statistique LASSO d'un modèle de prédiction de NAFLD.

Le programme utilisé dans cette investigation est LASSOSUM. Celui-ci est adapté, comme déjà mentionné, à la construction des modèles de prédiction de risque via l'exploitation des métadonnées génétiques (Mak, Porsch et al. 2017).

Quatre phénotypes ont été pris en considération dans la conception du modèle vu leur relation avec NAFLD :

- Le tour de taille : les données statistiques sommaires utilisées ici sont issues de la base de données UKBB (https://www.ukbiobank.ac.uk/). Le fichier statistique sommaire hébergé dans la base de données *GeneATLAS* (http://geneatlas.roslin.ed.ac.uk/downloads/?traits=750) comporte 9 millions de SNPs environ. Ces données de génotypage sont issues de 549598 individus d'origine caucasienne recrutés au Royaume-Uni.
- Les taux plasmatiques des triglycérides : le fichier des données statistiques sommaires utilisé pour ce phénotype est issu de l'investigation réalisée par Surakka et ses collègues (Surakka et al, 2015). Cette étude a focalisé sur 62166 individus d'origine européenne. Plus de 9 millions de SNPs sont rapportés dans le fichier.
- Les taux plasmatiques de ALT et de AST : les données statistiques sommaires proviennent de l'investigation de Prins et ses collègues (Prins, Kuchenbaecker et

al. 2017). Cette étude a concerné 9961 individus du Royaume-Uni. Les données sommaires comportent plus de 25 millions de SNPs.

A noter que la population de référence exploitée dans l'analyse est constituée de 1079 patients (patients ayant les taux de ALT mesurés parmi les 1092 individus de ADVANCE génotypés dans la puce *UKB-chip*). La population test, quant à elle, renferme 2220 patients ayant les taux de ALT rapportés parmi les 2228 individus de ADVANCE génotypés dans la puce *Affymetrix 6.0*.

Des analyses préliminaires ont permis de détecter un défaut dans le programme LASSOSUM dans le sens où les données phénotypiques manquantes, indiquées par défaut en chiffre « -9 » dans les tableaux « .bim » de Plink génèrent des résultats discordants. Le concepteur du programme LASSOSUM, Pr. Timothy Mak, a attesté cette incompatibilité. L'analyse a été reprise en indiquant en chiffre « 0 » les données phénotypiques manquantes.

Cette analyse a permis la construction de 4 PRS capables de prédire les patients ayant des taux plasmatiques élevés de ALT (ALT > 45 UI/L). Il s'agit de :

- Deux SNPs sélectionnés (rs3747207 dans le gène PNPLA3 et rs1352738 situé dans une région intergénique sur le chromosome 5) après exploitation des données statistiques sommaires (étude de Prins et ses collègues) liées aux niveaux de ALT (Prins et al, 2017) → **modèle de prédiction génétique 1.**
- Plus de 152000 SNPs extraits après exploitation des données statistiques sommaires de l'investigation de Prins et collègues en rapport avec la variation des niveaux de AST dans la population du Royaume-Uni (Prins et al, 2017) → **modèle de prédiction génétique 2.**
- Plus de 43000 SNPs sélectionnés à partir des données statistiques sommaires du phénotype « tour de taille » issues la base de données *UK-BioBank*. (http://geneatlas.roslin.ed.ac.uk/downloads/?traits=750) → **modèle de prédiction génétique 3.**
- Plus de 1,3 x 10^6 de SNPs en relation avec la triglycéridémie, extraits des données statistiques sommaires de l'étude de Surakka et ses collègues (Surakka et al, 2015) → **modèle de prédiction génétique 4.**

L'analyse a permis de constater que le modèle 1 a un potentiel de prédiction intéressant (AUC = 0,589, P = 1,4 x 10^{-7}) tandis que la capacité de prédiction par le modèle 2 est de : (AUC = 0,551, P = 0,016). Le modèle génétique 3 a un AUC de 0,555 (P = 0,017) alors que AUC du modèle 4 est = 0,552 (P = 0,021). La combinaison par l'addition des quatre modèles de risque a permis la prédiction des taux élevés de ALT avec un AUC de 0,622 (P significatifs pour chacun des 4 modèles combinés) (tableau 3).

Afin de tester la puissance de la méthode LASSOSUM dans la construction de modèles dans des populations de tailles relativement faibles comme ADVANCE, j'ai effectué une analyse de vérification. J'ai utilisé comme population test le groupe de patients que j'ai précédemment considéré comme référence (2220 patients) de même que j'ai pris comme population de référence les patients que j'ai précédemment utilisés pour tester le modèle (1079 patients). L'analyse des données statistiques sommaires liées à ALT (Prins, Kuchenbaecker et al. 2017) a permis de construire un PRS de 3071 SNPs prédisant les taux élevés de ALT avec un AUC = 0,55 [0,498 – 0,604], P = 0,056). Il apparait ainsi que le modèle obtenu après inversion des populations référence et test est différent en termes de nombre de SNPs et de puissance de prédiction. Cela pourrait être dû à une taille insuffisante des populations utilisées dans l'entrainement des modèles. En effet, les performances de certaines méthodes par apprentissage machine ont été rapportées comme limitées en présence de faibles quantités de données (Wei, Wang et al. 2013).

Phénotype testé	AUC	IC 95%	P
ALT (1)	0.589	0.553 - 0.625	$1,41 \times 10^{-7}$
AST (2)	0.551	0.512 - 0,590	0,016
WC (3)	0,555	0,517 - 0,593	0,017
TG (4)	0,552	0,513 - 0,590	0,021
1 + 2 + 3 + 4	0,622	0,585 - 0,659	$3,69 \times 10^{-7}$ (1) 0,037 (2) 0,014 (3) 0,023 (4)
1 + 2 + 3 + 4 + âge	0,658	0.622 - 0,694	$9,11 \times 10^{-7}$ (1) 0,035 (2) 0,019 (3) 0,057 (4) $8,8 \times 10^{-7}$ (âge)
1 + 2 + 3 + 4 + âge + sexe	0,671	0.636 - 0,707	$9,66 \times 10^{-7}$ (1) 0,044 (2) 0,016 (3) 0,042 (4) $1,28 \times 10^{-6}$ (âge) 3×10^{-4} (sexe)

Tableau 3. *Potentiel de prédiction des modèles construits par le programme LASSOSUM. Ces modèles incluent les données statistiques sommaires 1) des taux plasmatiques de ALT, 2) des taux plasmatiques de AST, 3) du tour de taille et 4) des taux plasmatiques de triglycérides. Les modèles sont éventuellement ajustés par l'âge et le sexe des patients.*

B- Exploration génétique via une approche ciblée

Deux investigations seront effectuées ici. Une première consistant en la sélection des SNPs déjà connus pour leur association avec la maladie ou avec l'une de ses composantes afin de construire un modèle de prédiction de NAFLD et une seconde étude tendant à bien caractériser l'association génétique du locus PNPLA3-SAMM50.

B-1- Conception, via sélection ciblée des SNPs, d'un modèle de prédiction de NAFLD et de la mortalité dans le DT2.

Conscient des limites de la conception d'un modèle de prédiction par apprentissage machine dans une population de taille relativement faible, j'ai décidé de concevoir un PRS en incluant des SNPs que j'aurai sélectionnés en fonction de leurs associations dans la littérature et de tester leur potentiel de prédiction et de stratification dans ADVANCE (figure 10). Les SNPs sélectionnés sont ceux, rapportés dans la littérature, comme associés à des phénotypes en relation avec la stéatose hépatique non-alcoolique.

Quatre groupes phénotypiques ont été pris en considération : 1) les phénotypes liés à la surcharge pondérale (obésité, IMC, tour de taille, rapport tour de taille / tour de hanche), 2) ceux liés au métabolisme lipidique surtout aux taux plasmatiques des triglycérides, 3) les phénotypes liés au profil du foie stéatosé (taux plasmatiques de ALT et de AST et foie stéatosé alcoolique) et 4) ceux en relation avec l'homéostasie glucidique (taux de HbA1c, taux de glucose plasmatique ou d'insuline).

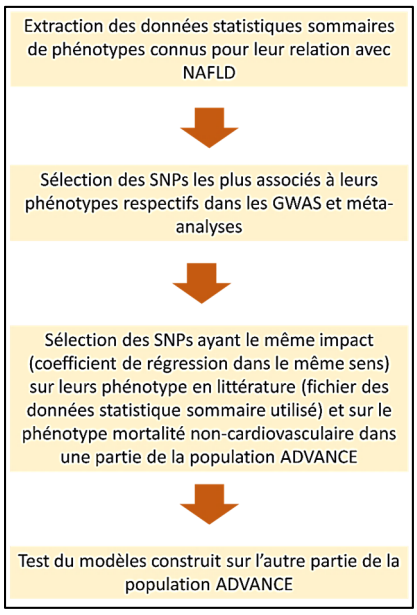

Figure 10. *Représentation schématique de la stratégie de construction du PRS via sélection ciblée des SNPs.*

Plus précisément, les fichiers statistiques sommaires issus de la base de données *UK Biobank* (assignés dans la base de données *GeneATLAS*) (http://geneatlas.roslin.ed.ac.uk/) ont été utilisés pour l'extraction de SNPs associés à l'obésité, à l'IMC, au tour de taille, et aux phénotypes fibrose-cirrhose et maladie du foie alcoolique. Par ailleurs, les données statistiques sommaires concernant le phénotype tour de taille/ tour de hanche (TT/TH) ajustés par IMC, issues de l'étude de Pulit et ses collègues (Pulit, Stoneman et al. 2019) ont été obtenues à partir de la base de données du Consortium *Giant* (https://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files#WHRadjBMI_.28download_GZIP.29). Les données statistiques générées par l'étude de Prins et ses collègues (Prins, Kuchenbaecker et al. 2017) et hébergées dans la base de données *GWAS catalog* (https://www.ebi.ac.uk/gwas/summary-statistics) ont été exploitées pour la sélection de SNPs associés aux taux plasmatiques de ALT, de AST, de triglycérides ou avec les niveaux de HbA1c. D'un autre côté, les données statistiques sommaires issues de l'étude de Willer et ses collègues (Willer, Schmidt et al. 2013) extraites de la base de données *Global Lipids Genetics Consortium* (http://csg.sph.umich.edu/willer/public/lipids2013/) ont été utilisées pour la sélection de SNPs associés aux taux plasmatiques de triglycérides. Enfin, les fichiers statistiques sommaires fournis par l'étude de Wheeler et ses collègues (Wheeler, Leong et al. 2017) et extraits de la base de données *Magic Consortiu*m (https://www.magicinvestigators.org/downloads/) ont été utilisés pour la sélection de SNPs liés aux taux de HbA1c.

Parallèlement, j'ai extrait des SNPs d'intérêt, en exploitant le fichier source des données d'associations répertoriées dans la base de donnes *GWAS catalog* (https://www.ebi.ac.uk/gwas/docs/file-downloads). Cela a été effectué en utilisant 28 mots-clés (figure 11). Une curation manuelle a été nécessaire par la suite afin d'éliminer, parmi les SNPs extraits, ceux associés à des phénotypes non liés à NAFLD. Ce processus a permis la restriction à 34 intitulés de phénotypes (figure 11).

La sélection via l'application d'un seuil d'association génétique $P = 5 \times 10^{-5}$ (*GWAS like level*) sur l'ensemble des données a permis l'extraction de 842513 SNPs (660788 SNPs uniques) (figure 12). Vu ce grand nombre de SNPs, j'ai jugé utile d'appliquer un seuil de

sélection plus stringent (seuil de signification GWAS : $P = 5 \times 10^{-8}$). Le croisement de la liste ainsi obtenue avec celle des SNPs disponibles dans ADVANCE a permis l'extraction de 111955 SNPs (figure 12).

Une dernière série de sélection a consisté en l'extraction des SNPs ayant les coefficients de régression (bêtas) les plus élevés parmi ceux avec lesquels ils sont en DL (figure 12). A noter que l'extraction des données de DL entre les 111955 a permis de générer des tableaux de tailles gigantesques (> 100 Gigabases) impossibles à exploiter dans le serveur du laboratoire. Pour y pallier, j'ai utilisé la commande --r2 du programme *Plink* afin d'extraire, au sein de ADVANCE, les données de DL dans des régions génomiques restreintes autour des 111955 SNP index. Deux cycles de sélection ont été ainsi réalisés ($r^2 \geq 0.8$ comme seuil de sélection) : les SNPs issus de la sélection lors du premier cycle ont servi de matrice pour effectuer le second cycle de sélection. Cela a permis de réduire la liste de 111955 à 6482 SNPs. Un troisième cycle a été effectué via l'exploitation des profils de DL de la population *1000 Genomes d'origine caucasienne*, répertoriée dans la base de données snipa ($r^2 \geq 0.8$ comme seuil de sélection) (Arnold, Raffler et al. 2015) (https://snipa.helmholtz-muenchen.de/snipa3/). Il a permis de ramener à 6129 le nombre de SNPs sélectionnés.

Dans la perspective d'optimiser la sélection des SNPs ayant un potentiel de prédiction de la mortalité non-cardiovasculaire, j'ai réparti de manière randomisée les individus de la population ADVANCE en deux groupes de tailles égales (2049 sujets dans chaque groupe). J'ai recherché ensuite, au sein du premier groupe de patients, des SNPs ayant des impacts sur la mortalité non-cardiovasculaire au sein de ADVANCE similaires à leurs effets sur leurs phénotypes respectifs dans la littérature (bêtas de même signe entre la littérature et ADVANCE) (figure 12). Cela a permis la construction d'un PRS de 3210 SNPs. Le test du PRS a été réalisé en 3 étapes au sein du deuxième groupe de patients :

1) Le calcul du modèle additif de chaque SNP pondéré par le coefficient de régression correspondant issu de la littérature, suivant l'équation :

$$\textbf{Modèle additif pondéré} = (1 * A + 2 * B) * |\textbf{bêta}|$$

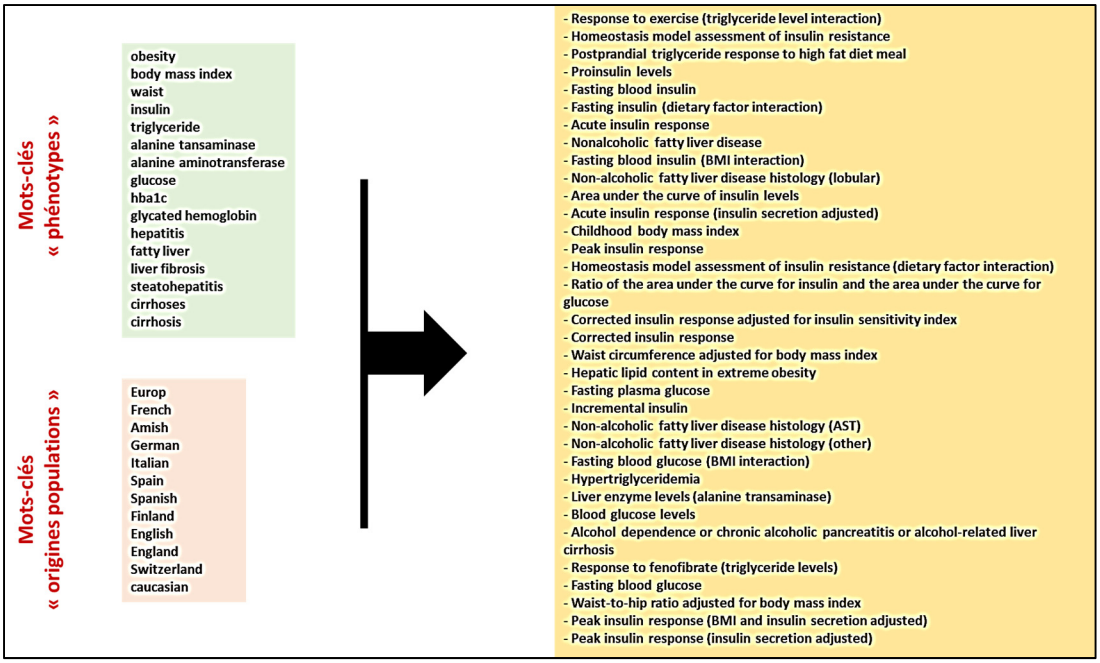

Figure 11. *Mots-clés des phénotypes et des origines des populations utilisés pour l'extraction des intitulés des phénotypes à partir des études par GWAS et des méta-analyses répertoriées dans la base de données GWAS catalog (https://www.ebi.ac.uk/gwas/).*

où **A** est la probabilité qu'a le patient d'avoir le génotype hétérozygote du SNP alors que **B** représente la probabilité que ce patient ait le génotype homozygote de l'allèle associé à la mortalité non-cardiovasculaire dans ADVANCE. **Bêta** est le coefficient de régression le plus élevé parmi ceux qui correspondent au SNP et rapportés dans les fichiers statistiques sommaires exploités dans l'étude ou dans la base de données *GWAS catalog*.

2) Les SNPs en relation avec des phénotypes identiques ou très proches ont été mis dans un même groupe de normalisation d'échelle (*scaling*). Neuf groupes de normalisation d'échelle ont été établis : **1) premier groupe** → comportant les SNPs associés taux plasmatiques des triglycérides, **2) un second groupe** → constitué de SNPs associés au phénotype binaire hypertriglycéridémie, **3) un troisième groupe** → renfermant des SNPs associés à l'IMC, **4) un quatrième groupe** → constitué de SNPs liés aux rapport tour de taille / tour de hanche, **5) un cinquième groupe** → comportant des SNPs liés au tour de

taille, **6) un sixième groupe →** renfermant des SNPs associés aux taux de ALT, 7) un septième groupe → formé de SNPs associés à la maladie du foie alcoolique, **8) un huitième groupe →** constitué de SNPs associés à la glycémie et **9) un dernier groupe →** renfermant de SNPs liés aux taux plasmatique de HbA1c.

La somme des modèles additifs pondérés des SNPs de chaque groupe de normalisation d'échelle a été calculée puis normalisée chez chaque individu afin d'obtenir un premier score. Ceci a été établi suivant l'équation :

$$\textbf{Score phénotype} = (M / S) * N$$

où **M** exprime la somme des modèles additifs pondérés des SNPs appartenant au même groupe de normalisation d'échelle, **S** est la somme des valeurs absolues des bêtas de ces polymorphismes et **N** est leur nombre.

Les SNPs ont été ensuite regroupés en 7 phénotypes et décrivent 4 groupes phénotypiques : l'obésité, la dyslipidémie, l'homéostasie glucidique et le profil stéatose du foie (tableau 4).

Enfin les SNPs appartenant à chacun des 4 groupes phénotypiques ont vu leurs scores précédemment normalisés additionnés chez chaque individu.

La prédiction par ce modèle des phénotypes « taux élevés de ALT » (> 45 UI/L), mortalité non-cardiovasculaire et mortalité cardiovasculaire a été testée.

Ces analyses ont mis en évidence un potentiel intéressant de prédiction du phénotype « taux élevés de ALT » (AUC : 0,60 IC% [0,557 – 0,640] et un peu moins en termes de prédiction de la mortalité non-cardiovasculaire 0,56 [0,510 – 0,611]. Le modèle a été encore plus faible dans la prédiction de la mortalité cardiovasculaire (AUC = 0,55 [0,496 – 0,595]) (tableau 5). L'addition au PRS des covariables « âge des patients », « sexe des patients », et « premier vecteur de la composante anthropogénétique des patients » a permis d'observer des potentiels de prédiction de 0,70 [0,65 – 0,75] de la mortalité cardiovasculaire, de 0,69 [0,66 – 0,73] du phénotype « taux élevés de ALT » et de 0,66 [0,61 – 0,70] de la mortalité non-cardiovasculaire (tableau 5).

L'analyse a révélé que le groupe phénotypique dont les scores sont les plus prédicteurs de la mortalité non-cardiovasculaire (AUC = 0.55 [0,50 – 0,60]) et du phénotype « taux élevés de ALT » (AUC = 0.60 [0,55 – 0,64]) est « profil stéatosé du foie ».

J'ai utilisé les scores de ce groupe phénotypique pour tester le potentiel de stratification du PRS. Cela a permis d'observer des valeurs plus élevées des niveaux de la mortalité non-cardiovasculaire et une fréquence plus importante des individus avec taux élevés de ALT parmi les patients ayant les 2 quintiles les plus hauts des scores.

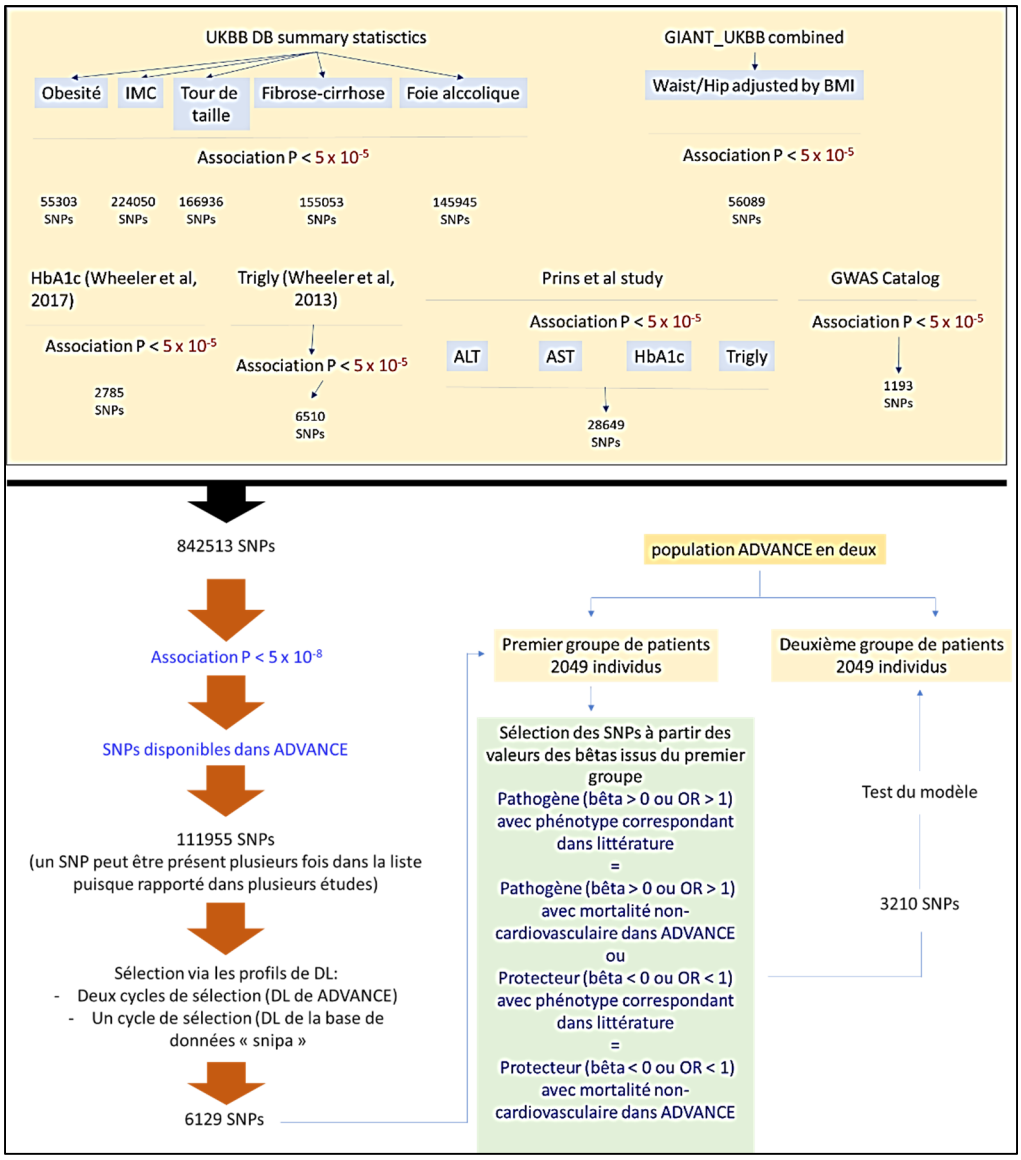

Figure 12. *Représentation des différentes étapes suivies pour la construction du modèle de prédiction de NAFLD et de la mortalité dans ADVANCE, via sélection ciblée des SNPs*

Le regroupement de ces deux quintiles et la comparaison avec le reste de la population a même mis en avant deux fois plus d'individus avec taux élevés de ALT (0,14 vs 0,07 ; P = 4,2 x 10^{-8}) et une plus grande présence de la mortalité non-cardiovasculaire (fréquence : 0,09 vs 0,06 ; P = 0,01) dans le groupe à risque (figure 13).

