

## THÈSE

pour obtenir le grade de :

DOCTEUR DE L'UNIVERSITÉ DE LILLE

dans la spécialité

« BIOSTATISTIQUE »

par

*MICKAËL CANOUIL*

### ***Développement et Application de Méthodologies Statistiques pour Études Multi-Omiques dans le Diabète de Type 2***

***Au-delà de l'Ère des Études d'Association Pangénomiques***

Thèse soutenue le 29 Septembre 2017 devant le jury composé de :

Pr. PHILIPPE	FROGUEL	( <i>Directeur de Thèse</i> )
Dr. GHISLAIN	ROCHELEAU	( <i>Co-Directeur de Thèse</i> )
Dr. HÉLÈNE	JACQMIN-GADDA	( <i>Rapporteur</i> )
Dr. MARIA	MARTINEZ	( <i>Rapporteur</i> )
Dr. GUILMETTE	MAROT-BRIEND	( <i>Examinateur</i> )



*“The best thing about being a statistician is that you get to play in everyone's backyard”*

—John Tukey



---

## **Table des matières**

---

<b>Remerciements</b>	v
<b>Résumé</b>	vii
<b>Abstract</b>	ix
<b>Introduction</b>	1
1    Préceptes . . . . .	1
1.1    Génome . . . . .	1
1.2    Transcriptome . . . . .	4
1.3    Épigénome et Méthylome . . . . .	6
1.4    Phénotype . . . . .	7
2    Le diabète de type 2 . . . . .	8
2.1    Définition et chiffres du diabète . . . . .	8
2.2    Physiopathologie du diabète de type 2 . . . . .	9
2.3    La maladie du foie non alcoolique . . . . .	11
2.4    La génétique et l'épigénétique du diabète de type 2 . . . . .	12
3    Méthodes statistiques : des données à la biologie . . . . .	16
3.1    La statistique génétique . . . . .	16
3.2    Recueil et prétraitement des données . . . . .	17
3.3    Analyses omique et multi-omique . . . . .	32
<b>Objectifs &amp; Plan</b>	51
<b>1    Variants génétiques associés à la trajectoire de la glycémie à jeun et à l'incidence du diabète de type 2 : Une approche par modèle joint</b>	55
1    Introduction . . . . .	55
1.1    Contexte/objectifs . . . . .	55
1.2    Méthodes . . . . .	55

1.3	Résultats . . . . .	56
1.4	Conclusion . . . . .	57
2	Article . . . . .	57
<b>2</b>	<b>L'Expression et l'Évaluation Fonctionnelle des Gènes de Susceptibilité au Diabète de Type 2 Identifient Quatre Nouveaux Gènes Contribuant à la Sécrétion d'Insuline Humaine</b>	<b>83</b>
1	Introduction . . . . .	84
1.1	Contexte/objectifs . . . . .	84
1.2	Méthodes . . . . .	84
1.3	Résultats . . . . .	86
1.4	Conclusion . . . . .	87
2	Article . . . . .	87
<b>3</b>	<b>La Surexpression Hépatique de PDGF-AA Affaiblit la Signalisation de l'Insuline dans le Diabète</b>	<b>89</b>
1	Introduction . . . . .	90
1.1	Contexte/objectifs . . . . .	90
1.2	Méthodes . . . . .	90
1.3	Résultats . . . . .	92
1.4	Conclusion . . . . .	93
2	Article . . . . .	94
<b>4</b>	<b>L'Exposition à Faible Dose aux Bisphénols A, F et S des Adipocytes Primaires Humains Modifie les Profils d'ARN Codant et Non-Codant</b>	<b>129</b>
1	Introduction . . . . .	130
1.1	Contexte/objectifs . . . . .	130
1.2	Méthodes . . . . .	130
1.3	Résultats . . . . .	131
1.4	Conclusion . . . . .	132
1.5	Note . . . . .	132
2	Article . . . . .	133
<b>Conclusion</b>		<b>135</b>
1	Développement méthodologique . . . . .	135
2	Support méthodologique . . . . .	136
3	Multi-omiques & Perspectives . . . . .	137

<i>Contents</i>	iii
<b>Liste des communications scientifiques</b>	<b>139</b>
1    Communications en lien avec la thèse . . . . .	139
1.1    Conférences . . . . .	139
1.2    Articles publiés dans des revues internationales à comité de lecture . . . . .	140
2    Autres communications dans le domaine de la génétique . . . . .	140
<b>Bibliographie</b>	<b>141</b>



---

## **Remerciements**

---

Je tiens en premier lieu à remercier Ghislain ROCHELEAU et Philippe FROGUEL de m'avoir donné l'opportunité de travailler sur ce thème de recherche et d'avoir supervisé celui-ci pendant trois ans.

Merci aux membres de ce jury : Alain DUHAMEL, Hélène JACQMIN-GADDA, Guillemette MAROT-BRIEND, Maria MARTINEZ, Cristian PREDA d'avoir accepté de juger mon travail.

Je remercie Loïc YENGO et Ghislain ROCHELEAU pour le soutien scientifique (et pas uniquement) qu'ils m'ont apporté avant le début et pendant les trois années de cette thèse, et pour avoir cru en moi non seulement au niveau de la thèse, mais plus généralement au niveau professionnel.

Merci également à Loïc YENGO et à Philippe FROGUEL de m'avoir accueilli dans cette équipe de recherche source de défi m'ayant permis de développer mes compétences en biostatistique et en management.

À cela s'ajoute des remerciements tout particuliers aux membres passés et présents de l'équipe de biostatistique : Boris S., Cécile L., Dorothée T., Ghislain R., Lijiao N., Loïc Y., Marie V. et Mathilde B. m'ayant permis d'améliorer mon travail, au gré de nombreux échanges, et de me concentrer sur ma thèse, en particulier sur cette dernière année.

Je tiens à remercier les différents chercheurs avec lesquels j'ai pu collaborer durant ce travail, notamment Amar ABDERRAHMANI, Amélie BONNEFOND et Odile POULAIN-CODEFROY.

Enfin, je remercie toutes les personnes qui ont contribué à ce travail en particulier celles intervenues aussi bien sur la scène que dans les coulisses (“pause-café” et “afterwork”) : Aurélie D., Cindy A., Clément D., David L. G., Franck D.G., Iandry R., Julie M., Julien D., Loïc D. S., Marie F., Marie V., Marine C., Mélanie H., Morgane B., Stefan G. et Véronique D.

Merci aux cinémas de la ville de Lille qui m'ont accueilli plus de 500 fois dans leurs salles obscures au cours des trois dernières années.



---

## Résumé

---

Les études d'association pangénomiques (GWAS) ont permis l'identification de plusieurs dizaines de gènes et de polymorphismes nucléotidiques (SNPs) contribuant au risque de diabète de type 2 (DT2). Plus généralement, les GWAS ont permis d'identifier des milliers de SNPs contribuant à des maladies complexes chez l'Homme. Cependant, la caractérisation fonctionnelle et les mécanismes biologiques impliquant ces SNPs et ces gènes restent en grande partie à explorer. En effet, les conséquences de ces polymorphismes sont complexes et peu connues. Une conséquence directe est l'altération de la protéine codée par un gène, voire une extinction complète de la transcription du gène (p. ex. via l'introduction d'un codon stop dans la séquence). Par ailleurs, ces polymorphismes peuvent avoir un rôle de régulation dans l'expression des gènes, par exemple, en perturbant la liaison de facteurs de transcription et d'enzymes impliqués dans la méthylation de l'ADN. Malgré des associations fortes des SNPs identifiés, ils ne peuvent expliquer la totalité de l'hérabilité du DT2, suggérant par le fait même des mécanismes d'interactions entre les différentes couches que représentent la génomique, la transcriptomique et l'épigénomique.

Le changement de paradigme en statistique génétique et la disponibilité de données transcriptomiques et épigénomiques sont responsables de l'évolution du domaine, passant des analyses d'associations à des analyses transversales de type multi-omique, et permettant de fournir des éléments de réponse sur l'aspect fonctionnel des SNPs ou des gènes impliqués, et dans certains cas, permettant d'évaluer le lien causal de ces variants sur la pathologie. Les développements et applications méthodologiques proposés dans cette thèse sont variés, allant d'une approche similaire aux GWAS, mettant à profit les données longitudinales disponibles dans certaines cohortes (p. ex. D.E.S.I.R.), au moyen d'un modèle joint; de la caractérisation fonctionnelle de gènes candidats, identifiés par GWAS, dans la sécrétion d'insuline par une étude transcriptomique multi tissu et dans un modèle cellulaire; de l'identification d'un nouveau gène candidat (PDGFA) impliqué dans la dérégulation de la voie de l'insuline dans le DT2 via des mécanismes épigénétiques et transcriptomiques; et enfin de la caractérisation de l'effet sur le transcriptome de deux substituts du bisphénol A dans un modèle d'adipocyte primaire.

L'augmentation des connaissances des processus biologiques dans lesquels sont impliqués les SNPs et gènes identifiés par GWAS pourrait permettre l'élaboration de stratégies diagnostiques plus efficaces, ainsi que l'identification de cibles thérapeutiques pour le traitement du DT2 et des complications associées (p. ex. insulino-résistance, NAFLD, cancer, etc.). Plus généralement, ces études multi-omiques ouvrent la voie à l'approche émergente que représente la médecine de précision, permettant le traitement et la prévention des pathologies tout en prenant en compte ce qui fait la spécificité d'un individu, à savoir son génome et son environnement, tous deux interagissant sur son transcriptome et son épigénome.