Phénotype	Nombre SNPs	Groupe phénotypique
Tour de taille et TT/TH	2032	Obésité
IMC	941	
Triglycerides	176	Dyslipidémie
Taux HbA1c	51	Homéostasie glucidique
Taux plasmatiques du glucose	6	
Taux ALT	3	Profil stéatosé
Maladie du foie alcoolique	1	

Tableau 4. *Phénotypes et groupes phénotypiques correspondant aux 3210 SNPs constitutifs du PRS.*

A noter que le score du groupe phénotypique « profil stéatosé du foie » est donné par 4 SNPs : rs6834314 (intergénique, proche de l'extrémité 5' du gène *hydroxysteroid 17-beta dehydrogenase 13*), rs2954021 (intergénique) ainsi que rs738409 et rs16991158, situés dans les troisièmes exon et intron du gène PNPLA3, respectivement (tableau 6).

Les SNPs rs738409 et rs16991158 ont été associés dans ADVANCE aux taux plasmatiques de ALT (P = 6.0 x 10^{-10} et P = 2.0 x 10^{-5}, respectivement) et à la mortalité non-cardiovasculaire (P = 5,2 x 10^{-5} et P = 7 x 10^{-6}, respectivement). Le retrait des deux SNPs du PRS entraine la perte de sa capacité de prédiction du phénotype « taux élevés de ALT » (AUC = 0,53 [0,49 – 0,57]) et de la mortalité non-cardiovasculaire (AUC = 0,54 [0,49 – 0,59]).

Afin de mieux comprendre ce potentiel de prédiction, j'ai testé le modèle sur les phénotypes qualitatifs dans ADVANCE. Cela a mis en avant une capacité de prédiction des évènements de l'infarctus du myocarde (AUC = 0,58 [0,52 – 0,64] et des évènements

de l'AVC (AUC = 0.56 [0,50 – 0,62]). Après ajustement par le sexe des patients, leurs âges et le premier vecteur (PC1) de leur composante anthropogénétique, les AUC des évènements de l'infarctus du myocarde et de l'AVC ont été de 0,62 [0,56 – 0,68] et 0,64 [0,58 – 0,70], respectivement.

			Groupes phénotypiques (P)			
Phénotype	**AUC**	**IC 95%**	**Homéostasie glucidique**	**Obésité**	**Dyslipidémie**	**Profil stéatosé du foie**
Taux élevés de ALT	0,60	0,557 - 0,640	0,60	0,21	0,63	**8,93E-07**
Mortalité non-cardiovasculaire	0,56	0,510 - 0,611	0,55	0,45	0,46	**0,01**
Mortalité totale	0,55	0,518 - 0,591	0,53	0,27	0,18	**0,01**
Mortalité cardiovasculaire	0,55	0,496 - 0,595	0,77	0,44	0,24	0,27

			Groupes phénotypiques (P)				Covariables (P)		
Phénotype	**AUC**	**IC 95%**	**Homéostasie glucidique**	**Obésité**	**Dyslipidémie**	**Profil stéatosé du foie**	**Age**	**Sexe**	**PC1**
Taux élevés de ALT	0,69	0,66 - 0,73	0,76	0,29	0,66	**2,94E-07**	**2,57E-09**	**8,13E-06**	0,11
Mortalité non-cardiovasculaire	0,66	0,61 - 0,70	0,39	0,60	0,49	**0,01**	**3,41E-08**	**0,001**	0,30
Mortalité totale	0,68	0,65 - 0,71	0,37	0,53	0,23	**0,01**	**9,46E-20**	**0,001**	**0,001**
Mortalité cardiovasculaire	0,70	0,65 - 0,75	0,59	0,73	0,31	0,25	**3,14E-14**	0,12	**3,31E-04**

Tableau 5. *Potentiel de prédiction du phénotype « taux élevés de ALT » et des mortalités totale, cardiovasculaire et non-cardiovasculaire par le PRS ajusté ou pas par l'âge, le sexe et la composante anthropogénétique des patients.*

L'analyse a révélé que la prédiction des évènements de l'infarctus du myocarde est déterminée par les scores des groupes phénotypiques « Obésité » et « Dyslipidémie » (AUC = 0,58 [0,516 - 0,642]) alors que les évènements de l'AVC sont déterminés en grande partie par les scores du profil phénotypique « homéostasie glucidique » (AUC = 0.56 [0.499 - 0.619]).

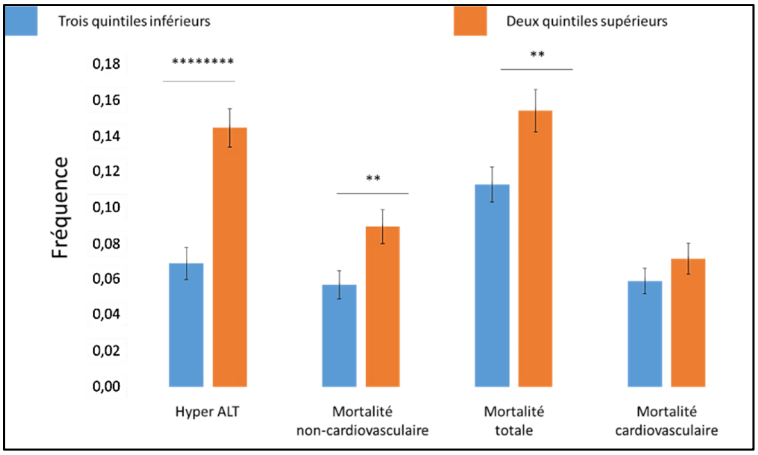

Figure 13. *Stratification par le PRS (deux quintiles supérieurs vs autres quintiles des scores du groupe phénotypique « profil stéatose hépatique ») des phénotypes « taux élevés de ALT » et mortalité non-cardiovasculaire. L'ajustement des valeurs a été effectué par l'âge, le sexe et le premier vecteur (PC1) de la composante anthropogénétique des patients.*

Pour attester de l'intérêt du PRS, j'ai recherché une plus-value en termes de marquage en comparaison avec la prédiction apportée par les paramètres clinico-biologiques correspondants au sein de ADVANCE. Les valeurs de tour de taille, d'IMC, et les taux de triglycérides, de HbA1c, de glucose et de ALT pris ensemble ont un potentiel de prédiction de la mortalité non-cardiovasculaire AUC de 0,60 [0,55 – 0,65] alors que l'addition du PRS à ces données clinico-biologiques amène le potentiel de prédiction à 0,62 [0,57 – 0,67]. Ce gain en potentiel de prédiction n'a pas été toutefois significatif (P = 0,48).

SNP	Chromosome	Gène	Phénotype associé (données de la littérature utilisées)
rs6834314	4	Intergénique	Taux de ALT
rs2954021	8	Intergénique	Taux de ALT
rs16991158	22	PNPLA3	Maladie du foie alcoolique
rs738409	22	PNPLA3	Taux de ALT

Tableau 6. *SNPs constituant le groupe phénotypique « profil stéatose hépatique » du PRS*

B-2- Criblage dense du locus PNPLA3-SAMM50

L'effet important des variations au sein du locus PNPLA3-SAMM50 sur le potentiel de prédiction des modèles que j'ai construits et son association à NAFLD et à ses composantes largement rapportée dans la littérature (Romeo, Kozlitina et al. 2008, Kotronen, Johansson et al. 2009) m'a motivé à explorer ce locus plus en détail.

Des études par GWAS réalisées avant mon arrivée au laboratoire, au sein de la population ADVANCE, ont permis de mettre en avant l'association ($P < 5.0 \times 10^{-8}$) d'un panel de SNPs, situé au niveau du locus PNPLA3-SAMM50, aux taux plasmatiques de l'enzyme ALT en ligne de base. Ce locus a été la seule région du génome associée à un niveau supérieur au seuil GWAS (figure 14).

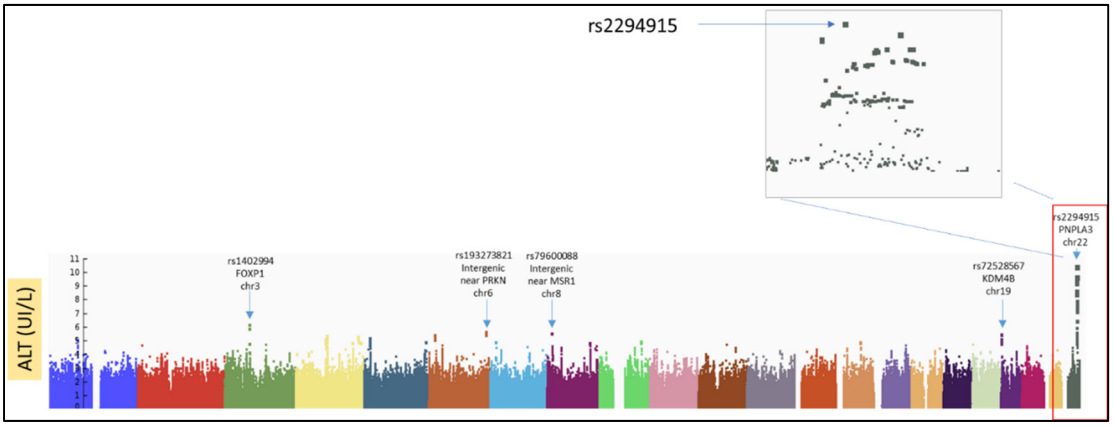

Figure 14. *Manhattan-Plot du GWAS des taux plasmatiques de l'enzyme ALT en ligne de base dans la population ADVANCE*

Les analyses par GWAS ont également mis en avant l'association du locus PNPLA3-SAMM50 au phénotype « mortalité globale » bien qu'avec un niveau de signification inférieur au seuil établi pour GWAS (association la plus significative : SNPs rs66933804 ; $P = 2{,}0 \times 10^{-5}$ puis rs1977081 ; $P = 4{,}4 \times 10^{-5}$). Le locus a été plus faiblement associé à la

mortalité d'origine cardiovasculaire (association la plus significative pour rs146570360 ; P = 0,007) (figure 15).

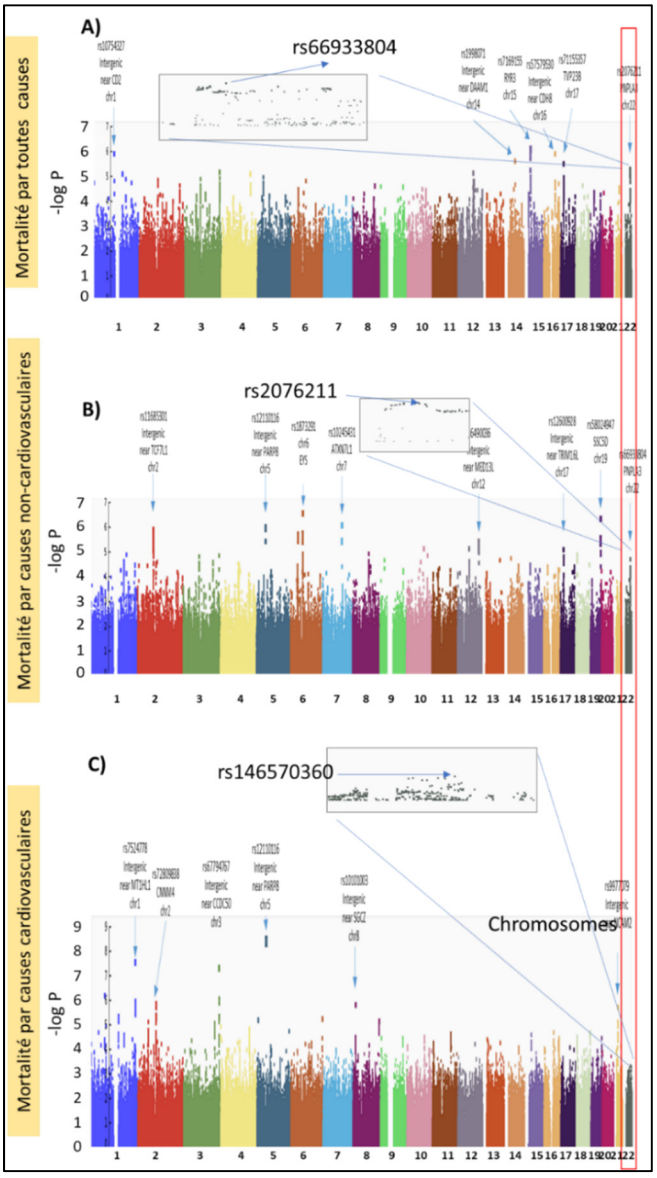

Figure 15. *Manhattan-plots des phénotypes A) « mortalité totale », B) « mortalité par causes non-cardiovasculaires » et C) « mortalité d'origine cardiovasculaire » au sein de la population ADVANCE.*

Motivé par ces résultats et afin de mieux caractériser le locus et sa relation avec la variation des taux de ALT, la mortalité et les composantes métaboliques qui y sont impliquées, j'ai analysé plus en détail les profils clinico-biologiques des patients de ADVANCE en fonction des variations de SNPs situés à différents endroits de PNPLA3-SAMM50.

Des données de GWAS disponibles dans le laboratoire ont été exploitées afin d'extraire des SNPs au sein du PNPLA3-SAMM50 en association avec des phénotypes cardiovasculaires, rénaux et de métabolisme glucidique. Deux-cent vingt et un (221) SNPs significativement associés ($P < 0,05$) à au moins un de ces phénotypes ont été utilisés pour établir une carte des associations phénotypiques au locus (figure 16). A noter que l'utilisation de l'outil *Variant Effect Predictor (VEP)* de la base de données *Ensembl* (https://grch37.ensembl.org/Homo_sapiens/Tools/VEP) a permis de constater qu'environ 1% des polymorphismes sont codants non-synonymes.

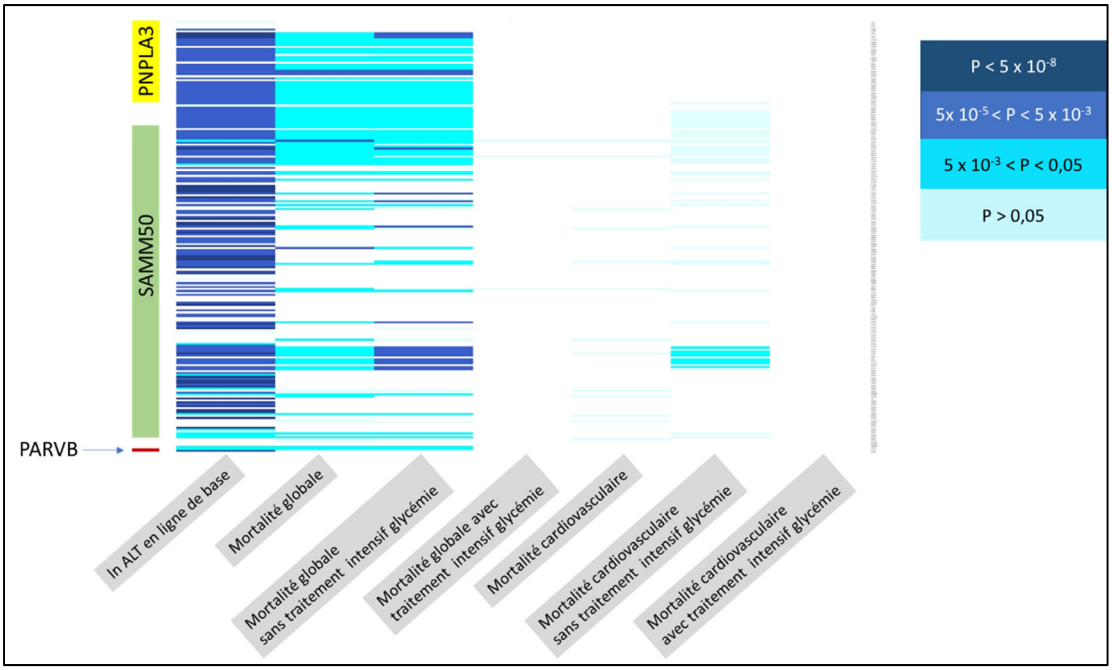

Figure 16. *Cartographie de l'association de 221 SNPs du locus PNPLA3-SAMM50 aux taux de ALT et à la mortalité globale ou cardiovasculaire au sein de la population ADVANCE en présence ou pas du traitement intensif hypoglycémiant.*

L'analyse a permis d'observer des associations génétiques fortes entre une grande partie des SNPs et les taux plasmatiques de ALT. Elle a également mis en avant l'association d'une bonne partie du locus à la mortalité globale alors que l'association à la mortalité cardiovasculaire a été rare et plutôt faible (figure 16). Cela démontre que le locus décrit la mortalité d'origine non-cardiovasculaire. Ces données démontrent donc la capacité de PNPLA3-SAMM50 à mettre en relation les variations des taux de ALT et la mortalité.

Afin de mieux apprécier le résultat, j'ai décidé d'effectuer un criblage de la région génomique. La fixation d'un seuil $P < 5 \times 10^{-3}$ a mis en avant 112 SNPs situés dans le locus significativement associés à la mortalité non-cardiovasculaire (tableau 7).

Le SNP rs66933804 a été écarté de l'analyse vu qu'il n'est pas répertorié dans la base de données BioMart (version GrCh37) (https://grch37.ensembl.org/biomart/martview/1fea0111f3ed65774d32f3befd655192).

Le traçage du profil de DL entre ces SNPs (indice D') a mis en avant un grand bloc de DL englobant cette région génomique (figure 17).

Dans le but de caractériser le locus à travers un nombre réduit de polymorphismes, j'ai sélectionné des SNPs tags en fixant le seuil de r^2 à 0,8. Ce seuil permettra de sélectionner un nombre minimum de SNPs capables de capturer une information génétique suffisante au sein de PNPLA3-SAMM50. Il représente d'ailleurs le niveau de DL au-delà duquel deux SNPs sont considérés en fort DL (HapMap 2005).

SNP	Position	Localisation
rs738407	44323955	PNPLA3 Intron 2
rs738409	44324727	PNPLA3 Exon 3
rs738408	44324730	
rs3747207	44324855	PNPLA3 Intron 3
rs201016637	44325480	
rs12485100	44325516	
rs12484801	44325565	
rs12484809	44325631	
rs12483959	44325996	
rs9625962	44326272	
rs11090617	44326700	
rs139052	44327012	
rs16991158	44327179	
rs36055245	44327192	
rs12484700	44327273	
rs4823173	44328730	
rs2076211	44329078	PNPLA3 Intron 4
rs2294433	44329275	
rs1977080	44330031	
rs1977081	44330128	
rs12484466	44330213	
rs1997693	44331513	PNPLA3 Intron 5
rs13056638	44331778	
rs1883348	44331815	
rs1883349	44331943	
rs2281138	44332477	
rs2281137	44332493	
rs2281135	44332570	
rs2072907	44332653	
rs34879941	44332878	
rs36038527	44332889	
rs2072906	44333172	PNPLA3 Intron 6
rs2076207	44333370	
rs2072905	44333479	
rs2896019	44333694	
rs2401512	44333945	
rs2896020	44333968	
rs4823176	44334476	
rs4823177	44334486	
rs4823178	44334529	
rs2281293	44334842	
rs16991175	44335331	
rs35621602	44335406	
rs34352134	44335416	
rs34376930	44335453	
rs2073081	44335744	
rs1010023	44336098	PNPLA3 Intron 7
rs1010022	44336310	
rs8142145	44336496	
rs73176497	44336957	
rs66933804	44337029	
rs926633	44337533	
rs3810622	44338134	
rs13056555	44339526	
rs2294915	44340904	PNPLA3 Intron 8
rs2294916	44340922	
rs4823179	44341193	
rs4823180	44341298	
rs4823181	44341606	
rs13055900	44341666	
rs13055874	44341672	
rs2294917	44341986	
rs2294919	44342325	PNPLA3 Intron 9
rs2008451	44342969	PNPLA3 3'UTR
rs1810508	44343151	
rs13054885	44345771	Intergénique
rs142354757	44345953	
rs7289329	44346639	
rs5764043	44346965	
rs5764044	44347137	
rs5764045	44347250	
rs2092501	44347251	
rs9614293	44347433	
rs11912828	44348116	
rs34912062	44348446	
rs1474745	44349236	

SNP	Position	Localisation
rs11705218	44354865	SAMM50 Intron 1
rs2401513	44355569	
rs2064361	44356349	
rs56373884	44356468	
rs9614294	44356566	
rs5764047	44358812	
rs16991236	44358997	
rs2073082	44360007	SAMM50 Intron 2
rs2073083	44360010	
rs5764430	44361497	SAMM50 Intron 3
rs2294921	44361842	
rs9614300	44362815	
rs3788603	44363939	
rs3761472	44368122	SAMM50 Exon 5
rs2235773	44370175	SAMM50 Intron 7
rs61473277	44371406	
rs2073088	44373579	SAMM50 Intron 9
rs2281295	44374195	SAMM50 Intron 10
rs2235776	44377999	SAMM50 Intron 11
rs4823183	44378672	
rs2235777	44378809	
rs2294922	44379565	
rs2294923	44379740	
rs9626079	44380009	SAMM50 Intron 12
rs71313378	44380171	
rs12167845	44380767	
rs4823108	44381340	
rs4823109	44381482	
rs2073079	44385594	SAMM50 Intron 13
rs7587	44386281	SAMM50 Exon 14
rs5764451	44388337	SAMM50 Intron 14
rs3827385	44388817	
rs2281296	44390547	
rs2281298	44391234	
rs2143571	44391686	
rs2401514	44394019	Intergénique
rs2073080	44394402	

Tableau 7. *Liste des 112 SNPs associés à la mortalité non-cardiovasculaire lors du GWAS dans ADVANCE, leurs coordonnées dans le chromosome 22 et leur localisation exonique, intronique ou intergénique.*

Huit SNPs ont été obtenus : rs1010022 (A/G), rs2073082(G/A), rs2143571 (G/A), rs2294915 (C/T), rs2294917 (T/C), rs738407 (T/C), rs3810622 (T/C) et rs16991236 (A/G). Ces derniers sont en équilibre de Hardy-Weinberg (HW) (seuil = 5,7 x10^{-7}) au sein de la population ADVANCE (les SNPs sont en équilibre de HW au sein de chacun des 3 groupes de patients génotypés dans les puces Affymetrix 5, 6 et UK-chip, respectivement).

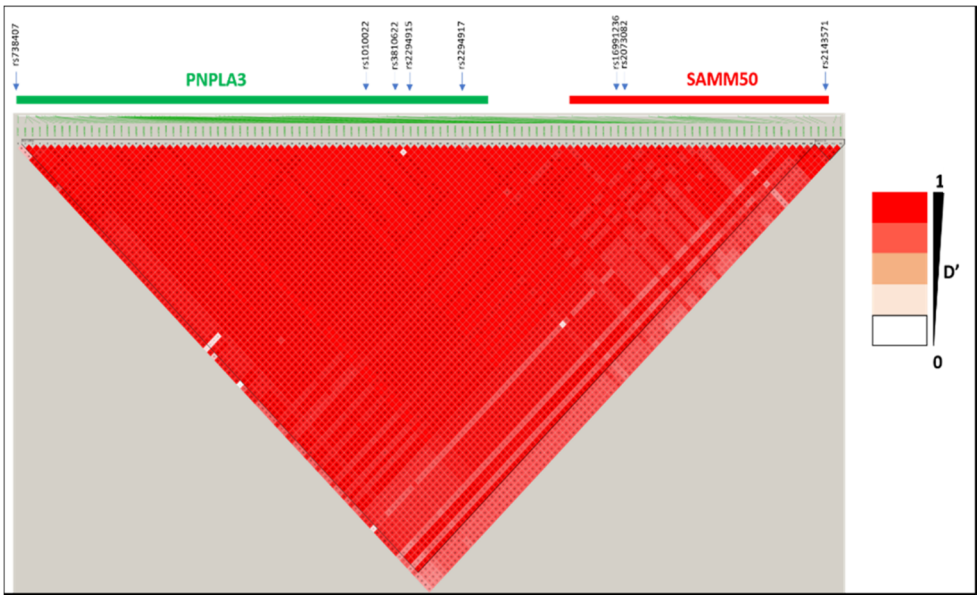

Figure 17. *Profil de DL déterminé par les 112 SNPs associés à la mortalité non-cardiovasculaire dans la population ADVANCE. Les positions des 8 SNPs tags sont indiqués dans la figure.*

L'analyse a permis d'observer la corrélation de 7 des 8 SNPs avec les taux plasmatiques de ALT (figure 18). Deux SNPs notamment ont eu des associations significatives sous le modèle additif : rs1010022 (P = 4,6 x10^{-4} (GG vs AA) ; P = 7,8 x 10^{-4} (AA vs AG)) et rs2294915 (P = 1,83 x 10^{-7} (TT vs CC) ; P = 9,4 x 10^{-8} (CC vs CT)). Après correction de Bonferroni, les différences entre les génotypes homozygotes neutres et pathogènes de chacun des deux SNPs demeurent significatives (P = 0,004 et P = 1,46 x10^{-6}, respectivement).

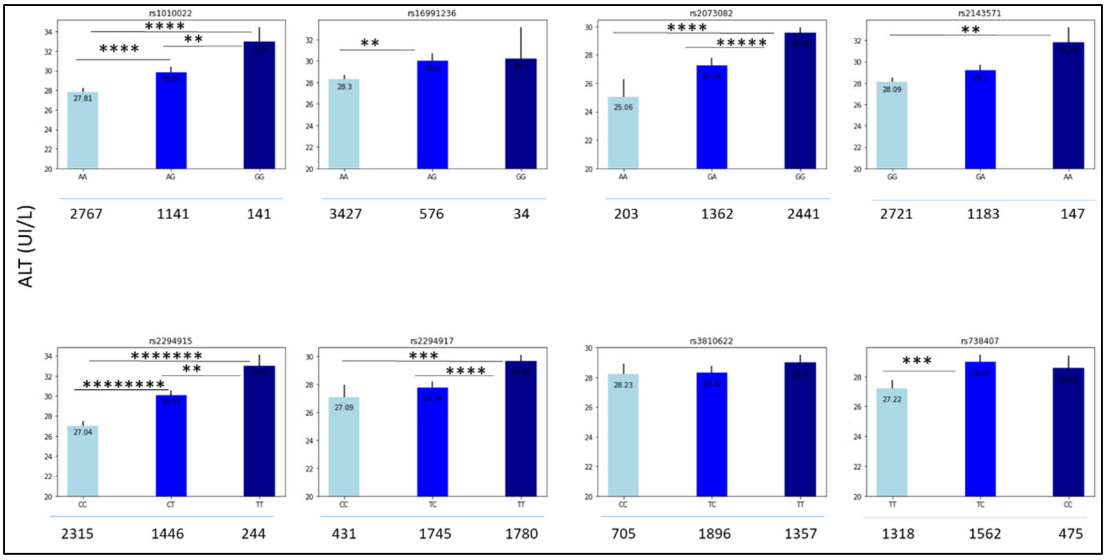

Figure 18. *Taux plasmatiques de ALT chez les patients porteurs de chacun des trois génotypes des 8 SNPs tags.*

L'investigation a également montré l'association de certains des SNPs tags à la mortalité non-cardiovasculaire (figure 19). Des différences significatives entre les génotypes neutres et les génotypes pathogènes ont été observées pour les SNPs rs1010022 (P = 0,02), rs16991236 (P = 0,03), rs2294915 (P = 0,01), rs2294917 (P = 3,9 x 10^{-4}), rs3810622 (P = 1,5 x 10^{-3}) et rs738407 (P = 0,045). Après correction de Bonferroni, ces différences restent significatives pour les SNPs rs2294917 et rs3810622 (P = 0,003 et P = 0,011, respectivement).

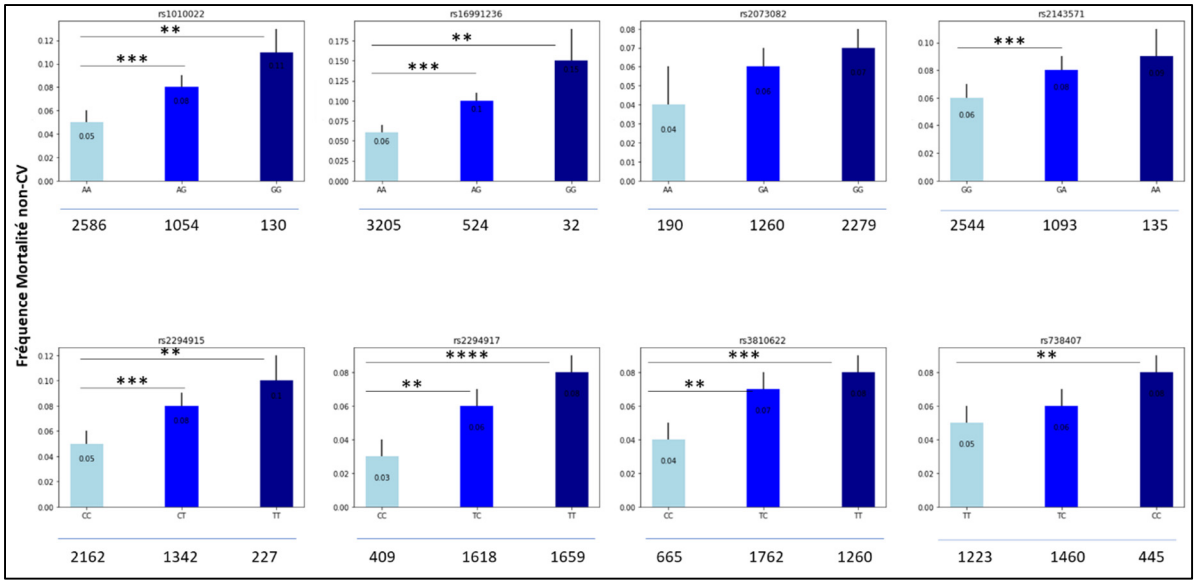

Figure 19. *Fréquence de la mortalité non-cardiovasculaire chez les patients porteurs de chacun des trois génotypes des 8 SNPs tags. Les nombres de patients dans chaque groupe génotypique sont indiqués sous les graphes (les individus décédés par causes cardiovasculaires sont exclus dans ce calcul des fréquences). Les fréquences de la mortalité non-cardiovasculaire sont rapportées dans les barres.*

L'analyse a démontré également des profils d'association variés des SNPs avec les différentes composantes du DT2. A titre d'exemple, le SNP rs16991236 a été associé à des anomalies cardiovasculaires telles que la pression artérielle diastolique en ligne de base (P = 0,01) alors qu'il n'avait pas d'association aux paramètres liés au poids ou à l'homéostasie glucidique. Le SNP rs738407 a été associé par exemple à la pression artérielle diastolique en ligne de base (P = 0,03) mais pas au profil lipidique chez les patients. Les associations les plus pertinentes des 8 SNPs tags sont rapportées dans la figure 20.

Cette variabilité d'association génétique aux composantes métaboliques révèle le potentiel qu'a le locus à décrire les différents aspects du DT2. Dans ce contexte, il est intéressant de rechercher une relation éventuelle entre PNPLA3-SAMM50 et la variabilité des effets des thérapies associées. Pour cela, je me suis focalisé sur la thérapie hypoglycémiante.

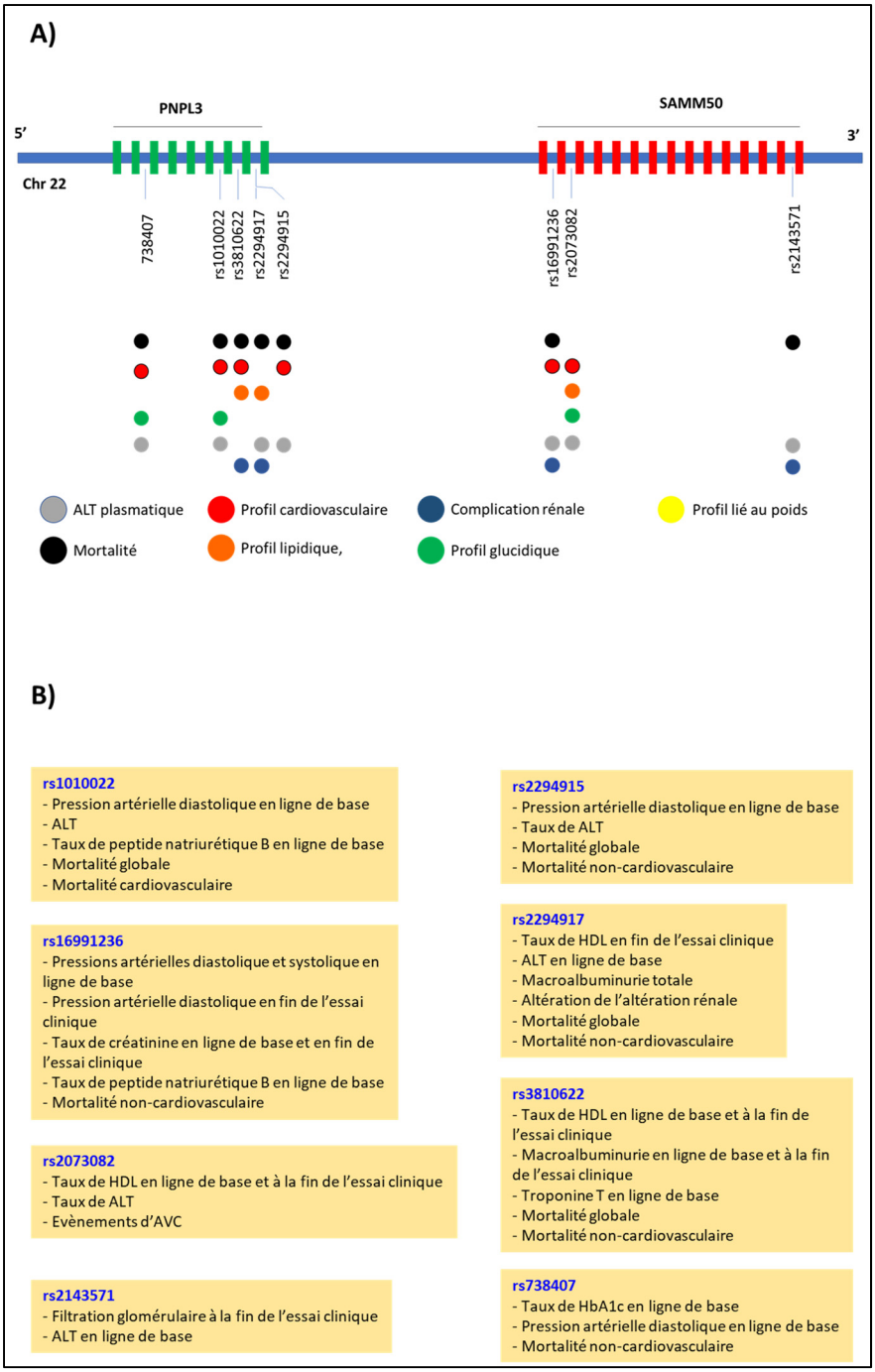

Figure 20. *Profils **(A)** et détails **(B)** des associations les plus pertinentes des 8 SNPs tags du locus PNPLA3-SAMM50 aux composantes du DT2.*

Ceci a permis de constater des intensités d'association à la mortalité non-cardiovasculaire différentes entre les SNPs. Plus précisément, il est apparu que le SNP rs16991236 (localisé

au niveau de l'intron 1 de SAMM50) arrive, via son génotype homozygote pathogène GG (double copie de l'allèle associé à ce phénotype dans notre analyse GWAS), à détecter des individus diabétiques ayant un risque élevé de mortalité non-cardiovasculaire en présence de la thérapie hypoglycémiante standard (fréquence = 0,29) alors que les individus porteurs du même génotype et soumis au traitement hypoglycémiant intensif voient ce risque diminuer de plus de 5 fois pour atteindre une fréquence de 0,05 (P = 0,007, P après correction de Bonferroni = 0,06; puissance de l'analyse = 0,74) (figure 21). Cette différence de mortalité est décrite avec moins d'intensité par les SNPs rs1010022 (fréquences 0,13 vs 0,08; P = 0,20; puissance = 0,25), rs2294915 (0,12 vs 0,08; P = 0,26; puissance = 0,20) et rs2143571 (0,11 vs 0,07; P = 0,45; puissance = 0,12). Les SNPs rs2073082, rs2294917, rs3810622 et rs738407 sont associés via leurs génotypes pathogènes à des taux de mortalité non-cardiovasculaires plus faibles et non significativement différents en présence des thérapies intensives ou standard (figure 21).

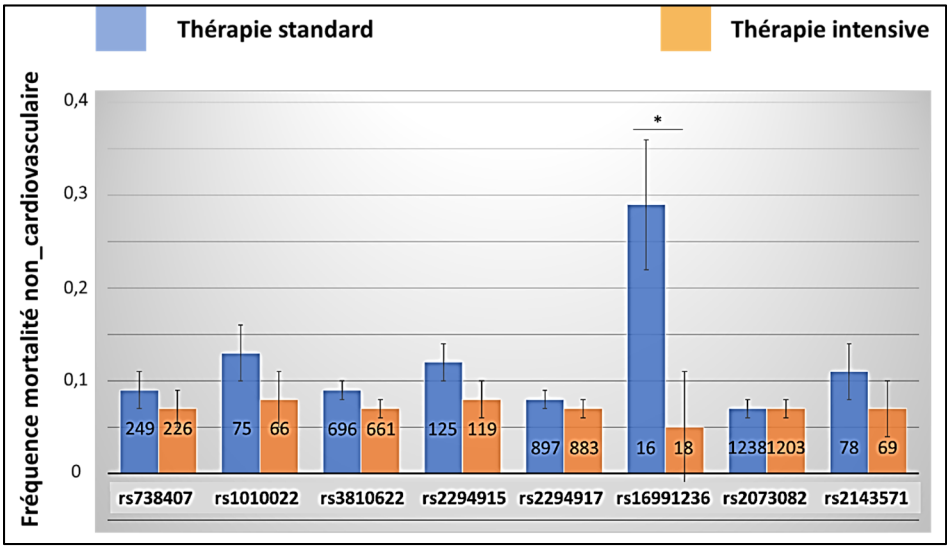

Figure 21. *Fréquence de la mortalité non-cardiovasculaire chez les patients de ADVANCE, soumis aux traitements hypoglycémiants standard ou intensif, porteurs des génotypes homozygotes pathogènes (homozygotes des allèles associés à la mortalité non-cardiovasculaire) de chacun des 8 SNPs tags. Le nombre des patients de chaque groupe génotypique est rapporté dans les barres.* Les différences significatives sont indiquées par l'étoile.

Afin d'avoir une vue d'ensemble sur le locus, j'ai exploré les fréquences de la mortalité non-cardiovasculaire associées aux 112 SNPs du locus, en présence du traitement hypoglycémiant intensif ou de la thérapie standard. Là aussi, les allèles considérés comme pathogènes sont ceux associés à la mortalité non-cardiovasculaire lors du GWAS dans ADVANCE. Cela a permis de distinguer trois groupes de patients en fonction du degré de diminution des taux de mortalité non-cardiovasculaire suite au traitement intensif :

1) Groupe 1 : inclut par exemple les patients porteurs du génotype homozygote pathogène de rs16991236 ou de rs12167845. Les taux de mortalité non-cardiovasculaire chez eux diminuent considérablement après thérapie intensive (P = 0,007 et P = 0,015, respectivement). Ils sont considérés donc comme des bons répondeurs au traitement (figure 22).

2) Groupe 2 : les d'individus appartenant à ce groupe (incluant à titre d'exemple les homozygotes pathogènes rs2294915 ou rs1977080) ont une diminution moins importante des fréquences de la mortalité non-cardiovasculaire après traitement intensif (P = 0,26 et P = 0,49, respectivement) et sont ainsi dits moins répondeurs (figure 22).

3) Groupe 3 : ce dernier groupe de patients a des taux de mortalité plus faibles et similaires en présence ou en absence de la thérapie intensive. Les patients sont par conséquence considérés comme non-répondeurs au traitement (figure 22).

Afin de mieux déterminer la composante physiopathologique qui serait liée à cette disparité en termes de réponse à la thérapie, j'ai comparé les données clinico-biologiques entre les patients porteurs du génotype homozygote pathogène (GG) du SNP rs16991236 (34 patients) au reste de la population ADVANCE. Les patients porteurs de ce génotype sont plus hypertendus (pression artérielle systolique en ligne de base: 155,3 ± 3,6 vs 147,1 ± 0,4 mmHg; P = 0,02 - pression artérielle diastolique en ligne de base: 85,4 ± 1,8 vs 80,9 ± 0,2 mmHg; P = 0,01 – pression artérielle diastolique en fin de l'essai : 77,8 ± 1,7 vs 74,2 ± 0,2 mmHg; P = 0,04). Ces individus ont également des taux de créatinine urinaire plus élevés en ligne de base (100,8 ± 4,0 vs 86,8 ± 0,5 µmol/L; P = 4,9 x 10^{-4}) et en fin d'étude (116,4 ± 7 vs 94,8 ± 0,9 µmol/L; P = 0,002). Les homozygotes GG ont par ailleurs des taux plus

importants de peptide natriurétique de type B (839, 6 ± 180,2 vs 262,4 ± 23,0 pmol/L; P = 0,001) et des fréquences plus élevées de mortalité non-cardiovasculaire (0,15 ± 0,04 vs 0,06 ± 0,01; P = 0,04).

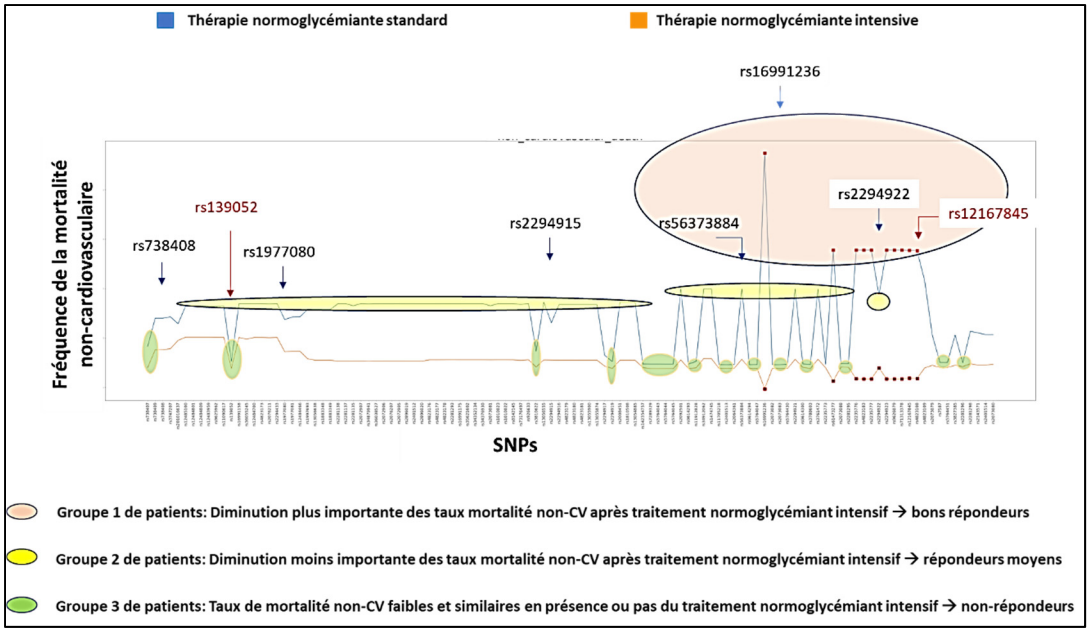

Figure 22. *Fréquence, en présence de la thérapie hypoglycémiante intensive et de la thérapie hypoglycémiante standard, de la mortalité non-cardiovasculaire chez les patients de ADVANCE porteurs des génotypes homozygotes pathogènes de chacun des 112 SNPs associés à ce phénotype. Les carrés rouges indiquent les différences significatives (P < 0,05) entre les deux thérapies.*

Discussion et conclusions

Des avancées intéressantes ont été réalisées dans le domaine de la prédiction des maladies complexes et divers outils et stratégies ont été établis. Il est actuellement possible de combiner des variations génétiques au sein des populations afin de construire des scores pour la prédiction des pathologies, leur évolution et l'efficacité des thérapies associées (Mega, Stitziel et al. 2015, Pilling, Kuo et al. 2017, Choi, Mak et al. 2018).

NAFLD est une altération hépatique complexe interagissant avec d'autres anomalies métaboliques telles que le DT2, les désordres cardiovasculaires et la mortalité (Byrne and Targher 2014, Amiri Dash Atan, Koushki et al. 2017, Hwang, Ahn et al. 2018). Elle est plus fréquente chez les diabétique de type 2 et est liée à un DT2 plus compliqué (Targher, Bertolini et al. 2006, Younossi, Koenig et al. 2016, Cusi, Sanyal et al. 2017, Wild, Walker et al. 2018). L'association de la stéatose hépatique avec les altérations cardiovasculaires serait déterminée par sa relation avec le syndrome métabolique (Targher, Bertolini et al. 2006). Vu l'absence de symptômes associés à NAFLD dans une grande partie des cas, la maladie est généralement diagnostiquée tardivement, à des stades où la prise en charge médicale devient difficile (Loguercio, De Simone et al. 2004, Friedman, Neuschwander-Tetri et al. 2018). Etablir des marqueurs non-invasifs de la pathologie est ainsi important afin de mettre des diagnostics précoces et suivre son avancement.

Diverses technologies d'imagerie ont été proposées dans le diagnostic et le suivi de la maladie. Tandis que certaines, comme l'ultrasonographie, se sont avérées plus adaptées à la détection des stéatoses du foie, les plus prononcées notamment (LaBrecque, Abbas et al. 2014), d'autres, telles que l'imagerie par élastographie, ont été indiquées dans la recherche et le suivi des fibroses hépatiques dans NAFLD (Imajo, Kessoku et al. 2016). Utiliser l'imagerie peut toutefois s'avérer coûteux dans certains pays (Papanicolas, Woskie et al. 2018) ce qui peut limiter son exploitation.

Des biomarqueurs tels que les niveaux plasmatiques de l'enzyme ALT ont été utilisés comme substitut dans le diagnostic de la maladie. Bien que les taux de l'enzyme aient été rapportés comme peu corrélés au diagnostic histopathologique de la stéatose hépatique non-alcoolique (Browning, Szczepaniak et al. 2004), ils ont été associés à diverses

altérations métaboliques et cardiovasculaires fortement présentes dans NAFLD, comme la résistance à l'insuline (Vozarova, Stefan et al. 2002, Martin-Rodriguez, Gonzalez-Cantero et al. 2017). Les niveaux de ALT ont été également liés à la mortalité surtout chez les individus appartenant aux tranches d'âges élevées (Liu, Ning et al. 2014, Schmilovitz-Weiss, Gingold-Belfer et al. 2018). Il me parait ainsi opportun qu'en absence de données de diagnostic par examen de biopsie hépatique, par imagerie (Chalasani, Younossi et al. 2018) ou par des scores clinico-biologiques comme *SteatoTest* (Poynard, Ratziu et al. 2005), d'explorer NAFLD en utilisant les taux plasmatiques de cette enzyme. Cela permet de suivre l'évolution de la pathologie et les complications qui lui sont associées et peut aider à mieux comprendre la mortalité d'origine hépatique notamment dans le DT2 (Wild, Walker et al. 2018).

Des modèles de prédiction génétique de NAFLD ont été également établis (Di Costanzo, Belardinilli et al. 2018). Le potentiel de prédiction de certains de ces modèles reste discutable vu un risque possible de surestimation des données survenant lors de leur conception (Di Costanzo, Belardinilli et al. 2018, Kawaguchi, Shima et al. 2018, Larrieta-Carrasco, Flores et al. 2018). Ces modèles sont par ailleurs constitués d'un nombre restreint de SNPs et seraient ainsi limités en termes de prédiction.

Dans ce contexte, mes travaux ont focalisé sur l'exploration du fond génétique de NAFLD au sein d'une population de diabétiques de type 2 (population ADVANCE) (Patel, MacMahon et al. 2008) en utilisant les taux plasmatiques de ALT comme substitut au diagnostic de la maladie. Le but a été de mettre au point des outils de prédiction de la pathologie et surtout des altérations métaboliques associées.