*Mots-clés :* Biostatistique; Génétique; Epigénétique; Transcriptomique; Diabète de type 2



---

## **Abstract**

---

Genome-wide association studies (GWAS) have resulted in the identification of several dozen of genes and single nucleotide polymorphisms (SNPs) contributing to type 2 diabetes (T2D). More generally, GWAS have identified thousands of SNPs contributing to complex diseases in human. However, the functional characterization and biological mechanisms involving these SNPs and genes remain to be explored. Indeed, the consequences of these polymorphisms are complex and little known. One direct consequence is the alteration of the protein encoded by a gene, or even a complete transcriptional gene silencing (e.g. codon stop in the sequence). Furthermore, these polymorphisms may have a regulatory role in gene expression, for example, by interfering with the binding of transcription factors and enzymes involved in DNA methylation. Despite the strong associations of SNPs identified, they cannot explain the full heritability of T2D, hence suggesting interaction mechanisms between the different layers of omics, such as genomics, transcriptomics and epigenomics.

The paradigm shift in statistical genetics and the availability of transcriptomic and epigenomic data are responsible for the evolution of the discipline, moving from association studies to multi-omics studies, and providing insights on the functional aspect of the SNPs or genes involved, and in some cases allowing to evaluate the causal link of these variants on the pathology. The methodological developments and their applications proposed in this thesis are various, ranging from a similar approach to GWAS, by leveraging the longitudinal data available in some cohorts (e.g. D.E.S.I.R.), using a joint model approach; to the functional characterisation of candidate genes, identified by GWAS, in insulin secretion by a multi tissue transcriptomic study and by study in a cell model; to the identification of a new candidate gene (PDGFA) involved in the deregulation of the insulin's pathway in T2D through epigenetic and transcriptomic mechanisms; and finally, to the characterisation of the effect on the transcriptome of two substitutes of bisphenol A in a primary adipocyte model.

The increase of knowledge in biological processes involving SNPs and genes identified by GWAS could enable the development of more effective diagnostic strategies, and the identification of therapeutic targets for the treatment of T2D and its associated complications (e.g., insulin resistance, NAFLD, cancer, etc.). More generally, these multi-omics studies pave the way for the emerging approach of precision medicine, allowing the treatment and prevention of pathologies while accounting for what makes the specificity of an individual, namely his genome and his environment, both interacting on his transcriptome and his epigenome.

*Keywords :* Biostatistics; Genetics; Epigenetics; Transcriptomics; Type 2 Diabetes



---

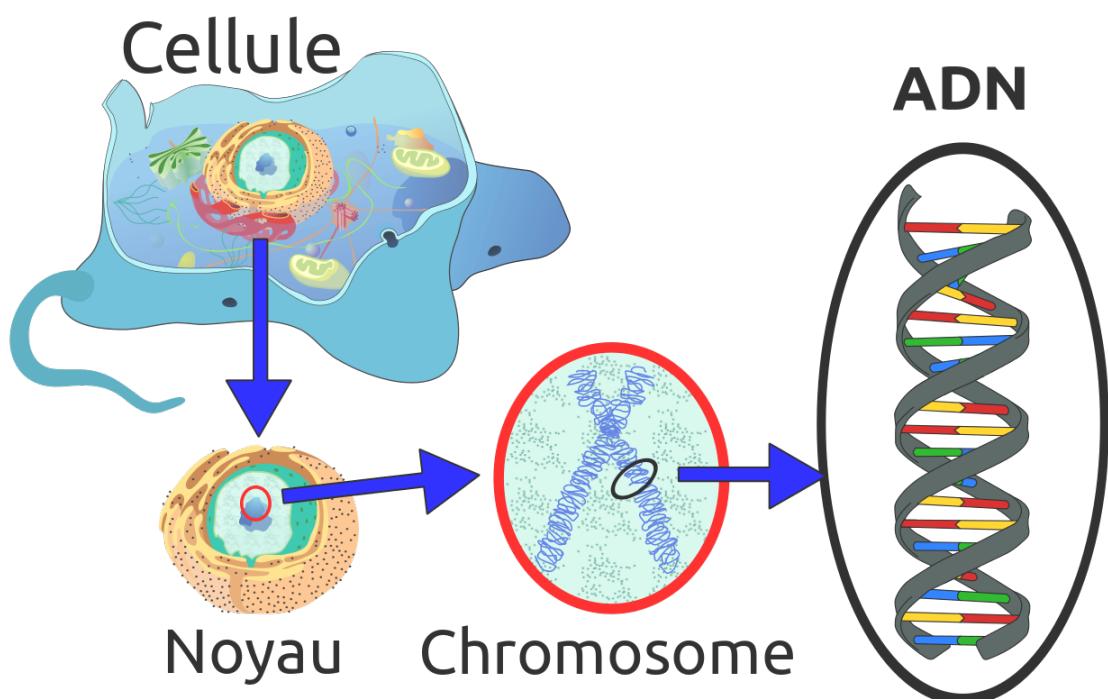
## **Introduction**

---

### **1 Préceptes**

Dans un premier temps, nous proposons de revenir sur quelques notions et définitions, qui pourront au besoin faire l'objet de simplifications.

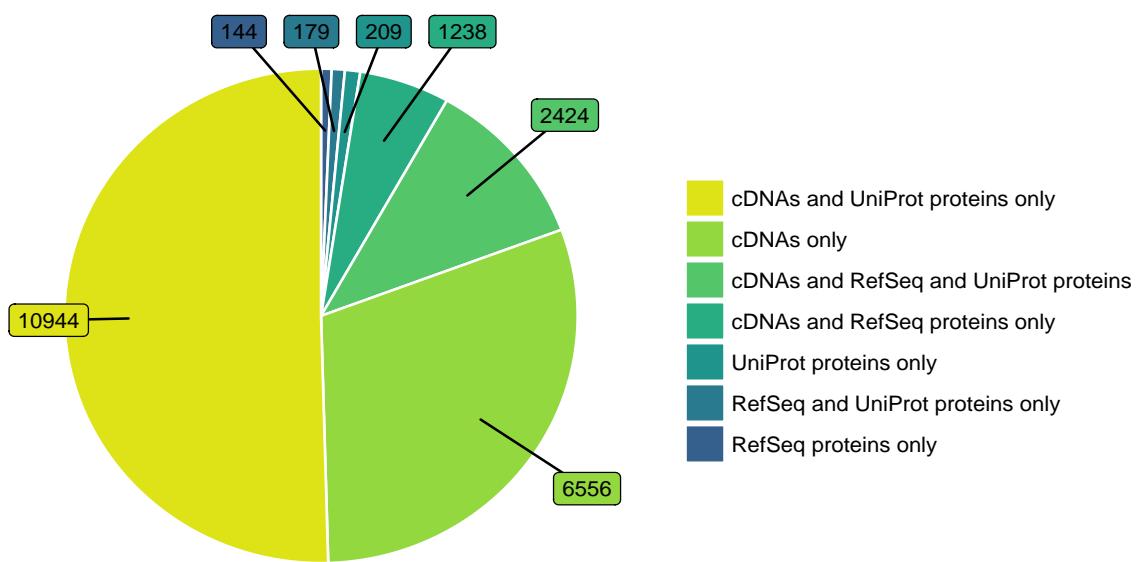
#### **1.1 Gé nome**



**FIGURE1.** Localisation de l'ADN dans une cellule eucaryote.

Le patrimoine génétique (génome) d'un individu est l'ensemble de l'information génétique présente sous forme de chromosomes dans le noyau des cellules eucaryotes (Figure 1). L'Homme dispose de 22 paires de chromosomes, appelés autosomes, et d'une paire de chromosomes sexuels ou gonosomes (notés XX chez la femme, et XY chez l'homme). Ces chromosomes constituent la forme condensée de deux molécules, appelées

brins, d'Acide DésoxyriboNucléique (ADN), et composées de la répétition de quatre nucléotides (ou bases nucléotidiques) : A, C, G et T, respectivement pour adénosine, cytosine, guanine et thymine. La position de l'une de ces bases, donnée en paire de base (pb), est appelée locus (loci, au pluriel). Il est à noter qu'un locus peut également désigner une région de plusieurs dizaines de paires de bases, voire d'un gène entier. Ces quatre nucléotides sont la base de l'information génétique et sont complémentaires pour une même paire : A est couplé à T, tandis que C est couplé à G. Cette complémentarité permet aux deux brins d'ADN de se lier l'un à l'autre via une liaison hydrogène et de former une structure en double-hélice. L'association de cette double-hélice avec des complexes protéiques, tels que les histones, permet à l'ADN d'être présent dans deux états de condensation différents : l'euchromatine, un état décondensé, et l'hétérochromatine, un état condensé où les histones sont très rapprochées les unes des autres. Ces deux états de condensation de la molécule d'ADN auront un effet sur les mécanismes de transcription de l'ADN. La transcription est un mécanisme de lecture de l'ADN et d'écriture d'une partie de l'information génétique se trouvant au niveau d'un gène sous une forme dérivée : l'Acide RiboNucléique (ARN). Les gènes sont le résultat de l'arrangement en séquence des 3,5 milliards de paires de bases du génome humain (Ensembl version 89, mai 2017). Cependant, l'ensemble de l'ADN n'est pas codant. Chez l'Homme, il existe environ 20 000 gènes codant pour des protéines, répartis de façon discontinue sur l'ensemble du génome (Ensembl version 89, mai 2017).

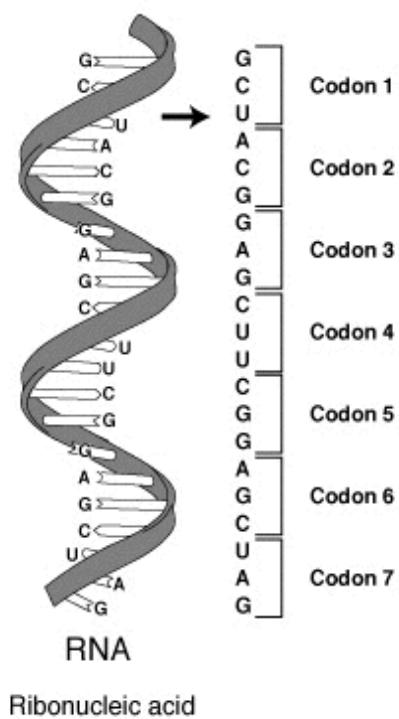


**FIGURE 2.** Diagramme des gènes répertoriés sur la base Ensembl.

D'une cellule à une autre dans un même organisme, le génome est le même. Cependant, il existe des disparités

entre le génome de deux individus d'une même espèce, et d'autant plus entre deux espèces. Ainsi, deux individus d'une même espèce vont partager pour plusieurs loci les mêmes allèles, mais pourront présenter des variations appelées polymorphismes. Pour un même locus, le génotype donne l'allèle présent sur chacun des deux chromosomes d'une même paire, et s'écrit sous la forme d'un couple d'allèles : AA, AB, ou BB, où A et B désignent les bases nucléotidiques (c.-à-d. A, C, G ou T).

Au sein d'une population, la variabilité génétique est engendrée principalement par l'intermédiaire de deux phénomènes : la mutation ou la recombinaison. La mutation est un mécanisme introduisant un polymorphisme, c'est-à-dire par l'introduction d'une nouvelle version d'un allèle au sein de la séquence, soit par l'ajout, la suppression ou l'insertion d'un ou plusieurs nucléotides, provoquant ainsi des changements dans la séquence d'ADN. Ces changements peuvent être classés en différentes catégories selon leur conséquence sur la synthèse de la protéine. Ainsi, une mutation est dite "silencieuse" ou "synonyme", lorsque l'acide aminé n'est pas changé.



**FIGURE 3.** Brin d'ARN et codons.

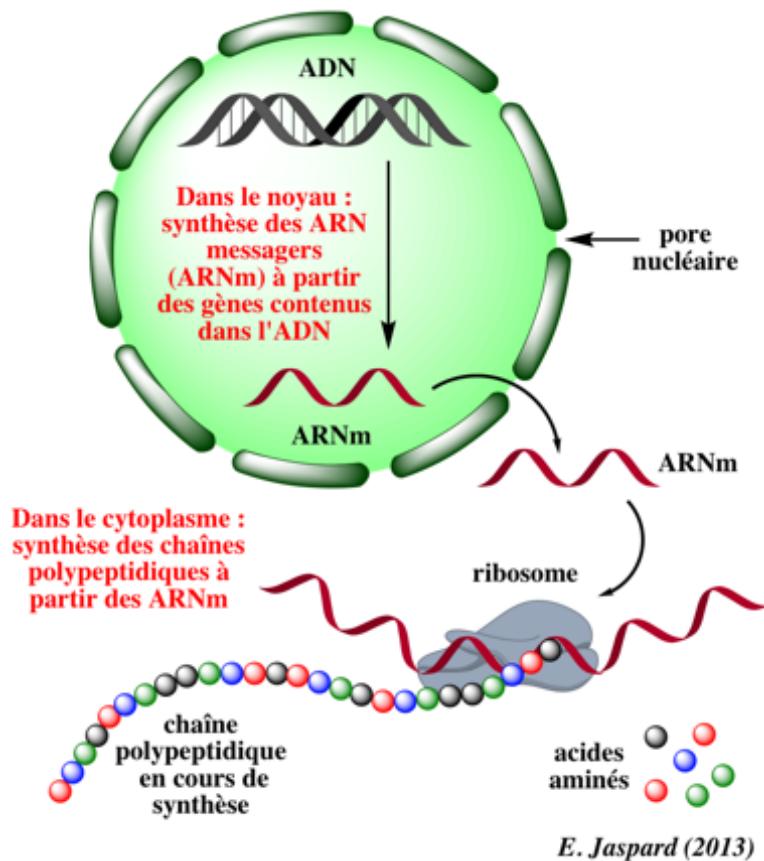
Un acide aminé est le résultat de la traduction, depuis un ARN messager, d'une séquence de trois bases nucléotidiques nommée "codon" (Figure 3). Il existe 64 codons ( $4^3$  combinaisons de bases), correspondant à 22 acides aminés uniques, ce qui permet à plusieurs codons d'être traduits en un même acide aminé. Quand la mutation engendre l'apparition d'un codon stop, arrêtant par le fait même la synthèse de la protéine avant la fin de la séquence d'ARN, elle est alors appelée "non-sens".

La recombinaison est un mécanisme se produisant lors de la méiose, c'est-à-dire lors du processus de formation

des gamètes, où les chromosomes homologues (c.-à-d. les chromosomes d'une même paire) se chevauchent, et peuvent alors échanger une partie de l'ADN les constituant. La recombinaison n'affecte pas l'ensemble du chromosome de façon homogène. En effet, les événements de recombinaison sont plus fréquents avec l'éloignement du centromère, c'est-à-dire à la position de jointure des chromosomes d'une même paire. On parle de déséquilibre de liaison lorsque la probabilité d'observer un allèle à un locus A n'est pas indépendante de celle d'observer un allèle à un locus B, autrement dit, lorsque la probabilité d'observer un certain couple d'allèle n'est pas égale au produit au produit des probabilités d'observé chaque allèle individuellement.

Pour la suite, nous nous intéresserons principalement aux variants génétiques polymorphiques au niveau d'une seule base nucléotidique, les SNPs ("Single Nucleotide Polymorphisms"), et omettrons les insertions/délétions (INDEL), ou encore les variations du nombre de copies ou CNVs ("Copy Number Variations").

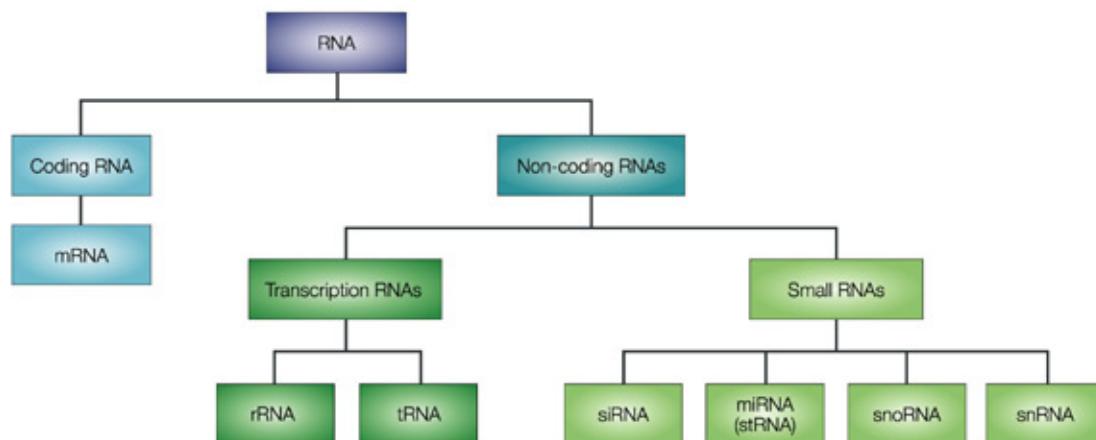
## 1.2 Transcriptome



**FIGURE 4.** Schéma simplifié de la synthèse des protéines chez les eucaryotes.

La transcriptomique étudie les ARN formés lors de la transcription d'un gène dans le noyau. La transcription est une étape indispensable pour la synthèse de protéines permettant de faire transiter l'information contenue dans l'ADN nucléaire vers le cytoplasme sous la forme d'ARN, où se trouve le matériel nécessaire à la traduction

en protéine (c.-à-d. les ribosomes et les acides aminés) (Figure 4). La transcription de l'ADN en ARN est réalisée par l'enzyme ARN polymérase. Il existe plusieurs classes d'ARN, dont la plus abondante est la classe des ARN messagers (mRNA).



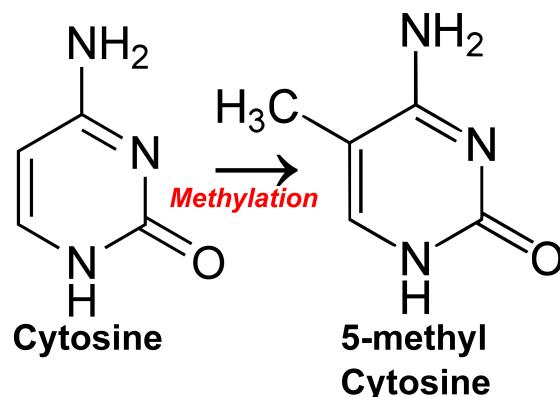
**FIGURE 5.** Classification des différents types d'ARN.

À cette classe s'ajoutent les ARN de transfert (tRNA), qui apportent les acides aminés nécessaires à la traduction des mRNA en protéines, et les ARN ribosomaux (rRNA), qui constituent les complexes protéiques que sont les ribosomes, ainsi que des ARN mesurant moins de 200 nucléotides, soit les petits ARN (“small RNA”), les miRNA (“microRNA”), les snoRNA (“small nucleolar RNA”), les siRNA (“small interfering RNA”), les piRNA (“piwi interfering RNA”), etc. (Figure 5). Ces derniers participent à divers mécanismes métaboliques, notamment la régulation de l'expression des gènes [Ambros, 2004; Bartel, 2004].