Contrairement aux données de la littérature (Liu, Ning et al. 2014, Schmilovitz-Weiss, Gingold-Belfer et al. 2018), mes analyses n'ont pas détecté d'association des taux de ALT à la mortalité totale, à la mortalité d'origine non-cardiovasculaire et à celle d'origine cardiovasculaire. Afin de mieux comprendre, j'ai exploité la méthode statistique LASSO (programme LASSOSUM) pour construire un PRS capable de prédire les taux élevés de ALT et la mortalité dans ADVANCE. Bien que les modèles construits aient été intéressants, l'inversion des populations référence et test a permis d'obtenir un PRS différent en termes de nombre de SNPs et de potentiel de prédiction. Cela serait lié à une

insuffisance d'apprentissage due à la faiblesse de la taille de population utilisée. Cette limite de performance de l'approche par apprentissage machine en présence de quantités limitées de données a été d'ailleurs rapportée dans la littérature (Wei, Wang et al. 2013). Il est donc important d'acquérir des populations et des cohortes de tailles suffisantes plus adaptées à ce genre d'analyses statistiques. Pour y pallier, j'ai appliqué une autre stratégie pour la construction du PRS. Celle-ci est basée sur une sélection ciblée des SNPs en tenant compte de leur association, rapportée dans la littérature, à des phénotypes liés à NAFLD. Une sélection supplémentaire a été faite en prenant en considération les SNPs dont les effets sur la mortalité non-cardiovasculaire dans ADVANCE et sur leurs phénotypes respectifs dans la littérature sont similaires. Cela a permis l'établissement d'un modèle de 3210 SNPs ayant un potentiel acceptable de prédiction de la mortalité non-cardiovasculaire et des taux élevés de ALT. Cependant, la capacité de prédiction est due en majeure partie aux variations au niveau du locus PNPLA3-SAMM50. Le modèle arrive toutefois via d'autres variations génétiques à prédire des altérations cardiovasculaires telles que les évènements d'infarctus du myocarde et d'AVC. A noter que l'optimisation dans la sélection des SNPs en fonction de la similarité de leurs effets sur la mortalité non-cardiovasculaire dans ADVANCE et sur leurs phénotypes dans la littérature peut être critiquée. Cependant, le test du PRS sur un groupe de patients autre que celui utilisé dans l'optimisation a montré l'efficacité du modèle. En outre, le modèle optimisé sur le phénotype mortalité non-cardiovasculaire est capable de prédire les patients ayant des taux élevés de ALT ainsi que les événements d'infarctus du myocarde et d'AVC. Cela atteste la justesse de la démarche d'optimisation et met en avant les liens étroits qui existent entre les altérations hépatiques, cardiovasculaires et le décès dans le DT2. Des analyses supplémentaires du potentiel du PRS seront effectuées. Il est en effet possible que le modèle puisse être plus performant pour prédire certaines complications au sein de certains sous-groupes de patients, ethniquement différents par exemple, et aider à titre d'exemple à comprendre les différences des profils d'altérations métaboliques observées entre les populations de ADVANCE d'origine celtique ou slave (Hamet, Haloui et al. 2017). Il est également possible que le potentiel de prédiction soit différent en fonction des thérapies prodiguées.

Le modèle comporte quatre panels de SNPs décrivant chacun un groupe phénotypique : un panel de polymorphismes associés au groupe phénotypique obésité, un deuxième panel en relation avec la dyslipidémie, un troisième lié à l'homéostasie glucidique et un panel en relation avec le phénotype « foie stéatosé ». La stratégie appliquée dans la conception du PRS a consisté en l'assignation de chacun des SNP à un de ces panels. Cependant, il serait plus approprié d'inclure chaque SNP dans l'ensemble des panels correspondant aux groupes phénotypiques auxquels ce polymorphisme a été associé dans la littérature. Cela peut améliorer la puissance du modèle et refléter davantage l'impact des groupes phénotypiques sur le potentiel de prédiction du PRS. D'autres améliorations peuvent également être apportées. En effet, les seuils de sélection des SNPs candidats à partir des données de la littérature ont été drastiques (niveaux d'association dans littérature $P < 5 \times 10^{-8}$). Il est donc possible qu'il y ait d'autres SNPs ayant des niveaux d'association plus faibles, toutefois intéressants dans la prédiction de la maladie. Des travaux futurs d'enrichissement du modèle seront réalisés et vont consister entre autres à ajouter des SNPs liés à d'autres phénotypes connus par leur association à NAFLD.

Bien que le PRS ait été construit et testé dans des groupes différents de la population ADVANCE (sélectionnés au hasard), il est essentiel de le valider dans d'autres groupes de patients. Des protocoles de collaboration ont été récemment mis en place afin d'acquérir des échantillons d'individus avec DT2 bien caractérisés sur le plan métabolique, issus de la base de données *UK Biobank* (Bycroft, Freeman et al. 2018) et de l'étude *CARTaGENE* (Awadalla, Boileau et al. 2013). Ceci permettra de tester le PRS sur des phénotypes non-rapportés dans la population ADVANCE, tels que le phénotype NAFLD attesté par histopathologie, par imagerie ou par des scores clinico-biologiques comme *SteatoTest* (Poynard, Ratziu et al. 2005) et confirmer ainsi son potentiel de prédiction. Il est également envisagé de combiner des données phénotypiques et génétiques afin d'améliorer la puissance du modèle. Cette stratégie a été d'ailleurs rapportée dans la littérature (Hyysalo, Mannisto et al. 2014). L'acquisition des populations *UK Biobank* et *CartaGene* (Awadalla, Boileau et al. 2013, Bycroft, Freeman et al. 2018) aura également comme intérêt de mettre à disposition des cohortes de plus grandes tailles plus adaptées à des analyses comme LASSO.

Afin de mieux comprendre le potentiel conféré au PRS par le locus PNPLA3-SAMM50, j'ai étudié ce dernier plus en détail dans la population ADVANCE. Le locus a été rapporté dans la littérature comme lié à l'apparition de la pathologie mais aussi à son évolution vers des stades plus compliqués comme NASH et l'hépatocarcinome du foie (Xu, Tao et al. 2015, Kanwal, Kramer et al. 2018). Toutefois, la plupart des études se sont intéressées à une variation génétique codante « SNP rs738409 » située dans la région génomique (Sookoian, Castano et al. 2009, Hotta, Yoneda et al. 2010). L'analyse d'autres polymorphismes dans le locus peut s'avérer utile afin d'améliorer le marquage de la pathologie et mettre en relief ses différents aspects.

Mes travaux ont permis de constater l'association de PNPLA3-SAMM50 à diverses composantes dans le DT2. Il s'agit notamment des variations des taux de ALT, de mortalité non-cardiovasculaire, des altérations cardiovasculaires, rénales et de l'homéostasie glucidique. Cela démontre une fois encore la relation très étroite entre les anomalies cardiométaboliques et l'atteinte hépatique surtout dans le DT2. Cette relation a été d'ailleurs rapportée dans la littérature et serait médiée par un fond de syndrome métabolique (Targher, Bertolini et al. 2006).

De manière intéressante, les profils d'association des variations génétiques au sein du locus ont été assez distincts. En effet, certaines d'entre elles ont été associées plus aux composantes cardiovasculaires alors que d'autres ont été plus en relation avec les complications rénales ou la dérégulation glycémique. Cette diversité de marquage dans la maladie est en adéquation avec les résultats de l'étude réalisée dans une population du Japon dans laquelle des groupes de polymorphismes au sein du locus ont été différemment associés à divers aspects de NAFLD (Kitamoto, Kitamoto et al. 2014).

L'analyse a rapporté également un potentiel très intéressant du locus dans la stratification des patients en fonction de leur degré de réponse au traitement hypoglycémiant intensif. En effet, PNPLA3-SAMM50 permet de distinguer des individus ayant de fortes diminutions des taux de mortalité non-cardiovasculaire suite au traitement, d'autres dont les taux diminuent plus faiblement et un troisième groupe dont les niveaux de décès d'origine non-cardiovasculaire sont faibles en présence ou en absence de la thérapie intensive. Bien que la puissance des résultats ait été limitée sur le plan statistique pour une

bonne partie des polymorphismes, l'association du SNP rs16991236 a été assez puissante et met en avant un sous-groupe de patients (34 individus) profitant amplement du traitement. Ces derniers voient les niveaux de mortalité non-cardiovasculaire diminuer 6 fois environ après thérapie hypoglycémiante intensive (puissance de l'analyse = 0,74). Ce SNP n'a pas de fonction connue dans la littérature. Il est localisé près de la partie 5' de l'exon 2 du gène SAMM50, ce qui laisse suggérer son implication dans la modification de l'épissage du gène. Dans ce contexte, il est primordial de répliquer les résultats au sein d'une autre population et de réaliser des études fonctionnelles, *in-vitro* ou dans des modèles animaux par exemple. L'ensemble des résultats démontre que l'exploitation de plusieurs polymorphismes du locus met en avant des associations non détectables par l'exploration d'un seul SNP. Il atteste également de l'intérêt de prendre en considération des SNPs non-codants pour la détection de profils d'association pas toujours observables via des variations génétiques codantes. Il est donc évident que les travaux explorant le gène PNPLA3 via un seul SNP codant (rs738409) soient limités.

Les résultats de mes études représentent une avancée de plus dans la compréhension de NAFLD et des altérations métaboliques associées, surtout la mortalité. Ils démontrent qu'il est inapproprié de séparer les composantes hépatiques de celles d'ordre cardiovasculaire ou métabolique. Les données montrent aussi que la génétique peut apporter davantage d'informations dans le sens où elle peut mieux stratifier les patients et détecter des individus à risque plus élevé ou encore des patients bénéficiant plus des thérapies prodiguées.

Les travaux effectués au sein de notre laboratoire s'inscrivent dans une dynamique internationale tendant à une meilleure exploitation des données pangénomiques (Zhou, Qi et al. 2014, Richardson, Harrison et al. 2019). En effet, des stratégies comme la combinaison des métadonnées génétiques, biologiques, des données cliniques et d'imagerie médicale sont appliquées afin d'améliorer la caractérisation des pathologies (Xu, Wu et al. 2017, Zhao, Jhamb et al. 2019). Malgré cela, les communautés scientifique et médicale restent conscientes des difficultés et des étapes à accomplir. Il est à titre d'exemple prématuré d'envisager la mise en place d'outils de prédiction appliqués en clinique dédiés à des groupes ethniques autres que les populations caucasiennes vu le

manque flagrant de données pangénomiques les concernant (Martin, Kanai et al. 2019). Par ailleurs, la diversité des outils statistiques et des cohortes mis à contribution dans la conception des modèles de risque peut être à l'origine de différences en termes de qualité de prédiction entre les kits dédiés à une même pathologie conçus par différents fabricants (Kalf, Mihaescu et al. 2014).

Des années sont encore nécessaires afin d'améliorer ces processus et mettre à la disposition des patients une médecine personnalisée plus équitable et de qualité.

Bibliographie

Abecasis, G. R., E. Noguchi, A. Heinzmann, J. A. Traherne, S. Bhattacharyya, N. I. Leaves, G. G. Anderson, Y. Zhang, N. J. Lench, A. Carey, L. R. Cardon, M. F. Moffatt and W. O. Cookson (2001). "Extent and distribution of linkage disequilibrium in three genomic regions." Am J Hum Genet **68**(1): 191-197.

Adams, L. A., S. Sanderson, K. D. Lindor and P. Angulo (2005). "The histological course of nonalcoholic fatty liver disease: a longitudinal study of 103 patients with sequential liver biopsies." J Hepatol **42**(1): 132-138.

ADVANCE-Group (2010). "Protection against cardiovascular and renal disease in type 2 diabetes: ADVANCEs in the control of blood pressure and blood glucose using Pretarax and Diamicron MR." Wolters Kluwer Pharma Solutions.

Afkarian, M., M. C. Sachs, B. Kestenbaum, I. B. Hirsch, K. R. Tuttle, J. Himmelfarb and I. H. de Boer (2013). "Kidney disease and increased mortality risk in type 2 diabetes." J Am Soc Nephrol **24**(2): 302-308.

Alexander, D. L., A. Tropsha and D. A. Winkler (2015). "Beware of R(2): Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models." J Chem Inf Model **55**(7): 1316-1322.

Altman, D. G. and J. M. Bland (1995). "Statistics notes: the normal distribution." Bmj **310**(6975): 298.

Ambler, G., S. Seaman and R. Z. Omar (2012). "An evaluation of penalised survival methods for developing prognostic models with rare events." Stat Med **31**(11-12): 1150-1161.

Amiri Dash Atan, N., M. Koushki, M. Motedayen, M. Dousti, F. Sayehmiri, R. Vafaee, M. Norouzinia and R. Gholami (2017). "Type 2 diabetes mellitus and non-alcoholic fatty liver disease: a systematic review and meta-analysis." Gastroenterol Hepatol Bed Bench **10**(Suppl1): S1-s7.

Anche, M. T., P. Bijma and M. C. De Jong (2015). "Genetic analysis of infectious diseases: estimating gene effects for susceptibility and infectivity." Genet Sel Evol **47**: 85.

Angulo, P., J. M. Hui, G. Marchesini, E. Bugianesi, J. George, G. C. Farrell, F. Enders, S. Saksena, A. D. Burt, J. P. Bida, K. Lindor, S. O. Sanderson, M. Lenzi, L. A. Adams, J. Kench, T. M. Therneau and C. P. Day (2007). "The NAFLD fibrosis score: a noninvasive system that identifies liver fibrosis in patients with NAFLD." Hepatology **45**(4): 846-854.

Arnold, M., J. Raffler, A. Pfeufer, K. Suhre and G. Kastenmuller (2015). "SNiPA: an interactive, genetic variant-centered annotation browser." Bioinformatics **31**(8): 1334-1336.

Auton, A., L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean and G. R. Abecasis (2015). "A global reference for human genetic variation." Nature **526**(7571): 68-74.

Awadalla, P., C. Boileau, Y. Payette, Y. Idaghdour, J. P. Goulet, B. Knoppers, P. Hamet and C. Laberge (2013). "Cohort profile of the CARTaGENE study: Quebec's population-based biobank for public health and personalized genomics." Int J Epidemiol **42**(5): 1285-1299.

Barreira, T. V., S. T. Broyles, A. K. Gupta and P. T. Katzmarzyk (2014). "Relationship of anthropometric indices to abdominal and total body fat in youth: sex and race differences." Obesity (Silver Spring) **22**(5): 1345-1350.

Barrett, J. C., B. Fry, J. Maller and M. J. Daly (2005). "Haploview: analysis and visualization of LD and haplotype maps." Bioinformatics **21**(2): 263-265.

Baulande, S., F. Lasnier, M. Lucas and J. Pairault (2001). "Adiponutrin, a transmembrane protein corresponding to a novel dietary- and obesity-linked mRNA specifically expressed in the adipose lineage." J Biol Chem **276**(36): 33336-33344.

Baynes, J. W. and S. R. Thorpe (1999). "Role of oxidative stress in diabetic complications: a new perspective on an old paradigm." Diabetes **48**(1): 1-9.
Bazick, J., M. Donithan, B. A. Neuschwander-Tetri, D. Kleiner, E. M. Brunt, L. Wilson, E. Doo, J. Lavine, J. Tonascia and R. Loomba (2015). "Clinical Model for NASH and Advanced Fibrosis in Adult Patients With Diabetes and NAFLD: Guidelines for Referral in NAFLD." Diabetes Care **38**(7): 1347-1355.
Bell, G. I., S. Horita and J. H. Karam (1984). "A polymorphic locus near the human insulin gene is associated with insulin-dependent diabetes mellitus." Diabetes **33**(2): 176-183.
Belsky, D. W. and S. Israel (2014). "Integrating genetics and social science: genetic risk scores." Biodemography Soc Biol **60**(2): 137-155.
Berisa, T. and J. K. Pickrell (2016). "Approximately independent linkage disequilibrium blocks in human populations." Bioinformatics **32**(2): 283-285.
Bertram, L., D. Blacker, K. Mullin, D. Keeney, J. Jones, S. Basu, S. Yhu, M. G. McInnis, R. C. Go, K. Vekrellis, D. J. Selkoe, A. J. Saunders and R. E. Tanzi (2000). "Evidence for genetic linkage of Alzheimer's disease to chromosome 10q." Science **290**(5500): 2302-2303.
Bottini, N., L. Musumeci, A. Alonso, S. Rahmouni, K. Nika, M. Rostamkhani, J. MacMurray, G. F. Meloni, P. Lucarelli, M. Pellecchia, G. S. Eisenbarth, D. Comings and T. Mustelin (2004). "A functional variant of lymphoid tyrosine phosphatase is associated with type I diabetes." Nat Genet **36**(4): 337-338.
Boyle, E. A., Y. I. Li and J. K. Pritchard (2017). "An Expanded View of Complex Traits: From Polygenic to Omnigenic." Cell **169**(7): 1177-1186.
Brenner, B. M., M. E. Cooper, D. de Zeeuw, W. F. Keane, W. E. Mitch, H. H. Parving, G. Remuzzi, S. M. Snapinn, Z. Zhang and S. Shahinfar (2001). "Effects of losartan on renal and cardiovascular outcomes in patients with type 2 diabetes and nephropathy." N Engl J Med **345**(12): 861-869.
Browning, J. D., L. S. Szczepaniak, R. Dobbins, P. Nuremberg, J. D. Horton, J. C. Cohen, S. M. Grundy and H. H. Hobbs (2004). "Prevalence of hepatic steatosis in an urban population in the United States: impact of ethnicity." Hepatology **40**(6): 1387-1395.
Bugianesi, E., A. Gastaldelli, E. Vanni, R. Gambino, M. Cassader, S. Baldi, V. Ponti, G. Pagano, E. Ferrannini and M. Rizzetto (2005). "Insulin resistance in non-diabetic patients with non-alcoholic fatty liver disease: sites and mechanisms." Diabetologia **48**(4): 634-642.
Bugianesi, E., U. Pagotto, R. Manini, E. Vanni, A. Gastaldelli, R. de Iasio, E. Gentilcore, S. Natale, M. Cassader, M. Rizzetto, R. Pasquali and G. Marchesini (2005). "Plasma adiponectin in nonalcoholic fatty liver is related to hepatic insulin resistance and hepatic fat content, not to liver disease severity." J Clin Endocrinol Metab **90**(6): 3498-3504.
Buzkova, P. (2013). "Linear regression in genetic association studies." PLoS One **8**(2): e56976.
Bycroft, C., C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O'Connell, A. Cortes, S. Welsh, A. Young, M. Effingham, G. McVean, S. Leslie, N. Allen, P. Donnelly and J. Marchini (2018). "The UK Biobank resource with deep phenotyping and genomic data." Nature **562**(7726): 203-209.
Byrne, C. D. and G. Targher (2014). "Ectopic fat, insulin resistance, and nonalcoholic fatty liver disease: implications for cardiovascular disease." Arterioscler Thromb Vasc Biol **34**(6): 1155-1161.
Caballeria, L., G. Pera, M. A. Auladell, P. Toran, L. Munoz, D. Miranda, A. Aluma, J. D. Casas, C. Sanchez, D. Gil, J. Auba, A. Tibau, S. Canut, J. Bernad and M. M. Aizpurua (2010). "Prevalence and factors associated with the presence of nonalcoholic fatty liver disease in an adult population in Spain." Eur J Gastroenterol Hepatol **22**(1): 24-32.
Campbell, P. T., C. C. Newton, A. V. Patel, E. J. Jacobs and S. M. Gapstur (2012). "Diabetes and cause-specific mortality in a prospective cohort of one million U.S. adults." Diabetes Care **35**(9): 1835-1844.

Carlson, C. S., M. A. Eberle, M. J. Rieder, Q. Yi, L. Kruglyak and D. A. Nickerson (2004). "Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium." Am J Hum Genet **74**(1): 106-120.
Carpino, G., D. Pastori, F. Baratta, D. Overi, G. Labbadia, L. Polimeni, A. Di Costanzo, G. Pannitteri, R. Carnevale, M. Del Ben, M. Arca, F. Violi, F. Angelico and E. Gaudio (2017). "PNPLA3 variant and portal/periportal histological pattern in patients with biopsy-proven non-alcoholic fatty liver disease: a possible role for oxidative stress." Sci Rep **7**(1): 15756.
Caussy, C., M. H. Alquiraish, P. Nguyen, C. Hernandez, S. Cepin, L. E. Fortney, V. Ajmera, R. Bettencourt, S. Collier, J. Hooker, E. Sy, E. Rizo, L. Richards, C. B. Sirlin and R. Loomba (2018). "Optimal threshold of controlled attenuation parameter with MRI-PDFF as the gold standard for the detection of hepatic steatosis." Hepatology **67**(4): 1348-1359.
Chalasani, N., Z. Younossi, J. E. Lavine, M. Charlton, K. Cusi, M. Rinella, S. A. Harrison, E. M. Brunt and A. J. Sanyal (2018). "The diagnosis and management of nonalcoholic fatty liver disease: Practice guidance from the American Association for the Study of Liver Diseases." Hepatology **67**(1): 328-357.
Chambers, J. C., W. Zhang, J. Sehmi, X. Li, M. N. Wass, P. Van der Harst, H. Holm, S. Sanna, M. Kavousi, S. E. Baumeister, L. J. Coin, G. Deng, C. Gieger, N. L. Heard-Costa, J. J. Hottenga, B. Kuhnel, V. Kumar, V. Lagou, L. Liang, J. Luan, P. M. Vidal, I. Mateo Leach, P. F. O'Reilly, J. F. Peden, N. Rahmioglu, P. Soininen, E. K. Speliotes, X. Yuan, G. Thorleifsson, B. Z. Alizadeh, L. D. Atwood, I. B. Borecki, M. J. Brown, P. Charoen, F. Cucca, D. Das, E. J. de Geus, A. L. Dixon, A. Doring, G. Ehret, G. I. Eyjolfsson, M. Farrall, N. G. Forouhi, N. Friedrich, W. Goessling, D. F. Gudbjartsson, T. B. Harris, A. L. Hartikainen, S. Heath, G. M. Hirschfield, A. Hofman, G. Homuth, E. Hypponen, H. L. Janssen, T. Johnson, A. J. Kangas, I. P. Kema, J. P. Kuhn, S. Lai, M. Lathrop, M. M. Lerch, Y. Li, T. J. Liang, J. P. Lin, R. J. Loos, N. G. Martin, M. F. Moffatt, G. W. Montgomery, P. B. Munroe, K. Musunuru, Y. Nakamura, C. J. O'Donnell, I. Olafsson, B. W. Penninx, A. Pouta, B. P. Prins, I. Prokopenko, R. Puls, A. Ruokonen, M. J. Savolainen, D. Schlessinger, J. N. Schouten, U. Seedorf, S. Sen-Chowdhry, K. A. Siminovitch, J. H. Smit, T. D. Spector, W. Tan, T. M. Teslovich, T. Tukiainen, A. G. Uitterlinden, M. M. Van der Klauw, R. S. Vasan, C. Wallace, H. Wallaschofski, H. E. Wichmann, G. Willemsen, P. Wurtz, C. Xu, L. M. Yerges-Armstrong, G. R. Abecasis, K. R. Ahmadi, D. I. Boomsma, M. Caulfield, W. O. Cookson, C. M. van Duijn, P. Froguel, K. Matsuda, M. I. McCarthy, C. Meisinger, V. Mooser, K. H. Pietilainen, G. Schumann, H. Snieder, M. J. Sternberg, R. P. Stolk, H. C. Thomas, U. Thorsteinsdottir, M. Uda, G. Waeber, N. J. Wareham, D. M. Waterworth, H. Watkins, J. B. Whitfield, J. C. Witteman, B. H. Wolffenbuttel, C. S. Fox, M. Ala-Korpela, K. Stefansson, P. Vollenweider, H. Volzke, E. E. Schadt, J. Scott, M. R. Jarvelin, P. Elliott and J. S. Kooner (2011). "Genome-wide association study identifies loci influencing concentrations of liver enzymes in plasma." Nat Genet **43**(11): 1131-1138.
Chan, P. H. (1996). "Role of oxidants in ischemic brain damage." Stroke **27**(6): 1124-1129.
Chatterjee, N., J. Shi and M. Garcia-Closas (2016). "Developing and evaluating polygenic risk prediction models for stratified disease prevention." Nat Rev Genet **17**(7): 392-406.
Chen, S. Y., Z. Feng and X. Yi (2017). "A general introduction to adjustment for multiple comparisons." J Thorac Dis **9**(6): 1725-1729.
Choi, S., T. Mak and P. O'Reilly (2018). "A guide to performing Polygenic Risk Score analyses." bioRxiv.
Claussnitzer, M., S. N. Dankel, K. H. Kim, G. Quon, W. Meuleman, C. Haugen, V. Glunk, I. S. Sousa, J. L. Beaudry, V. Puviindran, N. A. Abdennur, J. Liu, P. A. Svensson, Y. H. Hsu, D. J. Drucker, G. Mellgren, C. C. Hui, H. Hauner and M. Kellis (2015). "FTO Obesity Variant Circuitry and Adipocyte Browning in Humans." N Engl J Med **373**(10): 895-907.

Cohen, J. (1988). "Statistical Power Analysis for the Behavioral Sciences (Second Edition)." LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS.
Cooper, R. (2018). "Hypertension, Genes, and Environment: Challenges for Prevention and Risk Prediction." Circulation **137**(7): 662-664.
Costa-Urrutia, P., C. Abud, V. Franco-Trecu, V. Colistro, M. E. Rodriguez-Arellano, J. Vazquez-Perez, J. Granados and M. Seelaender (2017). "Genetic Obesity Risk and Attenuation Effect of Physical Fitness in Mexican-Mestizo Population: a Case-Control Study." Ann Hum Genet **81**(3): 106-116.
Cui, R., H. Iso, K. Yamagishi, I. Saito, Y. Kokubo, M. Inoue and S. Tsugane (2011). "Diabetes mellitus and risk of stroke and its subtypes among Japanese: the Japan public health center study." Stroke **42**(9): 2611-2614.
Cusi, K., A. J. Sanyal, S. Zhang, M. L. Hartman, J. M. Bue-Valleskey, B. J. Hoogwerf and A. Haupt (2017). "Non-alcoholic fatty liver disease (NAFLD) prevalence and its metabolic associations in patients with type 1 diabetes and type 2 diabetes." Diabetes Obes Metab **19**(11): 1630-1634.
Daly, M. J., J. D. Rioux, S. F. Schaffner, T. J. Hudson and E. S. Lander (2001). "High-resolution haplotype structure in the human genome." Nat Genet **29**(2): 229-232.
de Bakker, P. I., N. P. Burtt, R. R. Graham, C. Guiducci, R. Yelensky, J. A. Drake, T. Bersaglieri, K. L. Penney, J. Butler, S. Young, R. C. Onofrio, H. N. Lyon, D. O. Stram, C. A. Haiman, M. L. Freedman, X. Zhu, R. Cooper, L. Groop, L. N. Kolonel, B. E. Henderson, M. J. Daly, J. N. Hirschhorn and D. Altshuler (2006). "Transferability of tag SNPs in genetic association studies in multiple populations." Nat Genet **38**(11): 1298-1303.
De Taeye, B. M., T. Novitskaya, O. P. McGuinness, L. Gleaves, M. Medda, J. W. Covington and D. E. Vaughan (2007). "Macrophage TNF-alpha contributes to insulin resistance and hepatic steatosis in diet-induced obesity." Am J Physiol Endocrinol Metab **293**(3): E713-725.
Di Costanzo, A., F. Belardinilli, D. Bailetti, M. Sponziello, L. D'Erasmo, L. Polimeni, F. Baratta, D. Pastori, F. Ceci, A. Montali, G. Girelli, B. De Masi, A. Angeloni, G. Giannini, M. Del Ben, F. Angelico and M. Arca (2018). "Evaluation of Polygenic Determinants of Non-Alcoholic Fatty Liver Disease (NAFLD) By a Candidate Genes Resequencing Strategy." Sci Rep **8**(1): 3702.
Ding, X., N. K. Saxena, S. Lin, N. A. Gupta and F. A. Anania (2006). "Exendin-4, a glucagon-like protein-1 (GLP-1) receptor agonist, reverses hepatic steatosis in ob/ob mice." Hepatology **43**(1): 173-181.
Dobson, A. (2001). "An Introduction to Generalized Linear Models, Second Edition." CHAPMAN & HALL/CRC.
Domingue, B. W., D. W. Belsky, K. M. Harris, A. Smolen, M. B. McQueen and J. D. Boardman (2014). "Polygenic risk predicts obesity in both white and black young adults." PLoS One **9**(7): e101596.
Duggirala, R., J. Blangero, L. Almasy, T. D. Dyer, K. L. Williams, R. J. Leach, P. O'Connell and M. P. Stern (1999). "Linkage of type 2 diabetes mellitus and of age at onset to a genetic location on chromosome 10q in Mexican Americans." Am J Hum Genet **64**(4): 1127-1140.
EASL-EASD-EASO. (2016). "EASL-EASD-EASO Clinical Practice Guidelines for the management of non-alcoholic fatty liver disease." J Hepatol **64**(6): 1388-1402.
Eberle, M. A., M. J. Rieder, L. Kruglyak and D. A. Nickerson (2006). "Allele frequency matching between SNPs reveals an excess of linkage disequilibrium in genic regions of the human genome." PLoS Genet **2**(9): e142.
Eguchi, Y., H. Hyogo, M. Ono, T. Mizuta, N. Ono, K. Fujimoto, K. Chayama and T. Saibara (2012). "Prevalence and associated metabolic factors of nonalcoholic fatty liver disease in the general population from 2009 to 2010 in Japan: a multicenter large retrospective study." J Gastroenterol **47**(5): 586-595.

Einarson, T. R., A. Acs, C. Ludwig and U. H. Panton (2018). "Prevalence of cardiovascular disease in type 2 diabetes: a systematic literature review of scientific evidence from across the world in 2007-2017." Cardiovasc Diabetol **17**(1): 83.
Eissing, L., T. Scherer, K. Todter, U. Knippschild, J. W. Greve, W. A. Buurman, H. O. Pinnschmidt, S. S. Rensen, A. M. Wolf, A. Bartelt, J. Heeren, C. Buettner and L. Scheja (2013). "De novo lipogenesis in human fat and liver is linked to ChREBP-beta and metabolic health." Nat Commun **4**: 1528.
Ekstedt, M., H. Hagstrom, P. Nasr, M. Fredrikson, P. Stal, S. Kechagias and R. Hultcrantz (2015). "Fibrosis stage is the strongest predictor for disease-specific mortality in NAFLD after up to 33 years of follow-up." Hepatology **61**(5): 1547-1554.
Fawcett, T. (2003). "ROC Graphs: Notes and Practical Considerations for Data Mining Researc hers." Pattern Recognition Letters **31**: 1-38.
Fedchuk, L., F. Nascimbeni, R. Pais, F. Charlotte, C. Housset and V. Ratziu (2014). "Performance and limitations of steatosis biomarkers in patients with nonalcoholic fatty liver disease." Aliment Pharmacol Ther **40**(10): 1209-1222.
Feng, W., C. Gao, Y. Bi, M. Wu, P. Li, S. Shen, W. Chen, T. Yin and D. Zhu (2017). "Randomized trial comparing the effects of gliclazide, liraglutide, and metformin on diabetes with non-alcoholic fatty liver disease." J Diabetes **9**(8): 800-809.
Frayling, T. M., N. J. Timpson, M. N. Weedon, E. Zeggini, R. M. Freathy, C. M. Lindgren, J. R. Perry, K. S. Elliott, H. Lango, N. W. Rayner, B. Shields, L. W. Harries, J. C. Barrett, S. Ellard, C. J. Groves, B. Knight, A. M. Patch, A. R. Ness, S. Ebrahim, D. A. Lawlor, S. M. Ring, Y. Ben-Shlomo, M. R. Jarvelin, U. Sovio, A. J. Bennett, D. Melzer, L. Ferrucci, R. J. Loos, I. Barroso, N. J. Wareham, F. Karpe, K. R. Owen, L. R. Cardon, M. Walker, G. A. Hitman, C. N. Palmer, A. S. Doney, A. D. Morris, G. D. Smith, A. T. Hattersley and M. I. McCarthy (2007). "A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity." Science **316**(5826): 889-894.
Friedman, S. L., B. A. Neuschwander-Tetri, M. Rinella and A. J. Sanyal (2018). "Mechanisms of NAFLD development and therapeutic strategies." Nat Med **24**(7): 908-922.
Fuchsberger, C., J. Flannick, T. M. Teslovich, A. Mahajan, V. Agarwala, K. J. Gaulton, C. Ma, P. Fontanillas, L. Moutsianas, D. J. McCarthy, M. A. Rivas, J. R. B. Perry, X. Sim, T. W. Blackwell, N. R. Robertson, N. W. Rayner, P. Cingolani, A. E. Locke, J. F. Tajes, H. M. Highland, J. Dupuis, P. S. Chines, C. M. Lindgren, C. Hartl, A. U. Jackson, H. Chen, J. R. Huyghe, M. van de Bunt, R. D. Pearson, A. Kumar, M. Muller-Nurasyid, N. Grarup, H. M. Stringham, E. R. Gamazon, J. Lee, Y. Chen, R. A. Scott, J. E. Below, P. Chen, J. Huang, M. J. Go, M. L. Stitzel, D. Pasko, S. C. J. Parker, T. V. Varga, T. Green, N. L. Beer, A. G. Day-Williams, T. Ferreira, T. Fingerlin, M. Horikoshi, C. Hu, I. Huh, M. K. Ikram, B. J. Kim, Y. Kim, Y. J. Kim, M. S. Kwon, J. Lee, S. Lee, K. H. Lin, T. J. Maxwell, Y. Nagai, X. Wang, R. P. Welch, J. Yoon, W. Zhang, N. Barzilai, B. F. Voight, B. G. Han, C. P. Jenkinson, T. Kuulasmaa, J. Kuusisto, A. Manning, M. C. Y. Ng, N. D. Palmer, B. Balkau, A. Stancakova, H. E. Abboud, H. Boeing, V. Giedraitis, D. Prabhakaran, O. Gottesman, J. Scott, J. Carey, P. Kwan, G. Grant, J. D. Smith, B. M. Neale, S. Purcell, A. S. Butterworth, J. M. M. Howson, H. M. Lee, Y. Lu, S. H. Kwak, W. Zhao, J. Danesh, V. K. L. Lam, K. S. Park, D. Saleheen, W. Y. So, C. H. T. Tam, U. Afzal, D. Aguilar, R. Arya, T. Aung, E. Chan, C. Navarro, C. Y. Cheng, D. Palli, A. Correa, J. E. Curran, D. Rybin, V. S. Farook, S. P. Fowler, B. I. Freedman, M. Griswold, D. E. Hale, P. J. Hicks, C. C. Khor, S. Kumar, B. Lehne, D. Thuillier, W. Y. Lim, J. Liu, Y. T. van der Schouw, M. Loh, S. K. Musani, S. Puppala, W. R. Scott, L. Yengo, S. T. Tan, H. A. Taylor, Jr., F. Thameem, G. Wilson, Sr., T. Y. Wong, P. R. Njolstad, J. C. Levy, M. Mangino, L. L. Bonnycastle, T. Schwarzmayr, J. Fadista, G. L. Surdulescu, C. Herder, C. J. Groves, T. Wieland, J. Bork-Jensen, I. Brandslund, C. Christensen, H. A. Koistinen, A. S. F. Doney, L. Kinnunen, T. Esko, A. J. Farmer, L. Hakaste, D. Hodgkiss, J. Kravic, V. Lyssenko, M. Hollensted, M. E. Jorgensen, T. Jorgensen, C. Ladenvall, J. M. Justesen, A. Karajamaki, J. Kriebel, W. Rathmann, L. Lannfelt, T. Lauritzen, N. Narisu, A. Linneberg, O. Melander, L. Milani, M. Neville, M. Orho-

Melander, L. Qi, Q. Qi, M. Roden, O. Rolandsson, A. Swift, A. H. Rosengren, K. Stirrups, A. R. Wood, E. Mihailov, C. Blancher, M. O. Carneiro, J. Maguire, R. Poplin, K. Shakir, T. Fennell, M. DePristo, M. H. de Angelis, P. Deloukas, A. P. Gjesing, G. Jun, P. Nilsson, J. Murphy, R. Onofrio, B. Thorand, T. Hansen, C. Meisinger, F. B. Hu, B. Isomaa, F. Karpe, L. Liang, A. Peters, C. Huth, S. P. O'Rahilly, C. N. A. Palmer, O. Pedersen, R. Rauramaa, J. Tuomilehto, V. Salomaa, R. M. Watanabe, A. C. Syvanen, R. N. Bergman, D. Bharadwaj, E. P. Bottinger, Y. S. Cho, G. R. Chandak, J. C. N. Chan, K. S. Chia, M. J. Daly, S. B. Ebrahim, C. Langenberg, P. Elliott, K. A. Jablonski, D. M. Lehman, W. Jia, R. C. W. Ma, T. I. Pollin, M. Sandhu, N. Tandon, P. Froguel, I. Barroso, Y. Y. Teo, E. Zeggini, R. J. F. Loos, K. S. Small, J. S. Ried, R. A. DeFronzo, H. Grallert, B. Glaser, A. Metspalu, N. J. Wareham, M. Walker, E. Banks, C. Gieger, E. Ingelsson, H. K. Im, T. Illig, P. W. Franks, G. Buck, J. Trakalo, D. Buck, I. Prokopenko, R. Magi, L. Lind, Y. Farjoun, K. R. Owen, A. L. Gloyn, K. Strauch, T. Tuomi, J. S. Kooner, J. Y. Lee, T. Park, P. Donnelly, A. D. Morris, A. T. Hattersley, D. W. Bowden, F. S. Collins, G. Atzmon, J. C. Chambers, T. D. Spector, M. Laakso, T. M. Strom, G. I. Bell, J. Blangero, R. Duggirala, E. S. Tai, G. McVean, C. L. Hanis, J. G. Wilson, M. Seielstad, T. M. Frayling, J. B. Meigs, N. J. Cox, R. Sladek, E. S. Lander, S. Gabriel, N. P. Burtt, K. L. Mohlke, T. Meitinger, L. Groop, G. Abecasis, J. C. Florez, L. J. Scott, A. P. Morris, H. M. Kang, M. Boehnke, D. Altshuler and M. I. McCarthy (2016). "The genetic architecture of type 2 diabetes." Nature **536**(7614): 41-47.

Goldman, L. W. (2007). "Principles of CT and CT technology." J Nucl Med Technol **35**(3): 115-128; quiz 129-130.

Gordois, A., P. Scuffham, A. Shearer and A. Oglesby (2004). "The health care costs of diabetic nephropathy in the United States and the United Kingdom." J Diabetes Complications **18**(1): 18-26.

Gorski, M., A. Tin, M. Garnaas, G. M. McMahon, A. Y. Chu, B. O. Tayo, C. Pattaro, A. Teumer, D. I. Chasman, J. Chalmers, P. Hamet, J. Tremblay, M. Woodward, T. Aspelund, G. Eiriksdottir, V. Gudnason, T. B. Harris, L. J. Launer, A. V. Smith, B. D. Mitchell, J. R. O'Connell, A. R. Shuldiner, J. Coresh, M. Li, P. Freudenberger, E. Hofer, H. Schmidt, R. Schmidt, E. G. Holliday, P. Mitchell, J. J. Wang, I. H. de Boer, G. Li, D. S. Siscovick, Z. Kutalik, T. Corre, P. Vollenweider, G. Waeber, J. Gupta, P. A. Kanetsky, S. J. Hwang, M. Olden, Q. Yang, M. de Andrade, E. J. Atkinson, S. L. Kardia, S. T. Turner, J. M. Stafford, J. Ding, Y. Liu, C. Barlassina, D. Cusi, E. Salvi, J. A. Staessen, P. M. Ridker, H. Grallert, C. Meisinger, M. Muller-Nurasyid, B. K. Kramer, H. Kramer, S. E. Rosas, I. M. Nolte, B. W. Penninx, H. Snieder, M. Fabiola Del Greco, A. Franke, U. Nothlings, W. Lieb, S. J. Bakker, R. T. Gansevoort, P. van der Harst, A. Dehghan, O. H. Franco, A. Hofman, F. Rivadeneira, S. Sedaghat, A. G. Uitterlinden, S. Coassin, M. Haun, B. Kollerits, F. Kronenberg, B. Paulweber, N. Aumann, K. Endlich, M. Pietzner, U. Volker, R. Rettig, V. Chouraki, C. Helmer, J. C. Lambert, M. Metzger, B. Stengel, T. Lehtimaki, L. P. Lyytikainen, O. Raitakari, A. Johnson, A. Parsa, M. Bochud, I. M. Heid, W. Goessling, A. Kottgen, W. H. Kao, C. S. Fox and C. A. Boger (2015). "Genome-wide association study of kidney function decline in individuals of European descent." Kidney Int **87**(5): 1017-1029.

Gotfredsen, C. F. (1976). "Dynamics of sulfonylurea-induced insulin release from the isolated perfused rat pancreas." Diabetologia **12**(4): 339-342.

Goupil, R., S. Brachemi, A. C. Nadeau-Fredette, C. Deziel, Y. Troyanov, V. Lavergne and S. Troyanov (2013). "Lymphopenia and treatment-related infectious complications in ANCA-associated vasculitis." Clin J Am Soc Nephrol **8**(3): 416-423.

Grant, S. F., G. Thorleifsson, I. Reynisdottir, R. Benediktsson, A. Manolescu, J. Sainz, A. Helgason, H. Stefansson, V. Emilsson, A. Helgadottir, U. Styrkarsdottir, K. P. Magnusson, G. B. Walters, E. Palsdottir, T. Jonsdottir, T. Gudmundsdottir, A. Gylfason, J. Saemundsdottir, R. L. Wilensky, M. P. Reilly, D. J. Rader, Y. Bagger, C. Christiansen, V. Gudnason, G. Sigurdsson, U. Thorsteinsdottir, J. R. Gulcher, A. Kong and K. Stefansson (2006). "Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes." Nat Genet **38**(3): 320-323.

Grover, V. P., J. M. Tognarelli, M. M. Crossey, I. J. Cox, S. D. Taylor-Robinson and M. J. McPhail (2015). "Magnetic Resonance Imaging: Principles and Techniques: Lessons for Clinicians." J Clin Exp Hepatol **5**(3): 246-255.
Ha, N. T., S. Freytag and H. Bickeboeller (2014). "Coverage and efficiency in current SNP chips." Eur J Hum Genet **22**(9): 1124-1130.
Haflidadottir, S., J. G. Jonasson, H. Norland, S. O. Einarsdottir, D. E. Kleiner, S. H. Lund and E. S. Bjornsson (2014). "Long-term follow-up and liver-related death rate in patients with non-alcoholic and alcoholic related fatty liver disease." BMC Gastroenterol **14**: 166.
Hamet, P., M. Haloui, F. Harvey, F. C. Marois-Blanchet, M. P. Sylvestre, M. R. Tahir, P. H. Simon, B. S. Kanzki, J. Raelson, C. Long, J. Chalmers, M. Woodward, M. Marre, S. Harrap and J. Tremblay (2017). "PROX1 gene CC genotype as a major determinant of early onset of type 2 diabetes in slavic study participants from Action in Diabetes and Vascular Disease: Preterax and Diamicron MR Controlled Evaluation study." J Hypertens **35 Suppl 1**: S24-s32.
Hammes, H. P., S. Martin, K. Federlin, K. Geisen and M. Brownlee (1991). "Aminoguanidine treatment inhibits the development of experimental diabetic retinopathy." Proc Natl Acad Sci U S A **88**(24): 11555-11558.
Haneda, M., K. Utsunomiya, D. Koya, T. Babazono, T. Moriya, H. Makino, K. Kimura, Y. Suzuki, T. Wada, S. Ogawa, M. Inaba, Y. Kanno, T. Shigematsu, I. Masakane, K. Tsuchiya, K. Honda, K. Ichikawa and K. Shide (2015). "A new Classification of Diabetic Nephropathy 2014: a report from Joint Committee on Diabetic Nephropathy." J Diabetes Investig **6**(2): 242-246.
HapMap (2003). "The International HapMap Project." Nature **426**(6968): 789-796.
HapMap, C. (2005). "A haplotype map of the human genome." Nature **437**(7063): 1299-1320.
Hassan, M. M., A. Kaseb, C. J. Etzel, H. El-Serag, M. R. Spitz, P. Chang, K. S. Hale, M. Liu, A. Rashid, M. Shama, J. L. Abbruzzese, E. M. Loyer, H. Kaur, H. M. Hassabo, J. N. Vauthey, C. J. Wray, B. S. Hassan, Y. Z. Patt, E. Hawk, K. M. Soliman and D. Li (2013). "Genetic variation in the PNPLA3 gene and hepatocellular carcinoma in USA: risk and prognosis prediction." Mol Carcinog **52 Suppl 1**: E139-147.
Hassani, S. (1974). "Principles of ultrasonography." J Natl Med Assoc **66**(3): 205-207, 231.
Hedrick, P. W. (1987). "Gametic disequilibrium measures: proceed with caution." Genetics **117**(2): 331-341.
Hepp, T., M. Schmid, O. Gefeller, E. Waldmann and A. Mayr (2016). "Approaches to Regularized Regression - A Comparison between Gradient Boosting and the Lasso." Methods Inf Med **55**(5): 422-430.
Hill, W. G. and A. Robertson (1968). "Linkage disequilibrium in finite populations." Theor Appl Genet **38**(6): 226-231.
Hoffman, G. E. (2013). "Correcting for population structure and kinship using the linear mixed model: theory and extensions." PLoS One **8**(10): e75707.
Hoffmann, T. J., M. N. Kvale, S. E. Hesselson, Y. Zhan, C. Aquino, Y. Cao, S. Cawley, E. Chung, S. Connell, J. Eshragh, M. Ewing, J. Gollub, M. Henderson, E. Hubbell, C. Iribarren, J. Kaufman, R. Z. Lao, Y. Lu, D. Ludwig, G. K. Mathauda, W. McGuire, G. Mei, S. Miles, M. M. Purdy, C. Quesenberry, D. Ranatunga, S. Rowell, M. Sadler, M. H. Shapero, L. Shen, T. R. Shenoy, D. Smethurst, S. K. Van den Eeden, L. Walter, E. Wan, R. Wearley, T. Webster, C. C. Wen, L. Weng, R. A. Whitmer, A. Williams, S. C. Wong, C. Zau, A. Finn, C. Schaefer, P. Y. Kwok and N. Risch (2011). "Next generation genome-wide association tool: design and coverage of a high-throughput European-optimized SNP array." Genomics **98**(2): 79-89.
Hotta, K., M. Yoneda, H. Hyogo, H. Ochi, S. Mizusawa, T. Ueno, K. Chayama, A. Nakajima, K. Nakao and A. Sekine (2010). "Association of the rs738409 polymorphism in PNPLA3 with liver damage and the development of nonalcoholic fatty liver disease." BMC Med Genet **11**: 172.