L'expression des gènes est mesurée directement par les mRNA, qui représentent les ARN codant pour les protéines. Il est à noter que le nombre de protéines pouvant être synthétisé est supérieur au nombre de gènes. En effet, un gène se compose de plusieurs exons (parties codantes) et d'introns (parties non-codantes), et lors de la transcription, l'épissage du gène permet la création d'une molécule mRNA ne comportant que les parties codantes. Cette phase d'épissage peut être “alternative”, c'est-à-dire, que pourront être conservées, lors de la synthèse de mRNA, différentes combinaisons d'exons aboutissant à la synthèse de plusieurs mRNA ou transcrits, qui seront alors exportés en dehors du noyau pour être ensuite synthétisés en protéines dans le cytoplasme de la cellule. Ainsi, un gène peut produire plusieurs protéines différentes selon les besoins et la fonction de la cellule et du tissu. Le transcriptome est défini comme l'ensemble des ARN présents à un instant donné dans un tissu ou type cellulaire spécifique, et nécessaires à la synthèse protéique et à sa régulation en partie via les petits ARN.

### 1.3 Épigénome et Méthylome

Les mécanismes de régulation de l'expression des gènes sont nombreux. L'un de ces mécanismes passe par des marques épigénétiques modifiant la structure et la conformation de l'ADN, rendant de ce fait plus simple ou plus difficile selon les cas, la fixation des facteurs de transcription, et plus généralement de la machinerie cellulaire sur l'ADN. Il existe principalement deux types de modifications épigénétiques : la méthylation de l'ADN et les modifications d'histones. L'ensemble de ces modifications constitue l'épigénome d'un individu. Le méthylome constitue un sous-ensemble de l'épigénome regroupant uniquement les marques de méthylation. Nous nous concentrerons sur le méthylome impliqué principalement dans la régulation des gènes, la maintenance et la formation de la chromatine, constituant de ce fait un élément important de la régulation du transcriptome.

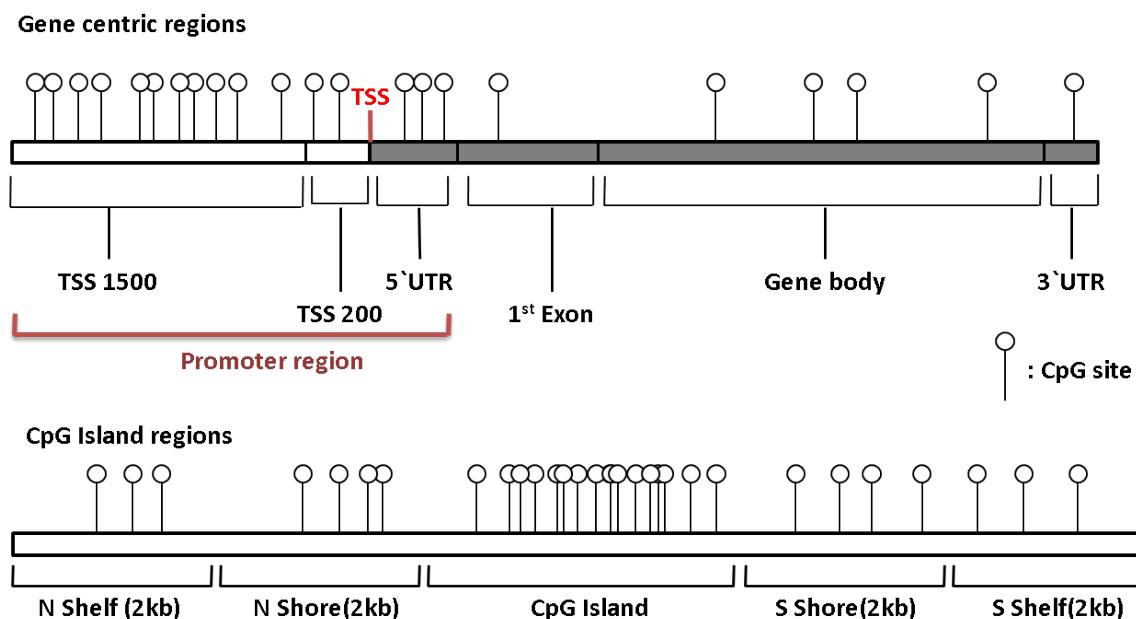


**FIGURE 6.** Formule structurelle de la méthylation de la cytosine.

La méthylation de l'ADN est l'ajout d'un groupement méthyl via une enzyme de la famille des ADN méthyl-transférases (DNMT). Cette enzyme va catalyser l'ajout d'un groupement méthyl sur le carbone en position 5 d'une cytosine (5mC), cytosine généralement suivie (5' vers 3') d'une guanine, formant ainsi un groupement CpG (cytosine-phosphate-guanine) (Figure 6). Ces groupements, ou sites CpG, ne sont pas les seuls groupements de dinucléotides pouvant faire l'objet d'une méthylation CpHpG (H = A, T, ou C) [Lister et al., 2009]. Cependant, ces marques ne sont pas les plus fréquentes chez les organismes eucaryotes [Bird, 1980].

Les sites CpG représentent une faible fraction du génome ( $\simeq 1\%$ ) et sont distribués de façon hétérogène sur l'ensemble du génome en raison d'une déamination spontanée, au cours du temps, des 5mC en thymine [Bird, 1980; Cooper and Krawczak, 1989].

Les sites CpG peuvent être regroupés en îlots, appelés îlots CpG [Deaton and Bird, 2011] (Figure 7). Ces îlots CpG sont des régions enrichies en dinucléotide guanine-cytosine (CC) et sont généralement non-méthylés. Un îlot CpG est défini comme une séquence d'une longueur de 200 à 1 000 paires de bases comportant plus de 50 % de GC et un ratio  $\frac{CpG_{observé}}{CpG_{attendu}} \geq 60\%$  [Antequera, 2003; Gardiner-Garden and Frommer, 1987]. Ces îlots CpG sont localisés dans les régions promotrices et exoniques pour 40 à 60 % d'entre eux [Larsen et al., 1992; Saxonov



**FIGURE 7.** Schéma représentant la localisation des sites CpG sur le génome.

et al., 2006], traduisant l'importance du processus de méthylation dans la régulation de l'expression des gènes. Les sites et îlots CpG sont principalement démethylés lorsqu'ils sont situés dans des régions promotrices (p. ex. site d'initiation de la transcription, en amont d'un gène) et des exons, et sont méthylés lorsqu'ils sont localisés dans le corps des gènes [Ball et al., 2009; Hellman and Chess, 2007; Jones, 1999]. En effet, une hypométhylation observée au niveau des régions promotrices de la transcription est généralement inversement corrélée avec l'expression des gènes [Schultz et al., 2015; Wagner et al., 2014], mais pas de façon systématique [Moarii et al., 2015], rendant complexe l'étude de l'épigénome conjointement avec le transcriptome. À cela s'ajoute l'impact potentiel des mutations survenant au niveau de la cytosine d'un site CpG. Ces polymorphismes sont appelés CpG-SNPs et permettent de mettre en évidence des mécanismes d'interaction entre génome et épigénome [Dayeh et al., 2013; Zhi et al., 2013].

#### 1.4 Phénotype

Le phénotype correspond à l'ensemble des caractères ou traits physiques observables chez un individu (p. ex. la taille, la couleur des yeux, le statut diabétique, etc.). Le phénotype est le résultat à la fois du génotype et des facteurs environnementaux, comme le mode de vie, le régime alimentaire ou l'activité physique par exemple. Un biomarqueur est une marque phénotypique particulière qui est la mesure d'un composé (p. ex. protéine, métabolite, etc.) présent dans le corps d'un individu et servant, en médecine, à des fins diagnostiques d'une pathologie.

**TABLEAU 1.** Critères glycémiques de l'organisation mondiale de la santé (OMS), définissant les statuts insulinorésistant et diabétique. (OGTT : test de tolérance au glucose par voie orale).

	Glycémie à jeun	Glycémie 2h après OGTT
Normoglycémique	< 6,1 mmol/L	< 7,7 mmol/L
Intolérant au glucose	6,1 - 6,9 mmol/L	7,7 - 11 mmol/L
Diabétique	> 7 mmol/L	> 11,1 mmol/L

**TABLEAU 2.** Critères glycémiques de l'association américaine pour le diabète (ADA), définissant les statuts insulinorésistant et diabétique.

	Glycémie à jeun	HbA1c
Normoglycémique	< 5,6 mmol/L	> 5,7 %
Intolérant au glucose	5,6 - 6,9 mmol/L	5,7 - 6,4 %
Diabétique	> 7 mmol/L	> 6,5 %

## 2 Le diabète de type 2

### 2.1 Définition et chiffres du diabète

Le diabète est défini par une hyperglycémie. Dès 1999, l'organisation Mondiale de la Santé (OMS) [World Health Organization, 1999] préconise deux mesures de glycémie pour diagnostiquer le diabète : la glycémie à jeun et la glycémie mesurée deux heures après un test de tolérance au glucose par voie orale (OGTT). Dans la pratique, le diagnostic du diabète s'effectue, dans certains cas, via la mesure d'hémoglobine glyquée (HbA1c).

C'est notamment le cas aux États-Unis, où la mesure d'HbA1c fait partie des critères de définition proposés par l'Association Américaine pour le Diabète (ADA). L'HbA1c est utilisée ici pour sa propriété à refléter l'évolution de la glycémie sur les trois derniers mois, ce qui correspond à la durée de vie moyenne d'un d'élément.

Même si l'ADA et l'OMS s'accordent pour définir le diabète à partir d'une glycémie à jeun supérieure à 7,0 mmol/L, les critères pour définir une glycémie normale diffèrent entre les deux organisations, avec un seuil de glycémie inférieure à 5,6 mmol/L pour l'ADA et 6,1 mmol/L pour l'OMS (Tableau 1 et 2). Cette phase, ainsi définie pour une glycémie entre 5,6 ou 6,1 mmol/L et 7 mmol/L, peut être transitoire vers un diabète, et est parfois appelée "prédiabète". Les patients diagnostiqués comme intolérants au glucose font l'objet d'une prise en charge préventive consistant principalement en une modification du comportement alimentaire et plus généralement des habitudes de vie.

L'hyperglycémie chronique peut, lorsqu'elle n'est pas traitée, provoquer des complications au niveau cardio-vasculaire, rénale, oculaire, et dans certains cas, conduire à une amputation d'un ou des membres inférieurs.

Selon le dernier rapport de l'OMS, plus de 400 millions de personnes en 2014 vivaient avec le diabète, contre seulement 108 millions en 1980 selon les estimations mondiales [Roglic and World Health Organization, 2016]. Depuis 1980, la prévalence du diabète est passée de 4,7 à 8,5 % chez la population adulte dans le monde.

En France, selon les derniers rapports de l'Institut de Veille Sanitaire (InVS) [Mandereau-Bruno et al., 2014; Ricci et al., 2010], la prévalence du diabète traité est passée de 4,6 % en 2012 à 5 % en 2015. En 2006-2007, la prévalence de l'intolérance au glucose représentait 5,6 %, ce qui en fait un véritable enjeu de santé publique.

Il existe 4 formes de diabète, le diabète de type 1, le diabète de type 2, le diabète gestationnel et les diabètes monogéniques.

Le diabète de type 1 est le diabète dit insulinodépendant et nécessite des injections régulières d'insuline. Ce diabète se développe généralement chez un individu jeune qui perd rapidement sa capacité à réguler sa glycémie, suite à une réaction auto-immune contre les cellules  $\beta$  du pancréas (cellules sécrétrices de l'insuline).

Le diabète de type 2 est parfois appelé diabète de l'adulte ou diabète non-insulinodépendant, par opposition au diabète de type 1. Il se caractérise principalement par un défaut du métabolisme de l'insuline d'un ou plusieurs organes. Le diabète de type 2 représente plus de 90 % des diabètes dans le monde [Lyssenko and Laakso, 2013].

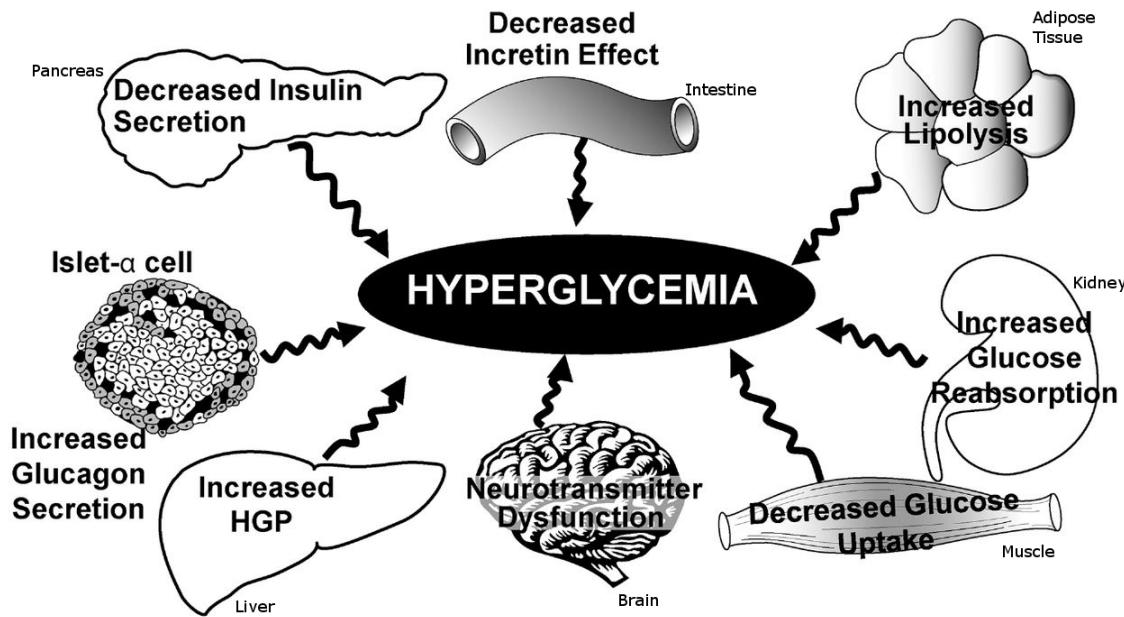
Parce que les symptômes du diabète de type 2 sont moins marqués que ceux du diabète de type 1, le diabète de type 2 est souvent diagnostiquée tardivement, et notamment suite aux complications résultantes de celui-ci.

Le diabète existe également sous une troisième forme, dit gestationnel. Ce diabète survient chez la femme durant la grossesse, aux environs de la 24ème semaine d'aménorrhée, et présente un facteur de risque accru du développement ultérieur d'un diabète de type 2, à la fois chez la mère et chez l'enfant [Case et al., 2006; Roglic and World Health Organization, 2016].

## 2.2 Physiopathologie du diabète de type 2

Le diabète de type 2 serait la conséquence d'une production insuffisante d'insuline en réponse à une demande accrue de l'organisme provenant d'une résistance à l'insuline [World Health Organization, 1999; World Health Organization and International Diabetes Federation, 2006]. Le diabète de type 2 est une pathologie complexe dont l'origine génétique est multiple et passe notamment par des interactions avec l'environnement. De nombreux traits ont été identifiés comme facteurs de risque, tels le sexe, l'âge ou encore l'Indice de Masse Corporel (IMC), mais aussi l'ethnicité (p. ex., population des indiens Pima [Diamond, 2003; Knowler et al., 1993]) et le manque d'activité physique font également partie de ces facteurs de risques [Lyssenko et al., 2008; Mykkänen et al., 1993; Noble et al., 2011].

L'hyperglycémie, dans le cadre du diabète de type 2, implique trois mécanismes principaux : i) une augmentation de la sécrétion de glucose par le foie (néoglucogenèse) ; ii) une diminution de l'entrée et donc du métabolisme



**FIGURE 8.** Tissus et organes impliqués dans l'hyperglycémie et le diabète de type 2 (HGP : “hepatic glucose production”, production hépatique de glucose).

du glucose dans les organes périphériques, comme le muscle (insulinorésistance); iii) une altération de la sécrétion d'insuline par le pancréas ou une altération de l'insuline elle-même (Figure 8).

L'insulinémie et la glycémie à jeun permettent de mesurer l'insulinorésistance sur la base des indices HOMA (“HOMeostasis Model Assessment”) [Matthews et al., 1985]. Trois indices HOMA ont été proposés, avec  $I$ , l'insulinémie à jeun (en mU/L) et avec  $G$ , la glycémie à jeun (en mmol/L) :

- l'HOMA-IR reflétant l'insulinorésistance, dont la valeur normale est établie à 1, et qui augmente avec la gravité de l'insulinorésistance ( $\text{HOMA-IR} = \frac{(I \times G)}{22,5}$ );
- l'HOMA-B augmentant avec l'altération de la fonction des cellules  $\beta$ , dont la valeur normale est fixée à 100 % ( $\text{HOMA-B} = \frac{(20 \times I)}{(G-3,5)}$ );
- l'HOMA-S étant l'inverse de l'HOMA-IR et représentant la sensibilité à l'insuline ( $\text{HOMA-S} = \frac{1}{\text{HOMA-IR}} \times 100$ ).

Depuis la formulation de ces indices, Levy et al. [1998] ont proposé un nouveau modèle HOMA2, permettant la prise en compte de la variabilité de la tolérance au glucose du foie et des tissus périphériques, de la contribution à l'homéostasie de la proinsuline circulante, ainsi qu'une meilleure définition de la courbe de sécrétion d'insuline en réponse au glucose, notamment pour des concentrations en glucose supérieures à 10 mmol/L. Les indices basés sur le modèle HOMA2 sont exprimés en pourcentage, contrairement aux indices HOMA.

L'insulinorésistance correspond à la perte de sensibilité des récepteurs cibles de l'insuline, ne permettant plus à

l'insuline de se fixer et bloquant ainsi l'entrée du glucose dans la cellule. Cette insulinorésistance engendre une sécrétion plus importante d'insuline par les cellules  $\beta$  pour compenser ce manque d'efficacité (partielle ou totale), entraînant à plus ou moins long terme la défaillance de ces cellules, et dans le même temps induisant une hyperglycémie, signe précurseur d'un potentiel diabète. L'insulinorésistance des tissus se traduit, en plus de l'augmentation de la néoglucogenèse (dans le foie) et la diminution de l'entrée du glucose dans les cellules, par une augmentation de la libération d'acide-gras dans le tissu adipeux (lipolyse). Cette augmentation de la lipolyse tend à aggraver les mêmes phénomènes ayant initialement induit celle-ci, à savoir l'aggravation de l'insulinorésistance, l'augmentation de la néoglucogenèse hépatique et la diminution de l'action et de la sécrétion de l'insuline.