Huang, Y., J. C. Cohen and H. H. Hobbs (2011). "Expression and characterization of a PNPLA3 protein isoform (I148M) associated with nonalcoholic fatty liver disease." J Biol Chem **286**(43): 37085-37093.
Hunter, C. M., M. C. Robinson, D. L. Aylor and N. D. Singh (2016). "Genetic Background, Maternal Age, and Interaction Effects Mediate Rates of Crossing Over in Drosophila melanogaster Females." G3 (Bethesda) **6**(5): 1409-1416.
Hwang, Y. C., H. Y. Ahn, S. W. Park and C. Y. Park (2018). "Nonalcoholic Fatty Liver Disease Associates With Increased Overall Mortality and Death From Cancer, Cardiovascular Disease, and Liver Disease in Women but Not Men." Clin Gastroenterol Hepatol **16**(7): 1131-1137.e1135.
Hyysalo, J., V. T. Mannisto, Y. Zhou, J. Arola, V. Karja, M. Leivonen, A. Juuti, N. Jaser, S. Lallukka, P. Kakela, S. Venesmaa, M. Simonen, J. Saltevo, L. Moilanen, E. Korpi-Hyovalti, S. Keinanen-Kiukaanniemi, H. Oksa, M. Orho-Melander, L. Valenti, S. Fargion, J. Pihlajamaki, M. Peltonen and H. Yki-Jarvinen (2014). "A population-based study on the prevalence of NASH using scores validated against liver histology." J Hepatol **60**(4): 839-846.
Imajo, K., T. Kessoku, Y. Honda, W. Tomeno, Y. Ogawa, H. Mawatari, K. Fujita, M. Yoneda, M. Taguri, H. Hyogo, Y. Sumida, M. Ono, Y. Eguchi, T. Inoue, T. Yamanaka, K. Wada, S. Saito and A. Nakajima (2016). "Magnetic Resonance Imaging More Accurately Classifies Steatosis and Fibrosis in Patients With Nonalcoholic Fatty Liver Disease Than Transient Elastography." Gastroenterology **150**(3): 626-637 e627.
Ishibashi, Y., S. Yamagishi, T. Matsui, K. Ohta, R. Tanoue, M. Takeuchi, S. Ueda, K. Nakamura and S. Okuda (2012). "Pravastatin inhibits advanced glycation end products (AGEs)-induced proximal tubular cell apoptosis and injury by reducing receptor for AGEs (RAGE) level." Metabolism **61**(8): 1067-1072.
Jarl, J., P. Desatnik, U. Peetz Hansson, K. G. Prutz and U. G. Gerdtham (2018). "Do kidney transplantations save money? A study using a before-after design and multiple register-based data from Sweden." Clin Kidney J **11**(2): 283-288.
Jenkins, C. M., D. J. Mancuso, W. Yan, H. F. Sims, B. Gibson and R. W. Gross (2004). "Identification, cloning, expression, and purification of three novel human calcium-independent phospholipase A2 family members possessing triacylglycerol lipase and acylglycerol transacylase activities." J Biol Chem **279**(47): 48968-48975.
Jiang, W., S. Huang, H. Teng, P. Wang, M. Wu, X. Zhou and H. Ran (2018). "Diagnostic accuracy of point shear wave elastography and transient elastography for staging hepatic fibrosis in patients with non-alcoholic fatty liver disease: a meta-analysis." BMJ Open **8**(8): e021787.
Jowett, J. B., J. E. Curran, M. P. Johnson, M. A. Carless, H. H. Goring, T. D. Dyer, S. A. Cole, A. G. Comuzzie, J. W. MacCluer, E. K. Moses and J. Blangero (2010). "Genetic variation at the FTO locus influences RBL2 gene expression." Diabetes **59**(3): 726-732.
Kabakov, E., C. Norymberg, E. Osher, M. Koffler, K. Tordjman, Y. Greenman and N. Stern (2006). "Prevalence of hypertension in type 2 diabetes mellitus: impact of the tightening definition of high blood pressure and association with confounding risk factors." J Cardiometab Syndr **1**(2): 95-101.
Kalf, R. R., R. Mihaescu, S. Kundu, P. de Knijff, R. C. Green and A. C. Janssens (2014). "Variations in predicted risks in personal genome testing for common complex diseases." Genet Med **16**(1): 85-91.
Kanwal, F., J. R. Kramer, S. Mapakshi, Y. Natarajan, M. Chayanupatkul, P. A. Richardson, L. Li, R. Desiderio, A. P. Thrift, S. M. Asch, J. Chu and H. B. El-Serag (2018). "Risk of Hepatocellular Cancer in Patients With Non-Alcoholic Fatty Liver Disease." Gastroenterology **155**(6): 1828-1837 e1822.
Kathiresan, S., A. K. Manning, S. Demissie, R. B. D'Agostino, A. Surti, C. Guiducci, L. Gianniny, N. P. Burtt, O. Melander, M. Orho-Melander, D. K. Arnett, G. M. Peloso, J. M. Ordovas and L. A. Cupples

(2007). "A genome-wide association study for blood lipid phenotypes in the Framingham Heart Study." BMC Med Genet **8 Suppl 1**: S17.
Kawaguchi, T., T. Shima, M. Mizuno, Y. Mitsumoto, A. Umemura, Y. Kanbara, S. Tanaka, Y. Sumida, K. Yasui, M. Takahashi, K. Matsuo, Y. Itoh, K. Tokushige, E. Hashimoto, K. Kiyosawa, M. Kawaguchi, H. Itoh, H. Uto, Y. Komorizono, K. Shirabe, S. Takami, T. Takamura, M. Kawanaka, R. Yamada, F. Matsuda and T. Okanoue (2018). "Risk estimation model for nonalcoholic fatty liver disease in the Japanese using multiple genetic markers." PLoS One **13**(1): e0185490.
Khera, A. V., M. Chaffin, K. G. Aragam, M. E. Haas, C. Roselli, S. H. Choi, P. Natarajan, E. S. Lander, S. A. Lubitz, P. T. Ellinor and S. Kathiresan (2018). "Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations." Nat Genet **50**(9): 1219-1224.
Khera, A. V., M. Chaffin, K. H. Wade, S. Zahid, J. Brancale, R. Xia, M. Distefano, O. Senol-Cosar, M. E. Haas, A. Bick, K. G. Aragam, E. S. Lander, G. D. Smith, H. Mason-Suares, M. Fornage, M. Lebo, N. J. Timpson, L. M. Kaplan and S. Kathiresan (2019). "Polygenic Prediction of Weight and Obesity Trajectories from Birth to Adulthood." Cell **177**(3): 587-596 e589.
Khoury, J. C., D. Kleindorfer, K. Alwell, C. J. Moomaw, D. Woo, O. Adeoye, M. L. Flaherty, P. Khatri, S. Ferioli, J. P. Broderick and B. M. Kissela (2013). "Diabetes mellitus: a risk factor for ischemic stroke in a large biracial population." Stroke **44**(6): 1500-1504.
Kim, D., W. Kim, A. C. Adejumo, G. Cholankeril, S. P. Tighe, R. J. Wong, S. A. Gonzalez, S. A. Harrison, Z. M. Younossi and A. Ahmed (2019). "Race/ethnicity-based temporal changes in prevalence of NAFLD-related advanced fibrosis in the United States, 2005-2016." Hepatol Int **13**(2): 205-213.
Kim, H. C., D. J. Kim and K. B. Huh (2009). "Association between nonalcoholic fatty liver disease and carotid intima-media thickness according to the presence of metabolic syndrome." Atherosclerosis **204**(2): 521-525.
Kitamoto, T., A. Kitamoto, M. Yoneda, H. Hyogo, H. Ochi, S. Mizusawa, T. Ueno, K. Nakao, A. Sekine, K. Chayama, A. Nakajima and K. Hotta (2014). "Targeted next-generation sequencing and fine linkage disequilibrium mapping reveals association of PNPLA3 and PARVB with the severity of nonalcoholic fatty liver disease." J Hum Genet **59**(5): 241-246.
Kolterman, O. G., R. S. Gray, G. Shapiro, J. A. Scarlett, J. Griffin and J. M. Olefsky (1984). "The acute and chronic effects of sulfonylurea therapy in type II diabetic subjects." Diabetes **33**(4): 346-354.
Kotronen, A., L. E. Johansson, L. M. Johansson, C. Roos, J. Westerbacka, A. Hamsten, R. Bergholm, P. Arkkila, J. Arola, T. Kiviluoto, R. M. Fisher, E. Ehrenborg, M. Orho-Melander, M. Ridderstrale, L. Groop and H. Yki-Jarvinen (2009). "A common variant in PNPLA3, which encodes adiponutrin, is associated with liver fat content in humans." Diabetologia **52**(6): 1056-1060.
Kozlitina, J., E. Smagris, S. Stender, B. G. Nordestgaard, H. H. Zhou, A. Tybjaerg-Hansen, T. F. Vogt, H. H. Hobbs and J. C. Cohen (2014). "Exome-wide association study identifies a TM6SF2 variant that confers susceptibility to nonalcoholic fatty liver disease." Nat Genet **46**(4): 352-356.
Kralovicova, J., T. R. Gaunt, S. Rodriguez, P. J. Wood, I. N. Day and I. Vorechovsky (2006). "Variants in the human insulin gene that affect pre-mRNA splicing: is -23HphI a functional single nucleotide polymorphism at IDDM2?" Diabetes **55**(1): 260-264.
Kvalseth, T. O. (1985). "Cautionary Note About R2." The American Statistician. Published by Taylor & Francis **39**(4): 279-285.
LaBrecque, D. R., Z. Abbas, F. Anania, P. Ferenci, A. G. Khan, K. L. Goh, S. S. Hamid, V. Isakov, M. Lizarzabal, M. M. Penaranda, J. F. Ramos, S. Sarin, D. Stimac, A. B. Thomson, M. Umar, J. Krabshuis and A. LeMair (2014). "World Gastroenterology Organisation global guidelines: Nonalcoholic fatty liver disease and nonalcoholic steatohepatitis." J Clin Gastroenterol **48**(6): 467-473.
Larrieta-Carrasco, E., Y. N. Flores, L. R. Macias-Kauffer, P. Ramirez-Palacios, M. Quiterio, E. G. Ramirez-Salazar, P. Leon-Mimila, B. Rivera-Paredez, G. Cabrera-Alvarez, S. Canizales-Quinteros, Z.

F. Zhang, T. V. Lopez-Perez, J. Salmeron and R. Velazquez-Cruz (2018). "Genetic variants in COL13A1, ADIPOQ and SAMM50, in addition to the PNPLA3 gene, confer susceptibility to elevated transaminase levels in an admixed Mexican population." Exp Mol Pathol **104**(1): 50-58.
Leasher, J. L., R. R. Bourne, S. R. Flaxman, J. B. Jonas, J. Keeffe, K. Naidoo, K. Pesudovs, H. Price, R. A. White, T. Y. Wong, S. Resnikoff and H. R. Taylor (2016). "Global Estimates on the Number of People Blind or Visually Impaired by Diabetic Retinopathy: A Meta-analysis From 1990 to 2010." Diabetes Care **39**(9): 1643-1649.
Lee, A. Y. and S. S. Chung (1999). "Contributions of polyol pathway to oxidative stress in diabetic cataract." FASEB J **13**(1): 23-30.
Lee, S. S. and S. H. Park (2014). "Radiologic evaluation of nonalcoholic fatty liver disease." World J Gastroenterol **20**(23): 7392-7402.
Lewis, C. M., S. C. Whitwell, A. Forbes, J. Sanderson, C. G. Mathew and T. M. Marteau (2007). "Estimating risks of common complex diseases across genetic and environmental factors: the example of Crohn disease." J Med Genet **44**(11): 689-694.
Lewontin, R. and K. Kojima (1960). "The Evolutionary Dynamics of Complex Polymorphisms." **14**(4): 458-472.
Lewontin, R. C. (1964). "The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models." Genetics **49**(1): 49-67.
Li, H., B. Hu, L. Wei, L. Zhou, L. Zhang, Y. Lin, B. Qin, Y. Dai and Z. Lu (2018). "Non-alcoholic fatty liver disease is associated with stroke severity and progression of brainstem infarctions." Eur J Neurol **25**(3): 577-e534.
Li, L., D. W. Liu, H. Y. Yan, Z. Y. Wang, S. H. Zhao and B. Wang (2016). "Obesity is an independent risk factor for non-alcoholic fatty liver disease: evidence from a meta-analysis of 21 cohort studies." Obes Rev **17**(6): 510-519.
Li, M., Y. Li, O. Weeks, V. Mijatovic, A. Teumer, J. E. Huffman, G. Tromp, C. Fuchsberger, M. Gorski, L. P. Lyytikainen, T. Nutile, S. Sedaghat, R. Sorice, A. Tin, Q. Yang, T. S. Ahluwalia, D. E. Arking, N. A. Bihlmeyer, C. A. Boger, R. J. Carroll, D. I. Chasman, M. C. Cornelis, A. Dehghan, J. D. Faul, M. F. Feitosa, G. Gambaro, P. Gasparini, F. Giulianini, I. Heid, J. Huang, M. Imboden, A. U. Jackson, J. Jeff, M. A. Jhun, R. Katz, A. Kifley, T. O. Kilpelainen, A. Kumar, M. Laakso, R. Li-Gao, K. Lohman, Y. Lu, R. Magi, G. Malerba, E. Mihailov, K. L. Mohlke, D. O. Mook-Kanamori, A. Robino, D. Ruderfer, E. Salvi, U. M. Schick, C. A. Schulz, A. V. Smith, J. A. Smith, M. Traglia, L. M. Yerges-Armstrong, W. Zhao, M. O. Goodarzi, A. T. Kraja, C. Liu, J. Wessel, E. Boerwinkle, I. B. Borecki, J. Bork-Jensen, E. P. Bottinger, D. Braga, I. Brandslund, J. A. Brody, A. Campbell, D. J. Carey, C. Christensen, J. Coresh, E. Crook, G. C. Curhan, D. Cusi, I. H. de Boer, A. P. de Vries, J. C. Denny, O. Devuyst, A. W. Dreisbach, K. Endlich, T. Esko, O. H. Franco, T. Fulop, G. S. Gerhard, C. Glumer, O. Gottesman, N. Grarup, V. Gudnason, T. Hansen, T. B. Harris, C. Hayward, L. Hocking, A. Hofman, F. B. Hu, L. L. Husemoen, R. D. Jackson, T. Jorgensen, M. E. Jorgensen, M. Kahonen, S. L. Kardia, W. Konig, C. Kooperberg, J. Kriebel, L. J. Launer, T. Lauritzen, T. Lehtimaki, D. Levy, P. Linksted, A. Linneberg, Y. Liu, R. J. Loos, A. Lupo, C. Meisinger, O. Melander, A. Metspalu, P. Mitchell, M. Nauck, P. Nurnberg, M. Orho-Melander, A. Parsa, O. Pedersen, A. Peters, U. Peters, O. Polasek, D. Porteous, N. M. Probst-Hensch, B. M. Psaty, L. Qi, O. T. Raitakari, A. P. Reiner, R. Rettig, P. M. Ridker, F. Rivadeneira, J. E. Rossouw, F. Schmidt, D. Siscovick, N. Soranzo, K. Strauch, D. Toniolo, S. T. Turner, A. G. Uitterlinden, S. Ulivi, D. Velayutham, U. Volker, H. Volzke, M. Waldenberger, J. J. Wang, D. R. Weir, D. Witte, H. Kuivaniemi, C. S. Fox, N. Franceschini, W. Goessling, A. Kottgen and A. Y. Chu (2017). "SOS2 and ACP1 Loci Identified through Large-Scale Exome Chip Analysis Regulate Kidney Development and Function." J Am Soc Nephrol **28**(3): 981-994.

Li, Q., H. Q. Qu, A. R. Rentfro, M. L. Grove, S. Mirza, Y. Lu, C. L. Hanis, M. B. Fallon, E. Boerwinkle, S. P. Fisher-Hoch and J. B. McCormick (2012). "PNPLA3 polymorphisms and liver aminotransferase levels in a Mexican American population." Clin Invest Med **35**(4): E237-245.
Liang, H., C. Vallarino, G. Joseph, S. Manne, A. Perez and S. Zhang (2014). "Increased risk of subsequent myocardial infarction in patients with type 2 diabetes: a retrospective cohort study using the U.K. General Practice Research Database." Diabetes Care **37**(5): 1329-1337.
Limpert, E., W. A. Stahel and M. Abbt (2001). "Log-normal Distributions across the Sciences: Keys and Clues: On the charms of statistics, and how mechanical models resembling gambling machines offer a link to a handy way to characterize log-normal distributions, which can provide deeper insight into variability and probability—normal or log-normal: That is the question." BioScience **51**(5): 341–352.
Liu, W., Q. M. Anstee, X. Wang, S. Gawrieh, E. R. Gamazon, S. Athinarayanan, Y. L. Liu, R. Darlay, H. J. Cordell, A. K. Daly, C. P. Day and N. Chalasani (2016). "Transcriptional regulation of PNPLA3 and its impact on susceptibility to nonalcoholic fatty liver Disease (NAFLD) in humans." Aging (Albany NY) **9**(1): 26-40.
Liu, Z., H. Ning, S. Que, L. Wang, X. Qin and T. Peng (2014). "Complex association between alanine aminotransferase activity and mortality in general population: a systematic review and meta-analysis of prospective studies." PLoS One **9**(3): e91410.
Loffroy, R., B. Terriat, V. Jooste, I. Robin, M. C. Brindisi, P. Hillon, B. Verges, J. P. Cercueil and J. M. Petit (2015). "Liver fat content is negatively associated with atherosclerotic carotid plaque in type 2 diabetic patients." Quant Imaging Med Surg **5**(6): 792-798.
Loguercio, C., T. De Simone, M. V. D'Auria, I. de Sio, A. Federico, C. Tuccillo, A. M. Abbatecola and C. Del Vecchio Blanco (2004). "Non-alcoholic fatty liver disease: a multicentre clinical study by the Italian Association for the Study of the Liver." Dig Liver Dis **36**(6): 398-405.
Loomba, R., N. Schork, C. H. Chen, R. Bettencourt, A. Bhatt, B. Ang, P. Nguyen, C. Hernandez, L. Richards, J. Salotti, S. Lin, E. Seki, K. E. Nelson, C. B. Sirlin and D. Brenner (2015). "Heritability of Hepatic Fibrosis and Steatosis Based on a Prospective Twin Study." Gastroenterology **149**(7): 1784-1793.
Lu, M., J. Zhou, C. Naylor, B. D. Kirkpatrick, R. Haque, W. A. Petri, Jr. and J. Z. Ma (2017). "Application of penalized linear regression methods to the selection of environmental enteropathy biomarkers." Biomark Res **5**: 9.
MacArthur, J., E. Bowler, M. Cerezo, L. Gil, P. Hall, E. Hastings, H. Junkins, A. McMahon, A. Milano, J. Morales, Z. M. Pendlington, D. Welter, T. Burdett, L. Hindorff, P. Flicek, F. Cunningham and H. Parkinson (2017). "The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)." Nucleic Acids Res **45**(D1): D896-d901.
MacMahon, S., R. Peto, J. Cutler, R. Collins, P. Sorlie, J. Neaton, R. Abbott, J. Godwin, A. Dyer and J. Stamler (1990). "Blood pressure, stroke, and coronary heart disease. Part 1, Prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias." Lancet **335**(8692): 765-774.
Mak, T. S. H., R. M. Porsch, S. W. Choi, X. Zhou and P. C. Sham (2017). "Polygenic scores via penalized regression on summary statistics." Genet Epidemiol **41**(6): 469-480.
Mandorfer, M., B. Scheiner, A. F. Stattermayer, P. Schwabl, R. Paternostro, D. Bauer, B. Schaefer, H. Zoller, M. Peck-Radosavljevic, M. Trauner, T. Reiberger, P. Ferenci and A. Ferlitsch (2018). "Impact of patatin-like phospholipase domain containing 3 rs738409 G/G genotype on hepatic decompensation and mortality in patients with portal hypertension." Aliment Pharmacol Ther **48**(4): 451-459.
Manolio, T. A. (2010). "Genomewide association studies and assessment of the risk of disease." N Engl J Med **363**(2): 166-176.

Martin-Rodriguez, J. L., J. Gonzalez-Cantero, A. Gonzalez-Cantero, J. P. Arrebola and J. L. Gonzalez-Calvin (2017). "Diagnostic accuracy of serum alanine aminotransferase as biomarker for nonalcoholic fatty liver disease and insulin resistance in healthy subjects, using 3T MR spectroscopy." Medicine (Baltimore) **96**(17): e6770.
Martin, A. R., M. Kanai, Y. Kamatani, Y. Okada, B. M. Neale and M. J. Daly (2019). "Clinical use of current polygenic risk scores may exacerbate health disparities." Nat Genet **51**(4): 584-591.
Mathiesen, U. L., L. E. Franzen, A. Fryden, U. Foberg and G. Bodemar (1999). "The clinical significance of slightly to moderately increased liver transaminase values in asymptomatic patients." Scand J Gastroenterol **34**(1): 85-91.
Mavaddat, N., K. Michailidou, J. Dennis, M. Lush, L. Fachal, A. Lee, J. P. Tyrer, T. H. Chen, Q. Wang, M. K. Bolla, X. Yang, M. A. Adank, T. Ahearn, K. Aittomaki, J. Allen, I. L. Andrulis, H. Anton-Culver, N. N. Antonenkova, V. Arndt, K. J. Aronson, P. L. Auer, P. Auvinen, M. Barrdahl, L. E. Beane Freeman, M. W. Beckmann, S. Behrens, J. Benitez, M. Bermisheva, L. Bernstein, C. Blomqvist, N. V. Bogdanova, S. E. Bojesen, B. Bonanni, A. L. Borresen-Dale, H. Brauch, M. Bremer, H. Brenner, A. Brentnall, I. W. Brock, A. Brooks-Wilson, S. Y. Brucker, T. Bruning, B. Burwinkel, D. Campa, B. D. Carter, J. E. Castelao, S. J. Chanock, R. Chlebowski, H. Christiansen, C. L. Clarke, J. M. Collee, E. Cordina-Duverger, S. Cornelissen, F. J. Couch, A. Cox, S. S. Cross, K. Czene, M. B. Daly, P. Devilee, T. Dork, I. Dos-Santos-Silva, M. Dumont, L. Durcan, M. Dwek, D. M. Eccles, A. B. Ekici, A. H. Eliassen, C. Ellberg, C. Engel, M. Eriksson, D. G. Evans, P. A. Fasching, J. Figueroa, O. Fletcher, H. Flyger, A. Forsti, L. Fritschi, M. Gabrielson, M. Gago-Dominguez, S. M. Gapstur, J. A. Garcia-Saenz, M. M. Gaudet, V. Georgoulias, G. G. Giles, I. R. Gilyazova, G. Glendon, M. S. Goldberg, D. E. Goldgar, A. Gonzalez-Neira, G. I. Grenaker Alnaes, M. Grip, J. Gronwald, A. Grundy, P. Guenel, L. Haeberle, E. Hahnen, C. A. Haiman, N. Hakansson, U. Hamann, S. E. Hankinson, E. F. Harkness, S. N. Hart, W. He, A. Hein, J. Heyworth, P. Hillemanns, A. Hollestelle, M. J. Hooning, R. N. Hoover, J. L. Hopper, A. Howell, G. Huang, K. Humphreys, D. J. Hunter, M. Jakimovska, A. Jakubowska, W. Janni, E. M. John, N. Johnson, M. E. Jones, A. Jukkola-Vuorinen, A. Jung, R. Kaaks, K. Kaczmarek, V. Kataja, R. Keeman, M. J. Kerin, E. Khusnutdinova, J. I. Kiiski, J. A. Knight, Y. D. Ko, V. M. Kosma, S. Koutros, V. N. Kristensen, U. Kruger, T. Kuhl, D. Lambrechts, L. Le Marchand, E. Lee, F. Lejbkowicz, J. Lilyquist, A. Lindblom, S. Lindstrom, J. Lissowska, W. Y. Lo, S. Loibl, J. Long, J. Lubinski, M. P. Lux, R. J. MacInnis, T. Maishman, E. Makalic, I. Maleva Kostovska, A. Mannermaa, S. Manoukian, S. Margolin, J. W. M. Martens, M. E. Martinez, D. Mavroudis, C. McLean, A. Meindl, U. Menon, P. Middha, N. Miller, F. Moreno, A. M. Mulligan, C. Mulot, V. M. Munoz-Garzon, S. L. Neuhausen, H. Nevanlinna, P. Neven, W. G. Newman, S. F. Nielsen, B. G. Nordestgaard, A. Norman, K. Offit, J. E. Olson, H. Olsson, N. Orr, V. S. Pankratz, T. W. Park-Simon, J. I. A. Perez, C. Perez-Barrios, P. Peterlongo, J. Peto, M. Pinchev, D. Plaseska-Karanfilska, E. C. Polley, R. Prentice, N. Presneau, D. Prokofyeva, K. Purrington, K. Pylkas, B. Rack, P. Radice, R. Rau-Murthy, G. Rennert, H. S. Rennert, V. Rhenius, M. Robson, A. Romero, K. J. Ruddy, M. Ruebner, E. Saloustros, D. P. Sandler, E. J. Sawyer, D. F. Schmidt, R. K. Schmutzler, A. Schneeweiss, M. J. Schoemaker, F. Schumacher, P. Schurmann, L. Schwentner, C. Scott, R. J. Scott, C. Seynaeve, M. Shah, M. E. Sherman, M. J. Shrubsole, X. O. Shu, S. Slager, A. Smeets, C. Sohn, P. Soucy, M. C. Southey, J. J. Spinelli, C. Stegmaier, J. Stone, A. J. Swerdlow, R. M. Tamimi, W. J. Tapper, J. A. Taylor, M. B. Terry, K. Thone, R. Tollenaar, I. Tomlinson, T. Truong, M. Tzardi, H. U. Ulmer, M. Untch, C. M. Vachon, E. M. van Veen, J. Vijai, C. R. Weinberg, C. Wendt, A. S. Whittemore, H. Wildiers, W. Willett, R. Winqvist, A. Wolk, X. R. Yang, D. Yannoukakos, Y. Zhang, W. Zheng, A. Ziogas, A. M. Dunning, D. J. Thompson, G. Chenevix-Trench, J. Chang-Claude, M. K. Schmidt, P. Hall, R. L. Milne, P. D. P. Pharoah, A. C. Antoniou, N. Chatterjee, P. Kraft, M. Garcia-Closas, J. Simard and D. F. Easton (2019). "Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes." Am J Hum Genet **104**(1): 21-34.