### 2.3 La maladie du foie non alcoolique

Les complications du diabète, et plus généralement les pathologies associées, sont variées et peuvent toucher tous les tissus. En conséquence, l'étude de la physiopathologie du diabète de type 2 nécessite de comprendre les mécanismes biologiques, génétiques et épigénétiques, impliqués dans l'ensemble des tissus connus comme ayant un rôle dans l'hyperglycémie et l'insulinorésistance (Figure 8).

Ainsi, les risques de maladies de peau, rétinopathie, neuropathie, néphropathie, ainsi que les pathologies cardiovasculaires (p. ex. accident cardiovasculaire, accident vasculaire cérébral et hypertension) sont accrus chez les individus diabétiques. À ces pathologies s'ajoutent des pathologies spécifiques du foie, rangées principalement sous l'appellation NAFLD ("Non-Alcoholic Fatty Liver Disease") et NASH ("Non-Alcoholic Steato-Hepatitis") dans les cas les plus sévères. Ces dernières pathologies se définissent à partir de l'évaluation de différents critères, au moyen d'une coupe histologique d'un échantillon de biopsie du foie :

- Stéatose : pourcentage d'accumulation de triglycérides, se caractérisant par la déformation des hépatocytes et l'apparition de taches blanches ;
- Inflammation lobulaire : inflammation et infiltration des lobules du foie, entraînant des lésions du tissu ;
- Ballonnemment hépatocytaire : comptage des hépatocytes présentant un gonflement anormal et une transparence accrue.

Ces trois critères servent à établir un score, le "NAFLD Activity Score" (NAS) (Tableau 3) [Kleiner et al., 2005]. À cela s'ajoute une potentielle fibrose du foie se caractérisant par la destruction d'une partie du tissu hépatique, et pouvant aboutir à une cirrhose, voire au développement d'un carcinome hépatique.

**TABLEAU 3.** Critères constituant le score NAS (“NAFLD Activity Score”).

Critère	Score	Catégorie
Stéatose	0	<5%
	1	5-33%
	2	>33-66%
	3	>66%
Inflammation Lobulaire	0	Pas de foci
	1	<2 foci/200x
	2	2-4 foci/200x
	3	>4 foci/200x
Ballonnement hépatocytaire	0	Aucune cellule
	1	Quelques cellules
	2	Beaucoup de cellules

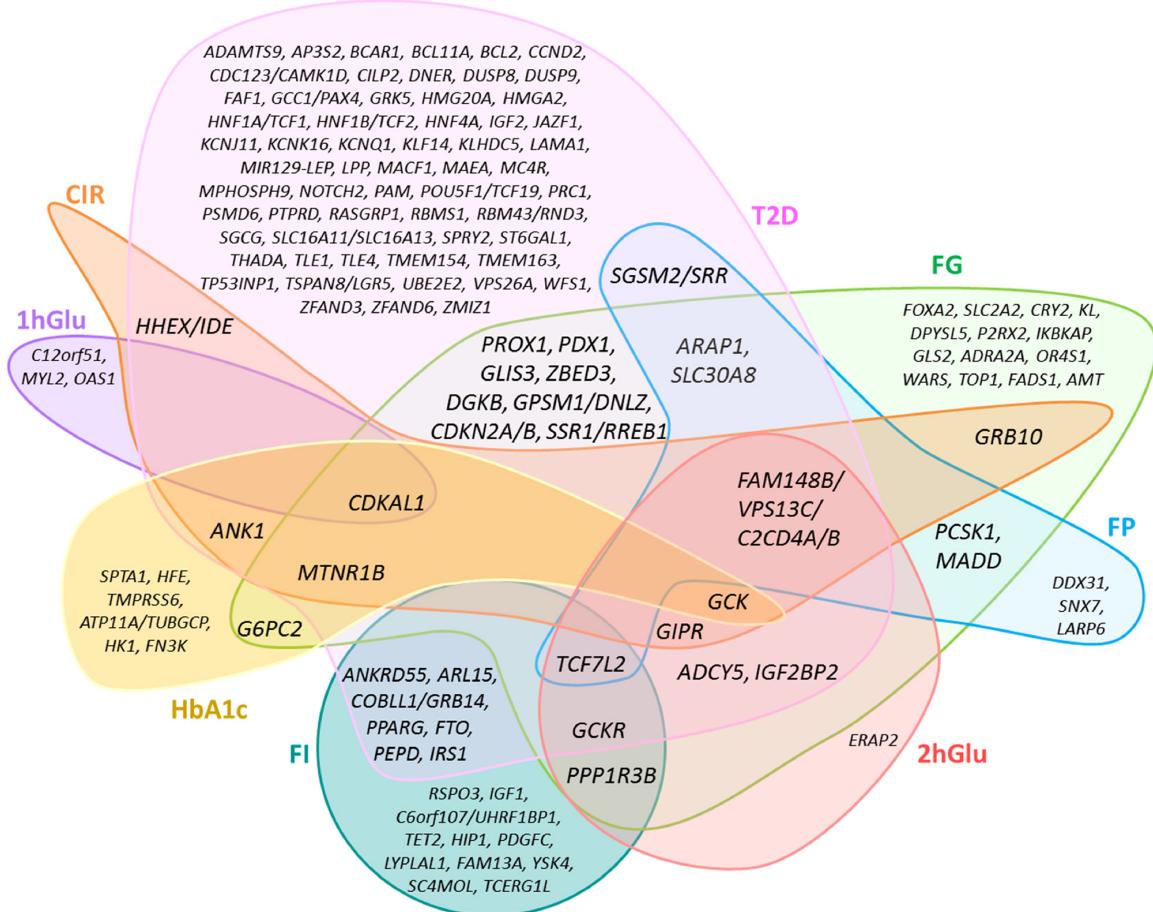
## 2.4 La génétique et l'épigénétique du diabète de type 2

**TABLEAU 4.** Gènes identifiés dans les diabètes de type MODY (“Maturity-Onset Diabetes of the Young”).

	Gène muté	Chromosome	Année de découverte
MODY 1	HNF-4-alfa	20	1991
MODY 2	Glucokinase	7	1992
MODY 3	HNF-1-alfa	12	1994
MODY 4	IPF-1	13	1997
MODY 5	HNF-1-beta	17	1997
MODY 6	NeuroD1	2	1999

Le diabète de type 2 présente une composante génétique dont les premiers éléments ont été mis en évidence dans des études portant sur des jumeaux (monozygotes et dizygotes) [Kaprio et al., 1992], des études d'agrégation familiale, ainsi que des études portant sur des formes monogéniques de diabète, comme les diabètes dits “Maturity-Onset Diabetes of the Young” (MODY) [Thanabalasingham and Owen, 2011], ou diabète de type adulte chez le jeune. Les diabètes de type MODY n’impliquent qu’un seul gène (ou quelques-uns) (Tableau 4). Par exemple, un individu caractérisé MODY 2 présentera une mutation au niveau du gène GCK (Glucokinase),

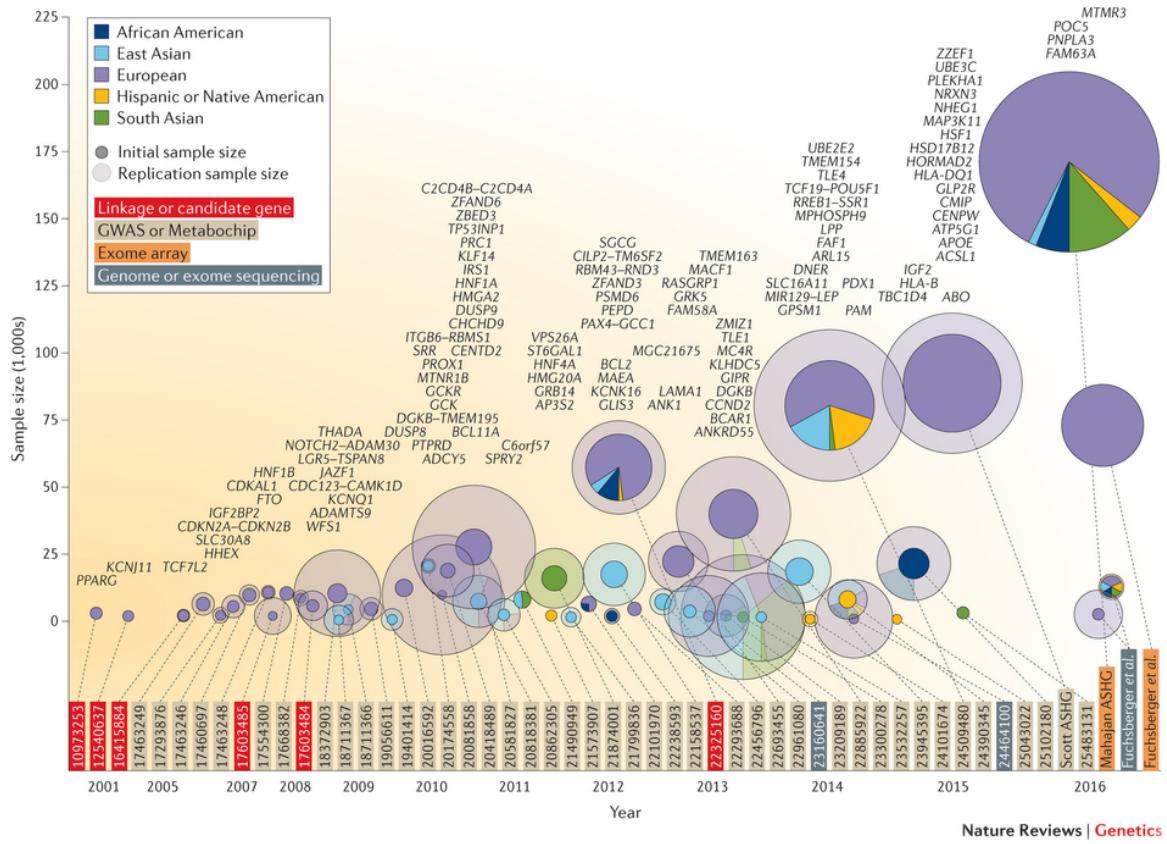
gène impliqué dans la régulation de la glycémie transformant le glucose en glucose-6-phosphate. Il a été montré qu'avoir un parent diabétique augmente le risque de développer un diabète de l'ordre de 30 à 40 %. Ce risque augmente à 70 % lorsque les deux parents sont diabétiques [Kobberling and Tillil, 1982; Meigs et al., 2000].