Mayerson, A. B., R. S. Hundal, S. Dufour, V. Lebon, D. Befroy, G. W. Cline, S. Enocksson, S. E. Inzucchi, G. I. Shulman and K. F. Petersen (2002). "The effects of rosiglitazone on insulin sensitivity, lipolysis, and hepatic and skeletal muscle triglyceride content in patients with type 2 diabetes." Diabetes **51**(3): 797-802.
McGuigan, F. E. and S. H. Ralston (2002). "Single nucleotide polymorphism detection: allelic discrimination using TaqMan." Psychiatr Genet **12**(3): 133-136.
Medici, F., M. Hawa, A. Ianari, D. A. Pyke and R. D. Leslie (1999). "Concordance rate for type II diabetes mellitus in monozygotic twins: actuarial analysis." Diabetologia **42**(2): 146-150.
Meffert, P. J., K. D. Repp, H. Volzke, F. U. Weiss, G. Homuth, J. P. Kuhn, M. M. Lerch and A. A. Aghdassi (2018). "The PNPLA3 SNP rs738409:G allele is associated with increased liver disease-associated mortality but reduced overall mortality in a population-based cohort." J Hepatol **68**(4): 858-860.
Mega, J. L., N. O. Stitziel, J. G. Smith, D. I. Chasman, M. Caulfield, J. J. Devlin, F. Nordio, C. Hyde, C. P. Cannon, F. Sacks, N. Poulter, P. Sever, P. M. Ridker, E. Braunwald, O. Melander, S. Kathiresan and M. S. Sabatine (2015). "Genetic risk, coronary heart disease events, and the clinical benefit of statin therapy: an analysis of primary and secondary prevention trials." Lancet **385**(9984): 2264-2271.
Mehta, S. H., F. L. Brancati, M. S. Sulkowski, S. A. Strathdee, M. Szklo and D. L. Thomas (2000). "Prevalence of type 2 diabetes mellitus among persons with hepatitis C virus infection in the United States." Ann Intern Med **133**(8): 592-599.
Migliore, L. and F. Coppede (2009). "Genetics, environmental factors and the emerging role of epigenetics in neurodegenerative diseases." Mutat Res **667**(1-2): 82-97.
Mills, K. T., J. D. Bundy, T. N. Kelly, J. E. Reed, P. M. Kearney, K. Reynolds, J. Chen and J. He (2016). "Global Disparities of Hypertension Prevalence and Control: A Systematic Analysis of Population-Based Studies From 90 Countries." Circulation **134**(6): 441-450.
Moses, A. (2016). "Statistical modeling and machine learning for molecular biology." Chapman and Hall/CRC.
Motamed, N., B. Rabiee, H. Poustchi, B. Dehestani, G. R. Hemasi, M. R. Khonsari, M. Maadi, F. S. Saeedian and F. Zamani (2017). "Non-alcoholic fatty liver disease (NAFLD) and 10-year risk of cardiovascular diseases." Clin Res Hepatol Gastroenterol **41**(1): 31-38.
Mueller, J. C. (2004). "Linkage disequilibrium for different scales and applications." Brief Bioinform **5**(4): 355-364.
Muller, M. (2004). "Generalized Linear Models." XploRe-Learning Guide, Springer: 204-228.
Musso, G., M. Cassader, E. Paschetta and R. Gambino (2017). "Thiazolidinediones and Advanced Liver Fibrosis in Nonalcoholic Steatohepatitis: A Meta-analysis." JAMA Intern Med **177**(5): 633-640.
Nayak, B. K. and A. Hazra (2011). "How to choose the right statistical test?" Indian J Ophthalmol **59**(2): 85-86.
NCBI (2015). "Etymologia: Bonferroni correction." Emerg Infect Dis **21**(2): 289.
Nelder, J. A. and R. W. M. Wedderburn (1972). "Generalized Linear Models." Journal of the Royal Statistical Society. Series A **135**(3): 370-384.
Nichols, G. A., S. Vupputuri and H. Lau (2011). "Medical care costs associated with progression of diabetic nephropathy." Diabetes Care **34**(11): 2374-2378.
Norton, B. J. and M. J. Strube (2001). "Understanding statistical power." J Orthop Sports Phys Ther **31**(6): 307-315.
Nseir, W. B., J. M. Mograbi, A. E. Amara, O. H. Abu Elheja and M. N. Mahamid (2019). "Non-alcoholic fatty liver disease and 30-day all-cause mortality in adult patients with community-acquired pneumonia." QJM **112**(2): 95-99.

Nunez, E., E. W. Steyerberg and J. Nunez (2011). "[Regression modeling strategies]." Rev Esp Cardiol **64**(6): 501-507.
Ogrodnik, M., S. Miwa, T. Tchkonia, D. Tiniakos, C. L. Wilson, A. Lahat, C. P. Day, A. Burt, A. Palmer, Q. M. Anstee, S. N. Grellscheid, J. H. J. Hoeijmakers, S. Barnhoorn, D. A. Mann, T. G. Bird, W. P. Vermeij, J. L. Kirkland, J. F. Passos, T. von Zglinicki and D. Jurk (2017). "Cellular senescence drives age-dependent hepatic steatosis." Nat Commun **8**: 15691.
Olefsky, J. M., J. W. Farquhar and G. M. Reaven (1974). "Reappraisal of the role of insulin in hypertriglyceridemia." Am J Med **57**(4): 551-560.
Ostchega, Y., K. S. Porter, J. Hughes, C. F. Dillon and T. Nwankwo (2011). "Resting pulse rate reference data for children, adolescents, and adults: United States, 1999-2008." Natl Health Stat Report(41): 1-16.
Ota, T., T. Takamura, S. Kurita, N. Matsuzawa, Y. Kita, M. Uno, H. Akahori, H. Misu, M. Sakurai, Y. Zen, Y. Nakanuma and S. Kaneko (2007). "Insulin resistance accelerates a dietary rat model of nonalcoholic steatohepatitis." Gastroenterology **132**(1): 282-293.
Papanicolas, I., L. R. Woskie and A. K. Jha (2018). "Health Care Spending in the United States and Other High-Income Countries." Jama **319**(10): 1024-1039.
Patel, A., S. MacMahon, J. Chalmers, B. Neal, L. Billot, M. Woodward, M. Marre, M. Cooper, P. Glasziou, D. Grobbee, P. Hamet, S. Harrap, S. Heller, L. Liu, G. Mancia, C. E. Mogensen, C. Pan, N. Poulter, A. Rodgers, B. Williams, S. Bompoint, B. E. de Galan, R. Joshi and F. Travert (2008). "Intensive blood glucose control and vascular outcomes in patients with type 2 diabetes." N Engl J Med **358**(24): 2560-2572.
Patel, A., S. MacMahon, J. Chalmers, B. Neal, M. Woodward, L. Billot, S. Harrap, N. Poulter, M. Marre, M. Cooper, P. Glasziou, D. E. Grobbee, P. Hamet, S. Heller, L. S. Liu, G. Mancia, C. E. Mogensen, C. Y. Pan, A. Rodgers and B. Williams (2007). "Effects of a fixed combination of perindopril and indapamide on macrovascular and microvascular outcomes in patients with type 2 diabetes mellitus (the ADVANCE trial): a randomised controlled trial." Lancet **370**(9590): 829-840.
Patel, C. J., R. Chen, K. Kodama, J. P. Ioannidis and A. J. Butte (2013). "Systematic identification of interaction effects between genome- and environment-wide associations in type 2 diabetes mellitus." Hum Genet **132**(5): 495-508.
Perley, M. and D. M. Kipnis (1966). "Plasma insulin responses to glucose and tolbutamide of normal weight and obese diabetic and nondiabetic subjects." Diabetes **15**(12): 867-874.
Petit, J. M., J. P. Cercueil, R. Loffroy, D. Denimal, B. Bouillet, C. Fourmont, O. Chevallier, L. Duvillard and B. Verges (2017). "Effect of Liraglutide Therapy on Liver Fat Content in Patients With Inadequately Controlled Type 2 Diabetes: The Lira-NAFLD Study." J Clin Endocrinol Metab **102**(2): 407-415.
Pilling, L. C., C. L. Kuo, K. Sicinski, J. Tamosauskaite, G. A. Kuchel, L. W. Harries, P. Herd, R. Wallace, L. Ferrucci and D. Melzer (2017). "Human longevity: 25 genetic loci associated in 389,166 UK biobank participants." Aging (Albany NY) **9**(12): 2504-2520.
Pinero, J., A. Bravo, N. Queralt-Rosinach, A. Gutierrez-Sacristan, J. Deu-Pons, E. Centeno, J. Garcia-Garcia, F. Sanz and L. I. Furlong (2017). "DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants." Nucleic Acids Res **45**(D1): D833-d839.
Pingitore, P., C. Pirazzi, R. M. Mancina, B. M. Motta, C. Indiveri, A. Pujia, T. Montalcini, K. Hedfalk and S. Romeo (2014). "Recombinant PNPLA3 protein shows triglyceride hydrolase activity and its I148M mutation results in loss of function." Biochim Biophys Acta **1841**(4): 574-580.

Poynard, T., V. Ratziu, S. Naveau, D. Thabut, F. Charlotte, D. Messous, D. Capron, A. Abella, J. Massard, Y. Ngo, M. Munteanu, A. Mercadier, M. Manns and J. Albrecht (2005). "The diagnostic value of biomarkers (SteatoTest) for the prediction of liver steatosis." Comp Hepatol **4**: 10.
Prins, B. P., K. B. Kuchenbaecker, Y. Bao, M. Smart, D. Zabaneh, G. Fatemifar, J. Luan, N. J. Wareham, R. A. Scott, J. R. B. Perry, C. Langenberg, M. Benzeval, M. Kumari and E. Zeggini (2017). "Genome-wide analysis of health-related biomarkers in the UK Household Longitudinal Study reveals novel associations." Sci Rep **7**(1): 11008.
Pulit, S. L., C. Stoneman, A. P. Morris, A. R. Wood, C. A. Glastonbury, J. Tyrrell, L. Yengo, T. Ferreira, E. Marouli, Y. Ji, J. Yang, S. Jones, R. Beaumont, D. C. Croteau-Chonka, T. W. Winkler, G. Consortium, A. T. Hattersley, R. J. F. Loos, J. N. Hirschhorn, P. M. Visscher, T. M. Frayling, H. Yaghootkar and C. M. Lindgren (2019). "Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry." Hum Mol Genet **28**(1): 166-174.
Reddy, M. V., H. Wang, S. Liu, B. Bode, J. C. Reed, R. D. Steed, S. W. Anderson, L. Steed, D. Hopkins and J. X. She (2011). "Association between type 1 diabetes and GWAS SNPs in the southeast US Caucasian population." Genes Immun **12**(3): 208-212.
Rich, N. E., S. Oji, A. R. Mufti, J. D. Browning, N. D. Parikh, M. Odewole, H. Mayo and A. G. Singal (2018). "Racial and Ethnic Disparities in Nonalcoholic Fatty Liver Disease Prevalence, Severity, and Outcomes in the United States: A Systematic Review and Meta-analysis." Clin Gastroenterol Hepatol **16**(2): 198-210.e192.
Richardson, T. G., S. Harrison, G. Hemani and G. Davey Smith (2019). "An atlas of polygenic risk score associations to highlight putative causal relationships across the human phenome." Elife **8**.
Romeo, S., J. Kozlitina, C. Xing, A. Pertsemlidis, D. Cox, L. A. Pennacchio, E. Boerwinkle, J. C. Cohen and H. H. Hobbs (2008). "Genetic variation in PNPLA3 confers susceptibility to nonalcoholic fatty liver disease." Nat Genet **40**(12): 1461-1465.
Rundek, T., H. Gardener, Q. Xu, R. B. Goldberg, C. B. Wright, B. Boden-Albala, N. Disla, M. C. Paik, M. S. Elkind and R. L. Sacco (2010). "Insulin resistance and risk of ischemic stroke among nondiabetic individuals from the northern Manhattan study." Arch Neurol **67**(10): 1195-1200.
Rung, J., S. Cauchi, A. Albrechtsen, L. Shen, G. Rocheleau, C. Cavalcanti-Proenca, F. Bacot, B. Balkau, A. Belisle, K. Borch-Johnsen, G. Charpentier, C. Dina, E. Durand, P. Elliott, S. Hadjadj, M. R. Jarvelin, J. Laitinen, T. Lauritzen, M. Marre, A. Mazur, D. Meyre, A. Montpetit, C. Pisinger, B. Posner, P. Poulsen, A. Pouta, M. Prentki, R. Ribel-Madsen, A. Ruokonen, A. Sandbaek, D. Serre, J. Tichet, M. Vaxillaire, J. F. Wojtaszewski, A. Vaag, T. Hansen, C. Polychronakos, O. Pedersen, P. Froguel and R. Sladek (2009). "Genetic variant near IRS1 is associated with type 2 diabetes, insulin resistance and hyperinsulinemia." Nat Genet **41**(10): 1110-1115.
Ryoo, J. H., Y. J. Suh, H. C. Shin, Y. K. Cho, J. M. Choi and S. K. Park (2014). "Clinical association between non-alcoholic fatty liver disease and the development of hypertension." J Gastroenterol Hepatol **29**(11): 1926-1931.
Sacerdote, C., F. Ricceri, O. Rolandsson, I. Baldi, M. D. Chirlaque, E. Feskens, B. Bendinelli, E. Ardanaz, L. Arriola, B. Balkau, M. Bergmann, J. W. Beulens, H. Boeing, F. Clavel-Chapelon, F. Crowe, B. de Lauzon-Guillain, N. Forouhi, P. W. Franks, V. Gallo, C. Gonzalez, J. Halkjaer, A. K. Illner, R. Kaaks, T. Key, K. T. Khaw, C. Navarro, P. M. Nilsson, S. O. Dal Ton, K. Overvad, V. Pala, D. Palli, S. Panico, S. Polidoro, J. R. Quiros, I. Romieu, M. J. Sanchez, N. Slimani, I. Sluijs, A. Spijkerman, B. Teucher, A. Tjonneland, R. Tumino, A. D. van der, A. C. Vergnaud, P. Wennberg, S. Sharp, C. Langenberg, E. Riboli, P. Vineis and N. Wareham (2012). "Lower educational level is a predictor of incident type 2 diabetes in European countries: the EPIC-InterAct study." Int J Epidemiol **41**(4): 1162-1173.

Santoro, N., G. Cirillo, A. Amato, C. Luongo, P. Raimondo, A. D'Aniello, L. Perrone and E. Miraglia del Giudice (2006). "Insulin gene variable number of tandem repeats (INS VNTR) genotype and metabolic syndrome in childhood obesity." J Clin Endocrinol Metab **91**(11): 4641-4644.
Sanyal, A. J., C. Campbell-Sargent, F. Mirshahi, W. B. Rizzo, M. J. Contos, R. K. Sterling, V. A. Luketic, M. L. Shiffman and J. N. Clore (2001). "Nonalcoholic steatohepatitis: association of insulin resistance and mitochondrial abnormalities." Gastroenterology **120**(5): 1183-1192.
Sato, H., S. Sato, R. Kawasaki, T. Yamamoto, T. Yamashita and H. Yamashita (2005). "Retinal Cell Damage Due to Oxidative Stress in Diabetic Retinopathy." ARVO Annual Meeting Abstract. Investigative Ophthalmology & Visual Science May 2005 **46**: 443.
Schafer, J. L. (1999). "Multiple imputation: a primer." Stat Methods Med Res **8**(1): 3-15.
Schmilovitz-Weiss, H., R. Gingold-Belfer, D. Boltin, Y. Beloosesky, J. Meyerovitch, R. Tor, N. Issa, A. Grossman, N. Koren-Morag and A. Weiss (2018). "Risk of mortality and level of serum alanine aminotransferase among community-dwelling elderly in Israel." Eur J Gastroenterol Hepatol **30**(12): 1428-1433.
Schneider, A., G. Hommel and M. Blettner (2010). "Linear regression analysis: part 14 of a series on evaluation of scientific publications." Dtsch Arztebl Int **107**(44): 776-782.
Schwimmer, J. B., M. A. Celedon, J. E. Lavine, R. Salem, N. Campbell, N. J. Schork, M. Shiehmorteza, T. Yokoo, A. Chavez, M. S. Middleton and C. B. Sirlin (2009). "Heritability of nonalcoholic fatty liver disease." Gastroenterology **136**(5): 1585-1592.
Shah, A. G., A. Lydecker, K. Murray, B. N. Tetri, M. J. Contos and A. J. Sanyal (2009). "Comparison of noninvasive markers of fibrosis in patients with nonalcoholic fatty liver disease." Clin Gastroenterol Hepatol **7**(10): 1104-1112.
Shah, C. P. and C. Chen (2011). "Review of therapeutic advances in diabetic retinopathy." Ther Adv Endocrinol Metab **2**(1): 39-53.
Shi, H., N. Moustaid-Moussa, W. O. Wilkison and M. B. Zemel (1999). "Role of the sulfonylurea receptor in regulating human adipocyte metabolism." Faseb j **13**(13): 1833-1838.
Shibata, H., T. Ichikawa, K. Nakao, H. Miyaaki, S. Takeshita, M. Akiyama, M. Fujimoto, S. Miuma, S. Kanda, H. Yamasaki and K. Eguchi (2008). "A high glucose condition sensitizes human hepatocytes to hydrogen peroxide-induced cell death." Mol Med Rep **1**(3): 379-385.
Shim, J. J., J. W. Kim, C. H. Oh, Y. R. Lee, J. S. Lee, S. Y. Park, B. H. Kim and I. H. Oh (2018). "Serum alanine aminotransferase level and liver-related mortality in patients with chronic hepatitis B: A large national cohort study." Liver Int **38**(10): 1751-1759.
Simon, P. H., M. P. Sylvestre, J. Tremblay and P. Hamet (2016). "Key Considerations and Methods in the Study of Gene-Environment Interactions." Am J Hypertens **29**(8): 891-899.
Smagris, E., S. BasuRay, J. Li, Y. Huang, K. M. Lai, J. Gromada, J. C. Cohen and H. H. Hobbs (2015). "Pnpla3I148M knockin mice accumulate PNPLA3 on lipid droplets and develop hepatic steatosis." Hepatology **61**(1): 108-118.
Sofer, E., M. Boaz, Z. Matas, M. Mashavi and M. Shargorodsky (2011). "Treatment with insulin sensitizer metformin improves arterial properties, metabolic parameters, and liver function in patients with nonalcoholic fatty liver disease: a randomized, placebo-controlled trial." Metabolism **60**(9): 1278-1284.
Sookoian, S., G. O. Castano, A. L. Burgueno, T. F. Gianotti, M. S. Rosselli and C. J. Pirola (2009). "A nonsynonymous gene variant in the adiponutrin gene is associated with nonalcoholic fatty liver disease severity." J Lipid Res **50**(10): 2111-2116.
Sookoian, S., G. O. Castano, R. Scian, T. Fernandez Gianotti, H. Dopazo, C. Rohr, G. Gaj, J. San Martino, I. Sevic, D. Flichman and C. J. Pirola (2016). "Serum aminotransferases in nonalcoholic fatty liver disease are a signature of liver metabolic perturbations at the amino acid and Krebs cycle level." Am J Clin Nutr **103**(2): 422-434.

Sorbi, D., J. Boynton and K. D. Lindor (1999). "The ratio of aspartate aminotransferase to alanine aminotransferase: potential value in differentiating nonalcoholic steatohepatitis from alcoholic liver disease." Am J Gastroenterol **94**(4): 1018-1022.
Stratigopoulos, G., S. L. Padilla, C. A. LeDuc, E. Watson, A. T. Hattersley, M. I. McCarthy, L. M. Zeltser, W. K. Chung and R. L. Leibel (2008). "Regulation of Fto/Ftm gene expression in mice and humans." Am J Physiol Regul Integr Comp Physiol **294**(4): R1185-1196.
Talmud, P. J., J. A. Cooper, R. W. Morris, F. Dudbridge, T. Shah, J. Engmann, C. Dale, J. White, S. McLachlan, D. Zabaneh, A. Wong, K. K. Ong, T. Gaunt, M. V. Holmes, D. A. Lawlor, M. Richards, R. Hardy, D. Kuh, N. Wareham, C. Langenberg, Y. Ben-Shlomo, S. G. Wannamethee, M. W. Strachan, M. Kumari, J. C. Whittaker, F. Drenos, M. Kivimaki, A. D. Hingorani, J. F. Price and S. E. Humphries (2015). "Sixty-five common genetic variants and prediction of type 2 diabetes." Diabetes **64**(5): 1830-1840.
Tan, H. L., S. M. Zain, R. Mohamed, S. Rampal, K. F. Chin, R. C. Basu, P. L. Cheah, S. Mahadeva and Z. Mohamed (2014). "Association of glucokinase regulatory gene polymorphisms with risk and severity of non-alcoholic fatty liver disease: an interaction study with adiponutrin gene." J Gastroenterol **49**(6): 1056-1064.
Targher, G., L. Bertolini, R. Padovani, F. Poli, L. Scala, R. Tessari, L. Zenari and G. Falezza (2006). "Increased prevalence of cardiovascular disease in Type 2 diabetic patients with non-alcoholic fatty liver disease." Diabet Med **23**(4): 403-409.
Taylor, R. (1990). "Interpretation of the Correlation Coefficient: A Basic Review." Journal of Diagnostic Medical Sonography **6**: 35-39.
Teuscher, A., H. Schnell and P. W. Wilson (1988). "Incidence of diabetic retinopathy and relationship to baseline plasma glucose and blood pressure." Diabetes Care **11**(3): 246-251.
Tibshirani, R. (1996). "Regression Shrinkage and Selection via the Lasso." Journal of the Royal Statistical Society. Series B (Methodological) **58**(1): 267-288.
Townsend, S. A. and P. N. Newsome (2016). "Non-alcoholic fatty liver disease in 2016." Br Med Bull **119**(1): 143-156.
Verboven, K., K. Wouters, K. Gaens, D. Hansen, M. Bijnen, S. Wetzels, C. D. Stehouwer, G. H. Goossens, C. G. Schalkwijk, E. E. Blaak and J. W. Jocken (2018). "Abdominal subcutaneous and visceral adipocyte size, lipolysis and inflammation relate to insulin resistance in male obese humans." Sci Rep **8**(1): 4677.
Vespasiani-Gentilucci, U., C. Dell'Unto, A. De Vincentis, A. Baiocchini, M. Delle Monache, R. Cecere, A. M. Pellicelli, V. Giannelli, S. Carotti, G. Galati, P. Gallo, F. Valentini, F. Del Nonno, D. Rosati, S. Morini, R. Antonelli-Incalzi and A. Picardi (2018). "Combining Genetic Variants to Improve Risk Prediction for NAFLD and Its Progression to Cirrhosis: A Proof of Concept Study." Can J Gastroenterol Hepatol **2018**: 7564835.
Vilhjalmsson, B. J., J. Yang, H. K. Finucane, A. Gusev, S. Lindstrom, S. Ripke, G. Genovese, P. R. Loh, G. Bhatia, R. Do, T. Hayeck, H. H. Won, S. Kathiresan, M. Pato, C. Pato, R. Tamimi, E. Stahl, N. Zaitlen, B. Pasaniuc, G. Belbin, E. E. Kenny, M. H. Schierup, P. De Jager, N. A. Patsopoulos, S. McCarroll, M. Daly, S. Purcell, D. Chasman, B. Neale, M. Goddard, P. M. Visscher, P. Kraft, N. Patterson and A. L. Price (2015). "Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores." Am J Hum Genet **97**(4): 576-592.
Vozarova, B., N. Stefan, R. S. Lindsay, A. Saremi, R. E. Pratley, C. Bogardus and P. A. Tataranni (2002). "High alanine aminotransferase is associated with decreased hepatic insulin sensitivity and predicts the development of type 2 diabetes." Diabetes **51**(6): 1889-1895.
Wang, S., F. B. Hu and J. Dupuis (2016). "Type 2 Diabetes Prediction." The Genetics of Type 2 Diabetes and Related Traits: Biology, Physiology and Translation, Editors: Florez, Jose C (Ed). 425-440.