**FIGURE 9.** Diagramme de Venn des loci identifiés par études d'association pangénomiques pour leur effet sur différents traits glycémiques et le diabète de type 2 [Marullo et al., 2014]. T2D : diabète de type 2; FG : glycémie à jeun; FI : insulinémie à jeun; FP : proinsulinémie à jeun; 2hGlu : glycémie à 2 heures; HbA1c : hémoglobine glyquée; CIR : ratio carbohydrate-insuline.

Au cours de la dernière décennie, et depuis la première étude d'association pangénomique (“Genome-Wide Association Study” ou GWAS) portant sur le diabète de type 2 [Sladek et al., 2007], à l'heure actuelle, plus de 100 loci ont été identifiés (Figure 10) comme étant associés au diabète de type 2, dont certains sont également associés à la glycémie ou l'insulinémie dans des populations normoglycémiques (Figure 9), notamment les gènes *MTNR1B* (“Melatonin Receptor 1B”) [Bouatia-Naji et al., 2009; Prokopenko et al., 2009; Sladek et al., 2007; Tam et al., 2010] et *TCF7L2* (“Transcription Factor 7-Like 2”) [Grant et al., 2006; Groves et al., 2006; Zhang et al., 2006].

Les études ayant mené à l'identification de ces variants regroupent des études de liaison et des études d'asso-



**FIGURE 10.** Historique des loci de susceptibilité au diabète de type 2 identifiés par études d'association pangénomiques [Flannick and Florez, 2016].

ciation de type gènes-candidats. Par exemple, en utilisant la puce Illumina Metabochip [Voight et al., 2012] qui inclut environ 200 000 variants préalablement sélectionnés de résultats provenant des études d'association sur des traits métaboliques, cardiovasculaires et anthropométriques, environ 5 000 variants susceptibles d'être associés au diabète de type 2, et 17 000 autres dans des régions déjà associées dans des études antérieures (par GWAS ou séquençage du génome) ont été testés.

Il est à noter que les loci identifiés par GWAS et par meta-analyses présentent, en premier lieu, des effets observés faibles sur le diabète de type 2 (odds ratio compris entre 1,1 et 1,4) et ne contribuent que faiblement à l'héritabilité de cette pathologie (10 à 15 %) [Scott et al., 2007; Morris et al., 2012]. Ces estimations de l'héritabilité ont conduit à l'émergence d'un débat portant sur "l'héritabilité manquante", ouvrant la voie vers de nouvelles pistes de recherches comme, par exemple, le séquençage de l'ensemble de l'exome ou du génome dans le but d'étudier des variants avec de faibles fréquences alléliques, et l'étude des CNV [Manolio et al., 2009]. La localisation intergénique ou intronique de ces loci ne permet pas d'identifier la fonction de ces variants de façon évidente, à quelques exceptions près, comme par exemple les loci au niveau de GCKR ("glucokinase regulatory protein") et SLC30A8 ("ZnT-8 zinc transporter"), qui entraînent une altération de la séquence codante du transcrit de ces gènes [Beer et al., 2009; McCarthy and Zeggini, 2009; Saxena et al., 2007; Sladek et al., 2007].

La problématique de l'hérabilité manquante a renforcé l'hypothèse de “maladie commune, variants rares”, laquelle stipule que des variants rares pourraient avoir une pénétrance plus forte et un effet plus important sur le risque de diabète de type 2 que les variants communs [Lupski et al., 2011; Schork et al., 2009]. Cette hypothèse fait contrepoids à l'hypothèse présupposée des études gènes-candidats et des GWAS, c'est-à-dire celle de “maladie commune, variant commun” [Schork et al., 2009]. Un variant commun est un variant dont la fréquence allélique est supérieure à 5 % dans la population générale, ce qui représente le seuil standard pour analyse statistique des SNPs dans les GWAS.

Ces études d'associations ont pu mettre en évidence le fait que certains variants pouvaient avoir un effet sur le risque de diabète de type 2, mais également avoir un effet sur des traits cliniques, telles l'insulinémie ou la glycémie [Dupuis et al., 2010; Voight et al., 2010; Yaghoobkar and Frayling, 2013]. Il est à noter qu'en raison de la prise de traitement influençant la glycémie et l'insulinémie, les études réalisées sur ces traits ne l'ont été que chez des individus normoglycémiques (non diabétiques). De plus, les variants présentant un effet à la fois sur la glycémie et sur le risque de diabète de type 2 ne représentent qu'une faible proportion des variants identifiés [Dupuis et al., 2010; Voight et al., 2010; Yaghoobkar and Frayling, 2013]. Cela indique que les mécanismes conduisant au diabète de type 2 et à l'élévation de la glycémie ne sont pas les mêmes, et paradoxalement qu'une élévation de la glycémie chez un individu pourrait ne pas augmenter son risque de développer un diabète de type 2.

L'épigénétique, principalement la méthylation de l'ADN et les modifications d'histones, est devenue une composante importante dans l'étude de la pathogenèse du diabète de type 2. En effet, ces modifications n'altèrent pas la séquence d'ADN et peuvent être transmises de génération en génération [Raciti et al., 2014]. Elles sont également le reflet de facteurs environnementaux et peuvent modifier l'expression, voire activer ou éteindre complètement certains gènes [Zierath and Barrès, 2011]. Dans un sens, ces modifications peuvent avoir un effet équivalent aux SNPs ou à d'autres mutations, en bloquant la transcription d'un gène. Plusieurs éléments viennent corroborer l'idée selon laquelle l'épigénétique pourrait expliquer une partie de “l'hérabilité manquante” dans le diabète de type 2, notamment en tant que reflet de l'environnement intra-utérin, comme cela a été montré dans des populations soumises à des contraintes alimentaires, où le risque de développement d'un diabète de type 2 était accru chez les enfants dont la mère avait connu une famine au moment de la grossesse [Hales and Barker, 1992; Pettitt et al., 1988; Ravelli et al., 1998]. Des études similaires menées chez les indiens Pima ont montré des risques de développement de diabète supérieurs chez l'enfant lorsque la mère présentait une hyperglycémie et/ou un diabète [Dabelea et al., 2000; Pavkov et al., 2010; Pettitt et al., 1983].

Une autre indication vient de l'étude des perturbateurs endocriniens et de polluants qui sont présents sous différentes formes dans divers produits et outils de la vie de tous les jours (p. ex. boîte alimentaire en plastique, produits d'entretien, peintures, etc.). Ces substances peuvent avoir un effet sur la méthylation de l'ADN, résultant

en un changement coordonné de l'expression des gènes (mRNA, miRNA), et ainsi produire un effet sur la sécrétion d'insuline [Hall et al., 2014] ou l'homéostasie du glucose, comme cela a été observé chez l'homme [Bi et al., 2015] et chez les rongeurs (rat et souris) [Li et al., 2014; Rajesh and Balasubramanian, 2015]. En raison du caractère tissu-spécifique de la méthylation, les premières études se sont focalisées sur les tissus dont les échantillons étaient facilement prélevables, tels que le sang [Bell et al., 2010; Canivell et al., 2014; Chambers et al., 2015; Dayeh et al., 2016; Toperoff et al., 2012] et le pancréas, notamment les îlots pancréatiques impliqués dans la sécrétion d'insuline [Dayeh et al., 2014; Hall et al., 2014; Stitzel et al., 2010; Volkmar et al., 2012]. Dans l'une des premières études de l'épigénome à grande échelle via l'utilisation de puce Illumina HumanMethylation450 BeadChip (~480 000 sites CpG couverts), plus de 1 600 CpG (~850 gènes), incluant des loci connus tels que *TCF7L2* et *KCNQ1*, ont été identifiés comme étant différentiellement méthylés entre des diabétiques et des non diabétiques [Dayeh et al., 2014]. Des études plus récentes ont apporté des pistes de réponse, quant à la nature causale de la méthylation, en considérant les polymorphismes identifiés dans le diabète de type 2 (Consortium DIAGRAM) [Morris et al., 2012]. Ainsi, la méthylation du locus *KCNQ1* serait causale dans le développement du diabète de type 2 [Elliott et al., 2017].