Warner, R. M. (2013). "Applied statistics: From bivariate through multivariate techniques (2nd ed.)." Sage Publications, Inc **Chapter 4**.
Warrier, V. and S. Baron-Cohen (2018). "Genetic contribution to 'theory of mind' in adolescence." Sci Rep **8**(1): 3465.
Wei, Z., W. Wang, J. Bradfield, J. Li, C. Cardinale, E. Frackelton, C. Kim, F. Mentch, K. Van Steen, P. M. Visscher, R. N. Baldassano and H. Hakonarson (2013). "Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease." Am J Hum Genet **92**(6): 1008-1012.
Wells, M. M., Z. Li, B. Addeman, C. A. McKenzie, A. Mujoomdar, M. Beaton and J. Bird (2016). "Computed Tomography Measurement of Hepatic Steatosis: Prevalence of Hepatic Steatosis in a Canadian Population." Can J Gastroenterol Hepatol **2016**: 4930987.
Wheeler, E., A. Leong, C. T. Liu, M. F. Hivert, R. J. Strawbridge, C. Podmore, M. Li, J. Yao, X. Sim, J. Hong, A. Y. Chu, W. Zhang, X. Wang, P. Chen, N. M. Maruthur, B. C. Porneala, S. J. Sharp, Y. Jia, E. K. Kabagambe, L. C. Chang, W. M. Chen, C. E. Elks, D. S. Evans, Q. Fan, F. Giulianini, M. J. Go, J. J. Hottenga, Y. Hu, A. U. Jackson, S. Kanoni, Y. J. Kim, M. E. Kleber, C. Ladenvall, C. Lecoeur, S. H. Lim, Y. Lu, A. Mahajan, C. Marzi, M. A. Nalls, P. Navarro, I. M. Nolte, L. M. Rose, D. V. Rybin, S. Sanna, Y. Shi, D. O. Stram, F. Takeuchi, S. P. Tan, P. J. van der Most, J. V. Van Vliet-Ostaptchouk, A. Wong, L. Yengo, W. Zhao, A. Goel, M. T. Martinez Larrad, D. Radke, P. Salo, T. Tanaka, E. P. A. van Iperen, G. Abecasis, S. Afaq, B. Z. Alizadeh, A. G. Bertoni, A. Bonnefond, Y. Bottcher, E. P. Bottinger, H. Campbell, O. D. Carlson, C. H. Chen, Y. S. Cho, W. T. Garvey, C. Gieger, M. O. Goodarzi, H. Grallert, A. Hamsten, C. A. Hartman, C. Herder, C. A. Hsiung, J. Huang, M. Igase, M. Isono, T. Katsuya, C. C. Khor, W. Kiess, K. Kohara, P. Kovacs, J. Lee, W. J. Lee, B. Lehne, H. Li, J. Liu, S. Lobbens, J. Luan, V. Lyssenko, T. Meitinger, T. Miki, I. Miljkovic, S. Moon, A. Mulas, G. Muller, M. Muller-Nurasyid, R. Nagaraja, M. Nauck, J. S. Pankow, O. Polasek, I. Prokopenko, P. S. Ramos, L. Rasmussen-Torvik, W. Rathmann, S. S. Rich, N. R. Robertson, M. Roden, R. Roussel, I. Rudan, R. A. Scott, W. R. Scott, B. Sennblad, D. S. Siscovick, K. Strauch, L. Sun, M. Swertz, S. M. Tajuddin, K. D. Taylor, Y. Y. Teo, Y. C. Tham, A. Tonjes, N. J. Wareham, G. Willemsen, T. Wilsgaard, A. D. Hingorani, J. Egan, L. Ferrucci, G. K. Hovingh, A. Jula, M. Kivimaki, M. Kumari, I. Njolstad, C. N. A. Palmer, M. Serrano Rios, M. Stumvoll, H. Watkins, T. Aung, M. Bluher, M. Boehnke, D. I. Boomsma, S. R. Bornstein, J. C. Chambers, D. I. Chasman, Y. I. Chen, Y. T. Chen, C. Y. Cheng, F. Cucca, E. J. C. de Geus, P. Deloukas, M. K. Evans, M. Fornage, Y. Friedlander, P. Froguel, L. Groop, M. D. Gross, T. B. Harris, C. Hayward, C. K. Heng, E. Ingelsson, N. Kato, B. J. Kim, W. P. Koh, J. S. Kooner, A. Korner, D. Kuh, J. Kuusisto, M. Laakso, X. Lin, Y. Liu, R. J. F. Loos, P. K. E. Magnusson, W. Marz, M. I. McCarthy, A. J. Oldehinkel, K. K. Ong, N. L. Pedersen, M. A. Pereira, A. Peters, P. M. Ridker, C. Sabanayagam, M. Sale, D. Saleheen, J. Saltevo, P. E. Schwarz, W. H. H. Sheu, H. Snieder, T. D. Spector, Y. Tabara, J. Tuomilehto, R. M. van Dam, J. G. Wilson, J. F. Wilson, B. H. R. Wolffenbuttel, T. Y. Wong, J. Y. Wu, J. M. Yuan, A. B. Zonderman, N. Soranzo, X. Guo, D. J. Roberts, J. C. Florez, R. Sladek, J. Dupuis, A. P. Morris, E. S. Tai, E. Selvin, J. I. Rotter, C. Langenberg, I. Barroso and J. B. Meigs (2017). "Impact of common genetic determinants of Hemoglobin A1c on type 2 diabetes risk and diagnosis in ancestrally diverse populations: A transethnic genome-wide meta-analysis." PLoS Med **14**(9): e1002383.
Wild, S. H., J. J. Walker, J. R. Morling, D. A. McAllister, H. M. Colhoun, B. Farran, S. McGurnaghan, R. McCrimmon, S. H. Read, N. Sattar and C. D. Byrne (2018). "Cardiovascular Disease, Cancer, and Mortality Among People With Type 2 Diabetes and Alcoholic or Nonalcoholic Fatty Liver Disease Hospital Admission." Diabetes Care **41**(2): 341-347.
Willer, C. J., E. M. Schmidt, S. Sengupta, G. M. Peloso, S. Gustafsson, S. Kanoni, A. Ganna, J. Chen, M. L. Buchkovich, S. Mora, J. S. Beckmann, J. L. Bragg-Gresham, H. Y. Chang, A. Demirkan, H. M. Den Hertog, R. Do, L. A. Donnelly, G. B. Ehret, T. Esko, M. F. Feitosa, T. Ferreira, K. Fischer, P.

Fontanillas, R. M. Fraser, D. F. Freitag, D. Gurdasani, K. Heikkila, E. Hypponen, A. Isaacs, A. U. Jackson, A. Johansson, T. Johnson, M. Kaakinen, J. Kettunen, M. E. Kleber, X. Li, J. Luan, L. P. Lyytikainen, P. K. E. Magnusson, M. Mangino, E. Mihailov, M. E. Montasser, M. Muller-Nurasyid, I. M. Nolte, J. R. O'Connell, C. D. Palmer, M. Perola, A. K. Petersen, S. Sanna, R. Saxena, S. K. Service, S. Shah, D. Shungin, C. Sidore, C. Song, R. J. Strawbridge, I. Surakka, T. Tanaka, T. M. Teslovich, G. Thorleifsson, E. G. Van den Herik, B. F. Voight, K. A. Volcik, L. L. Waite, A. Wong, Y. Wu, W. Zhang, D. Absher, G. Asiki, I. Barroso, L. F. Been, J. L. Bolton, L. L. Bonnycastle, P. Brambilla, M. S. Burnett, G. Cesana, M. Dimitriou, A. S. F. Doney, A. Doring, P. Elliott, S. E. Epstein, G. Ingi Eyjolfsson, B. Gigante, M. O. Goodarzi, H. Grallert, M. L. Gravito, C. J. Groves, G. Hallmans, A. L. Hartikainen, C. Hayward, D. Hernandez, A. A. Hicks, H. Holm, Y. J. Hung, T. Illig, M. R. Jones, P. Kaleebu, J. J. P. Kastelein, K. T. Khaw, E. Kim, N. Klopp, P. Komulainen, M. Kumari, C. Langenberg, T. Lehtimaki, S. Y. Lin, J. Lindstrom, R. J. F. Loos, F. Mach, W. L. McArdle, C. Meisinger, B. D. Mitchell, G. Muller, R. Nagaraja, N. Narisu, T. V. M. Nieminen, R. N. Nsubuga, I. Olafsson, K. K. Ong, A. Palotie, T. Papamarkou, C. Pomilla, A. Pouta, D. J. Rader, M. P. Reilly, P. M. Ridker, F. Rivadeneira, I. Rudan, A. Ruokonen, N. Samani, H. Scharnagl, J. Seeley, K. Silander, A. Stancakova, K. Stirrups, A. J. Swift, L. Tiret, A. G. Uitterlinden, L. J. van Pelt, S. Vedantam, N. Wainwright, C. Wijmenga, S. H. Wild, G. Willemsen, T. Wilsgaard, J. F. Wilson, E. H. Young, J. H. Zhao, L. S. Adair, D. Arveiler, T. L. Assimes, S. Bandinelli, F. Bennett, M. Bochud, B. O. Boehm, D. I. Boomsma, I. B. Borecki, S. R. Bornstein, P. Bovet, M. Burnier, H. Campbell, A. Chakravarti, J. C. Chambers, Y. I. Chen, F. S. Collins, R. S. Cooper, J. Danesh, G. Dedoussis, U. de Faire, A. B. Feranil, J. Ferrieres, L. Ferrucci, N. B. Freimer, C. Gieger, L. C. Groop, V. Gudnason, U. Gyllensten, A. Hamsten, T. B. Harris, A. Hingorani, J. N. Hirschhorn, A. Hofman, G. K. Hovingh, C. A. Hsiung, S. E. Humphries, S. C. Hunt, K. Hveem, C. Iribarren, M. R. Jarvelin, A. Jula, M. Kahonen, J. Kaprio, A. Kesaniemi, M. Kivimaki, J. S. Kooner, P. J. Koudstaal, R. M. Krauss, D. Kuh, J. Kuusisto, K. O. Kyvik, M. Laakso, T. A. Lakka, L. Lind, C. M. Lindgren, N. G. Martin, W. Marz, M. I. McCarthy, C. A. McKenzie, P. Meneton, A. Metspalu, L. Moilanen, A. D. Morris, P. B. Munroe, I. Njolstad, N. L. Pedersen, C. Power, P. P. Pramstaller, J. F. Price, B. M. Psaty, T. Quertermous, R. Rauramaa, D. Saleheen, V. Salomaa, D. K. Sanghera, J. Saramies, P. E. H. Schwarz, W. H. Sheu, A. R. Shuldiner, A. Siegbahn, T. D. Spector, K. Stefansson, D. P. Strachan, B. O. Tayo, E. Tremoli, J. Tuomilehto, M. Uusitupa, C. M. van Duijn, P. Vollenweider, L. Wallentin, N. J. Wareham, J. B. Whitfield, B. H. R. Wolffenbuttel, J. M. Ordovas, E. Boerwinkle, C. N. A. Palmer, U. Thorsteinsdottir, D. I. Chasman, J. I. Rotter, P. W. Franks, S. Ripatti, L. A. Cupples, M. S. Sandhu, S. S. Rich, M. Boehnke, P. Deloukas, S. Kathiresan, K. L. Mohlke, E. Ingelsson and G. R. Abecasis (2013). "Discovery and refinement of loci associated with lipid levels." Nat Genet **45**(11): 1274-1283.

Winsor, E. (1988). "Mendelian genetics." Can Fam Physician **34**: 859-862.

Wray, N. R., J. Yang, M. E. Goddard and P. M. Visscher (2010). "The genetic interpretation of area under the ROC curve in genomic profiling." PLoS Genet **6**(2): e1000864.

Wuttke, M., Y. Li, M. Li, K. B. Sieber, M. F. Feitosa, M. Gorski, A. Tin, L. Wang, A. Y. Chu, A. Hoppmann, H. Kirsten, A. Giri, J. F. Chai, G. Sveinbjornsson, B. O. Tayo, T. Nutile, C. Fuchsberger, J. Marten, M. Cocca, S. Ghasemi, Y. Xu, K. Horn, D. Noce, P. J. van der Most, S. Sedaghat, Z. Yu, M. Akiyama, S. Afaq, T. S. Ahluwalia, P. Almgren, N. Amin, J. Arnlov, S. J. L. Bakker, N. Bansal, D. Baptista, S. Bergmann, M. L. Biggs, G. Biino, M. Boehnke, E. Boerwinkle, M. Boissel, E. P. Bottinger, T. S. Boutin, H. Brenner, M. Brumat, R. Burkhardt, A. S. Butterworth, E. Campana, A. Campbell, H. Campbell, M. Canouil, R. J. Carroll, E. Catamo, J. C. Chambers, M. L. Chee, M. L. Chee, X. Chen, C. Y. Cheng, Y. Cheng, K. Christensen, R. Cifkova, M. Ciullo, M. P. Concas, J. P. Cook, J. Coresh, T. Corre, C. F. Sala, D. Cusi, J. Danesh, E. W. Daw, M. H. de Borst, A. De Grandi, R. de Mutsert, A. P. J. de Vries, F. Degenhardt, G. Delgado, A. Demirkan, E. Di Angelantonio, K. Dittrich, J. Divers, R. Dorajoo, K. U. Eckardt, G. Ehret, P. Elliott, K. Endlich, M. K. Evans, J. F. Felix, V. H. X. Foo, O. H.

Franco, A. Franke, B. I. Freedman, S. Freitag-Wolf, Y. Friedlander, P. Froguel, R. T. Gansevoort, H. Gao, P. Gasparini, J. M. Gaziano, V. Giedraitis, C. Gieger, G. Girotto, F. Giulianini, M. Gogele, S. D. Gordon, D. F. Gudbjartsson, V. Gudnason, T. Haller, P. Hamet, T. B. Harris, C. A. Hartman, C. Hayward, J. N. Hellwege, C. K. Heng, A. A. Hicks, E. Hofer, W. Huang, N. Hutri-Kahonen, S. J. Hwang, M. A. Ikram, O. S. Indridason, E. Ingelsson, M. Ising, V. W. V. Jaddoe, J. Jakobsdottir, J. B. Jonas, P. K. Joshi, N. S. Josyula, B. Jung, M. Kahonen, Y. Kamatani, C. M. Kammerer, M. Kanai, M. Kastarinen, S. M. Kerr, C. C. Khor, W. Kiess, M. E. Kleber, W. Koenig, J. S. Kooner, A. Korner, P. Kovacs, A. T. Kraja, A. Krajcoviechova, H. Kramer, B. K. Kramer, F. Kronenberg, M. Kubo, B. Kuhnel, M. Kuokkanen, J. Kuusisto, M. La Bianca, M. Laakso, L. A. Lange, C. D. Langefeld, J. J. Lee, B. Lehne, T. Lehtimaki, W. Lieb, S. C. Lim, L. Lind, C. M. Lindgren, J. Liu, J. Liu, M. Loeffler, R. J. F. Loos, S. Lucae, M. A. Lukas, L. P. Lyytikainen, R. Magi, P. K. E. Magnusson, A. Mahajan, N. G. Martin, J. Martins, W. Marz, D. Mascalzoni, K. Matsuda, C. Meisinger, T. Meitinger, O. Melander, A. Metspalu, E. K. Mikaelsdottir, Y. Milaneschi, K. Miliku, P. P. Mishra, K. L. Mohlke, N. Mononen, G. W. Montgomery, D. O. Mook-Kanamori, J. C. Mychaleckyj, G. N. Nadkarni, M. A. Nalls, M. Nauck, K. Nikus, B. Ning, I. M. Nolte, R. Noordam, J. O'Connell, M. L. O'Donoghue, I. Olafsson, A. J. Oldehinkel, M. Orho-Melander, W. H. Ouwehand, S. Padmanabhan, N. D. Palmer, R. Palsson, B. Penninx, T. Perls, M. Perola, M. Pirastu, N. Pirastu, G. Pistis, A. I. Podgornaia, O. Polasek, B. Ponte, D. J. Porteous, T. Poulain, P. P. Pramstaller, M. H. Preuss, B. P. Prins, M. A. Province, T. J. Rabelink, L. M. Raffield, O. T. Raitakari, D. F. Reilly, R. Rettig, M. Rheinberger, K. M. Rice, P. M. Ridker, F. Rivadeneira, F. Rizzi, D. J. Roberts, A. Robino, P. Rossing, I. Rudan, R. Rueedi, D. Ruggiero, K. A. Ryan, Y. Saba, C. Sabanayagam, V. Salomaa, E. Salvi, K. U. Saum, H. Schmidt, R. Schmidt, B. Schottker, C. A. Schulz, N. Schupf, C. M. Shaffer, Y. Shi, A. V. Smith, B. H. Smith, N. Soranzo, C. N. Spracklen, K. Strauch, H. M. Stringham, M. Stumvoll, P. O. Svensson, S. Szymczak, E. S. Tai, S. M. Tajuddin, N. Y. Q. Tan, K. D. Taylor, A. Teren, Y. C. Tham, J. Thiery, C. H. L. Thio, H. Thomsen, G. Thorleifsson, D. Toniolo, A. Tonjes, J. Tremblay, I. Tzoulaki, A. G. Uitterlinden, S. Vaccargiu, R. M. van Dam, P. van der Harst, C. M. van Duijn, D. R. Velez Edward, N. Verweij, S. Vogelezang, U. Volker, P. Vollenweider, G. Waeber, M. Waldenberger, L. Wallentin, Y. X. Wang, C. Wang, D. M. Waterworth, W. Bin Wei, H. White, J. B. Whitfield, S. H. Wild, J. F. Wilson, M. K. Wojczynski, C. Wong, T. Y. Wong, L. Xu, Q. Yang, M. Yasuda, L. M. Yerges-Armstrong, W. Zhang, A. B. Zonderman, J. I. Rotter, M. Bochud, B. M. Psaty, V. Vitart, J. G. Wilson, A. Dehghan, A. Parsa, D. I. Chasman, K. Ho, A. P. Morris, O. Devuyst, S. Akilesh, S. A. Pendergrass, X. Sim, C. A. Boger, Y. Okada, T. L. Edwards, H. Snieder, K. Stefansson, A. M. Hung, I. M. Heid, M. Scholz, A. Teumer, A. Kottgen and C. Pattaro (2019). "A catalog of genetic loci associated with kidney function from analyses of a million individuals." Nat Genet **51**(6): 957-972.

Xu, R., A. Tao, S. Zhang, Y. Deng and G. Chen (2015). "Association between patatin-like phospholipase domain containing 3 gene (PNPLA3) polymorphisms and nonalcoholic fatty liver disease: a HuGE review and meta-analysis." Sci Rep **5**: 9284.

Xu, Z., C. Wu and W. Pan (2017). "Imaging-wide association study: Integrating imaging endophenotypes in GWAS." Neuroimage **159**: 159-169.

Yasuda, T., H. Takeshita, E. Nakazato, T. Nakajima, Y. Nakashima, S. Mori, K. Mogi and K. Kishi (2000). "The molecular basis for genetic polymorphism of human deoxyribonuclease II (DNase II): a single nucleotide substitution in the promoter region of human DNase II changes the promoter activity." FEBS Lett **467**(2-3): 231-234.

Yoo, J., S. Lee, K. Kim, S. Yoo, E. Sung and J. Yim (2008). "Relationship between insulin resistance and serum alanine aminotransferase as a surrogate of NAFLD (nonalcoholic fatty liver disease) in obese Korean children." Diabetes Res Clin Pract **81**(3): 321-326.

Younossi, Z. M., A. B. Koenig, D. Abdelatif, Y. Fazel, L. Henry and M. Wymer (2016). "Global epidemiology of nonalcoholic fatty liver disease-Meta-analytic assessment of prevalence, incidence, and outcomes." Hepatology **64**(1): 73-84.
Younossi, Z. M., M. Otgonsuren, L. Henry, C. Venkatesan, A. Mishra, M. Erario and S. Hunt (2015). "Association of nonalcoholic fatty liver disease (NAFLD) with hepatocellular carcinoma (HCC) in the United States from 2004 to 2009." Hepatology **62**(6): 1723-1730.
Younossi, Z. M., M. Stepanova, F. Negro, S. Hallaji, Y. Younossi, B. Lam and M. Srishord (2012). "Nonalcoholic fatty liver disease in lean individuals in the United States." Medicine (Baltimore) **91**(6): 319-327.
Yu, C., W. Yao and X. Bai (2014). "Robust Linear Regression: A Review and Comparison." arXiv.
Yun, K. E., C. Y. Shin, Y. S. Yoon and H. S. Park (2009). "Elevated alanine aminotransferase levels predict mortality from cardiovascular disease and diabetes in Koreans." Atherosclerosis **205**(2): 533-537.
Zeggini, E., L. J. Scott, R. Saxena, B. F. Voight, J. L. Marchini, T. Hu, P. I. de Bakker, G. R. Abecasis, P. Almgren, G. Andersen, K. Ardlie, K. B. Bostrom, R. N. Bergman, L. L. Bonnycastle, K. Borch-Johnsen, N. P. Burtt, H. Chen, P. S. Chines, M. J. Daly, P. Deodhar, C. J. Ding, A. S. Doney, W. L. Duren, K. S. Elliott, M. R. Erdos, T. M. Frayling, R. M. Freathy, L. Gianniny, H. Grallert, N. Grarup, C. J. Groves, C. Guiducci, T. Hansen, C. Herder, G. A. Hitman, T. E. Hughes, B. Isomaa, A. U. Jackson, T. Jorgensen, A. Kong, K. Kubalanza, F. G. Kuruvilla, J. Kuusisto, C. Langenberg, H. Lango, T. Lauritzen, Y. Li, C. M. Lindgren, V. Lyssenko, A. F. Marvelle, C. Meisinger, K. Midthjell, K. L. Mohlke, M. A. Morken, A. D. Morris, N. Narisu, P. Nilsson, K. R. Owen, C. N. Palmer, F. Payne, J. R. Perry, E. Pettersen, C. Platou, I. Prokopenko, L. Qi, L. Qin, N. W. Rayner, M. Rees, J. J. Roix, A. Sandbaek, B. Shields, M. Sjogren, V. Steinthorsdottir, H. M. Stringham, A. J. Swift, G. Thorleifsson, U. Thorsteinsdottir, N. J. Timpson, T. Tuomi, J. Tuomilehto, M. Walker, R. M. Watanabe, M. N. Weedon, C. J. Willer, T. Illig, K. Hveem, F. B. Hu, M. Laakso, K. Stefansson, O. Pedersen, N. J. Wareham, I. Barroso, A. T. Hattersley, F. S. Collins, L. Groop, M. I. McCarthy, M. Boehnke and D. Altshuler (2008). "Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes." Nat Genet **40**(5): 638-645.
Zhao, Y., D. Jhamb, L. Shu, D. Arneson, D. K. Rajpal and X. Yang (2019). "Multi-omics integration reveals molecular networks and regulators of psoriasis." BMC Syst Biol **13**(1): 8.
Zhou, X. J., Y. Y. Qi, P. Hou, J. C. Lv, S. F. Shi, L. J. Liu, N. Zhao and H. Zhang (2014). "Cumulative effects of variants identified by genome-wide association studies in IgA nephropathy." Sci Rep **4**: 4904.
Zhou, Y., Y. Liang, K. Li, X. Bai, G. Chen, Z. Xing and J. Xiao (2012). "The phenotypic distribution of quantitative traits in a wild mouse F1 population." Mamm Genome **23**(3-4): 232-240.
Zou, H. and T. Hastie (2005). "Regularization and Variable Selection via the Elastic Net." Journal of the Royal Statistical Society. Series B (Statistical Methodology) **67**(2): 301-320.
Zoungas, S., J. Chalmers, B. Neal, L. Billot, Q. Li, Y. Hirakawa, H. Arima, H. Monaghan, R. Joshi, S. Colagiuri, M. E. Cooper, P. Glasziou, D. Grobbee, P. Hamet, S. Harrap, S. Heller, L. Lisheng, G. Mancia, M. Marre, D. R. Matthews, C. E. Mogensen, V. Perkovic, N. Poulter, A. Rodgers, B. Williams, S. MacMahon, A. Patel and M. Woodward (2014). "Follow-up of blood-pressure lowering and glucose control in type 2 diabetes." N Engl J Med **371**(15): 1392-1406.
Zoungas, S., B. E. de Galan, T. Ninomiya, D. Grobbee, P. Hamet, S. Heller, S. MacMahon, M. Marre, B. Neal, A. Patel, M. Woodward, J. Chalmers, A. Cass, P. Glasziou, S. Harrap, L. Lisheng, G. Mancia, A. Pillai, N. Poulter, V. Perkovic and F. Travert (2009). "Combined effects of routine blood pressure lowering and intensive glucose control on macrovascular and microvascular outcomes in patients with type 2 diabetes: New results from the ADVANCE trial." Diabetes Care **32**(11): 2068-2074.

Zuk, O., S. F. Schaffner, K. Samocha, R. Do, E. Hechter, S. Kathiresan, M. J. Daly, B. M. Neale, S. R. Sunyaev and E. S. Lander (2014). "Searching for missing heritability: designing rare variant association studies." Proc Natl Acad Sci U S A **111**(4): E455-464.