Bien que l'ère des GWAS ait permis d'identifier plus de 100 loci associés au diabète de type 2, les mécanismes liant ces variants à sa pathogenèse restent méconnus pour une grande partie d'entre eux. À cela s'ajoute que ces variants ne constituent qu'une faible part de l'héritabilité de cette maladie complexe. Ainsi, au cours des dernières années, les axes de recherches se sont progressivement déplacés vers l'étude d'autres “-omiques” comme la transcriptomique, l'épigénomique, ou encore la métabolomique. La grande diversité des organes/tissus impliqués dans la pathogenèse du diabète de type 2 renforce la nécessité de recueillir et d'étudier en détail le caractère spécifique des tissus et la fonction des cellules qui les composent, afin de pouvoir établir une carte détaillée des mécanismes sous-jacents au développement du diabète de type 2. Avec le développement des techniques et technologies, il est également possible non seulement d'étudier séparément la génétique, l'épigénétique et les facteurs environnementaux afin de classer les individus selon leur risque de développer un diabète, mais aussi d'étudier les interactions et les connections entre ces différentes composantes. Les études fonctionnelles représentent également une étape importante dans la compréhension des mécanismes biologiques des loci identifiés à l'aide des études “-omiques”. La variété et la croissance de la quantité des données générées nécessitent un développement constant d'outils et de méthodes statistiques visant à identifier des gènes ou loci candidats. L'intégration des différentes données “-omiques” offre la possibilité de mettre au jour de nouvelles connaissances sur la chronologie des mécanismes en amont et en aval du développement d'une pathologie, mais également de révéler les liens unissant le génotype et le phénotype.

### 3 Méthodes statistiques : des données à la biologie

#### 3.1 La statistique génétique

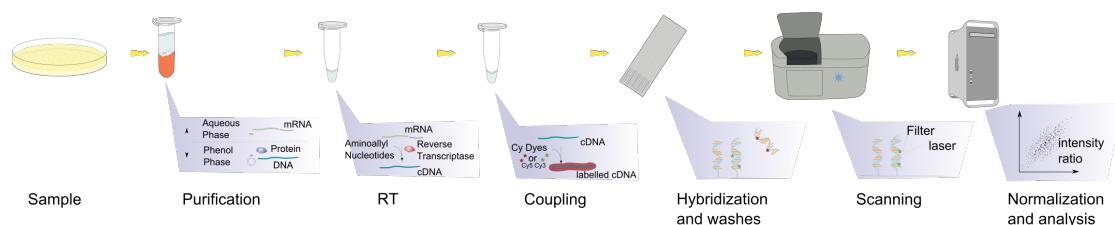
Les principes de l'hérabilité développés par Gregor Mendel (1822-1884) à la fin du XIXème siècle ont servi de fondements à la plupart des connaissances actuelles sur la transmission des traits des parents à leurs enfants. Au fur et à mesure du développement du concept de traits hérités s'est développée la génétique, la science qui étudie la transmission de ces traits et du matériel biologique dans les organismes vivants. La génétique est devenue partie intégrante de la recherche sur l'origine de certaines maladies, comme l'obésité et le diabète. Contrairement à l'étude des végétaux et des petits animaux (rat, souris, etc.), où la croissance et les croisements peuvent être contrôlés de façon expérimentale, et où la transmission des caractères étudiés est rendue possible dans un temps limité (en particulier grâce à un temps réduit de passage d'une génération à la suivante, p. ex. 2-4 mois pour le poisson-zèbre), la situation est plus complexe chez l'Homme, puisqu'il faut entre 20 et 30 ans pour qu'une nouvelle génération voit le jour.

Aujourd'hui, avec les évolutions technologiques au niveau des plateformes moléculaires, ayant notamment permis le séquençage du génome humain ("Human Genome Project" [Sawicki et al., 1993]), le volume des données génomiques a augmenté au cours des dernières décennies et a ainsi permis le développement d'une branche de la statistique, soit la statistique génétique. Cette nouvelle branche vise à développer des outils d'analyses des caractères hérités et plus généralement des données génétiques, permettant en outre l'identification de facteurs de risque ou de déterminants génétiques pour des maladies complexes, tels que les cancers, les diabètes, les maladies cardio-vasculaires ou les troubles psychiatriques. Une maladie complexe est définie comme une pathologie dont les causes sont multiples, lesquelles peuvent être le fruit d'une interaction de facteurs comportementaux, environnementaux et génétiques pouvant produire un effet sur plusieurs gènes de façon simultanée. Ces maladies complexes s'opposent aux maladies dites mendéliennes, dont l'origine s'explique principalement par la génétique, via des processus de transmission conjecturés par Gregor Mendel. Ces maladies se caractérisent par la transmission d'un gène délétère à la descendance, comme c'est le cas pour les formes de diabète MODY discutées précédemment.

### 3.2 Recueil et prétraitement des données

#### 3.2.1 Puce-à-ADN

L'émergence des puces à ADN ("DNA micro-array"), notamment propulsées par les grands projets internationaux de séquençage tels le "Human Genome Project" [Sawicki et al., 1993], "HapMap Project" [Gibbs et al., 2003], "1 000 Genomes Project" [Siva, 2008; The 1000 Genomes Project Consortium, 2015], a permis d'étendre le champ d'application de la statistique grâce à la disponibilité et la variété des données issue de ces puces, à savoir aussi bien des données de transcriptomique, de génomique et d'épigénomique. En effet, les puces à ADN utilisent le principe d'hybridation de l'ADN reposant sur la complémentarité de ces bases. Rappelons le fonctionnement des puces à ADN d'un point de vue général.



**FIGURE 11.** Schéma du protocole des puces à ADN.

L'ADN est extrait et purifié à partir d'un échantillon de tissu (p. ex. prélèvement sanguin ou salivaire). L'ADN purifié peut ensuite être amplifié au moyen d'une réaction en chaîne par polymérase (PCR), un procédé qui permet d'augmenter la quantité d'ADN. Un marquage des séquences par le remplacement de certaines bases nucléotidiques par leur analogue radioactif est ensuite réalisé. Les séquences d'ADN marquées sont ensuite hybridées sur des sondes spécifiques disposées sur les puces à ADN, puces pouvant contenir des milliers de ces sondes complémentaires de séquences d'ADN. Une fois l'hybridation des séquences d'intérêts réalisée, la puce est passée dans un outil de visualisation permettant de lire et quantifier la fluorescence émise par chaque puits (position d'une sonde sur la puce). Selon le type d'omique, le protocole utilisé dans les puces varie et peut inclure des étapes spécifiques, telles qu'une étape de conversion de l'ARN en cDNA (ADN complémentaire de l'ARN reconstitué par une étape de transcription inverse) pour une étude transcriptomique, ou bien une étape de bisulfitation dans le cas d'une étude méthylomique, permettant le changement de la cytosine non-méthylée d'un groupement CpG en uracile avant de passer à l'étape d'amplification PCR. C'est à l'issue de ces étapes (Figure 11), que l'information génomique, transcriptomique ou méthylomique est disponible sous forme de données numériques pouvant être analysées après un prétraitement et un contrôle qualité.

### 3.2.2 Prétraitement

Comme dans toute analyse statistique, la validité des résultats est conditionnée par la qualité des données. De ce fait, une étape de contrôle de qualité et de plausibilité des données est indispensable. En plus des problématiques génériques telles que les valeurs extrêmes, les études omiques soulèvent quelques niveaux de complexité supplémentaires provenant en grande partie des protocoles complexes générant ces données. Ainsi, dans le cas du génotypage, la qualité peut être influencée par plusieurs facteurs n'étant pas toujours sous contrôle, comme la qualité de l'ADN qui dépend du type de prélèvement de l'échantillon (p. ex. échantillon sanguin ou buccal, biopsie, etc.), le stockage et la conservation de l'échantillon (p. ex. température, stockage paraffine, etc.), et la plateforme de génotypage (c.-à-d. la technologie utilisée par le manufacturier).

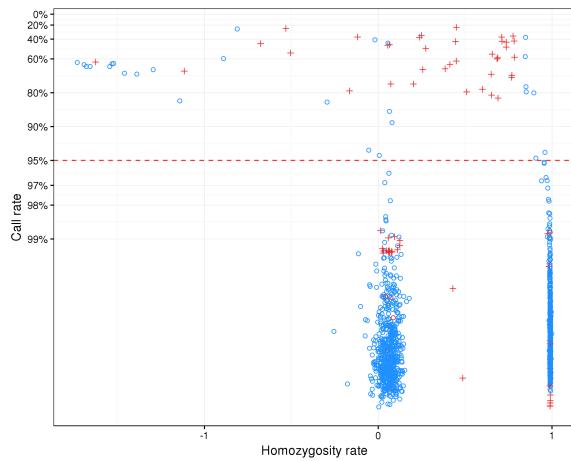
#### a Génomique

Dans les études populationnelles, les erreurs de génotypage survenant indépendamment du statut des individus (p. ex. malade/non malade) ou de leur génotype, peuvent occasionner une diminution de la puissance statistique sans pour autant modifier le risque de première espèce des tests [Fardo et al., 2009; Gordon and Ott, 2001; Marquard et al., 2009], ce qui peut ne pas être le cas dans les études familiales où le taux de faux positifs peut alors être augmenté [Yan et al., 2016; Abecasis et al., 2001]. Cette diminution de la puissance statistique et augmentation de l'erreur de type 1 sont d'autant plus importantes lorsque les erreurs de génotypage se trouvent être associées aux génotypes et/ou phénotypes, voire au plan expérimental (c.-à-d. différents techniciens, séparation complète des groupes étudiés sur des puces différentes, etc.). Une étape de contrôle qualité peut consister en l'application de plusieurs filtres successifs, en particulier au moyen du logiciel PLINK [Chang et al., 2015; Purcell and Chang, 2015] qui dispose de nombreuses fonctionnalités pour la manipulation de fichiers génomiques. Ces filtres peuvent être regroupés en deux catégories, d'une part sur les individus, et d'autre part sur les variants génétiques.

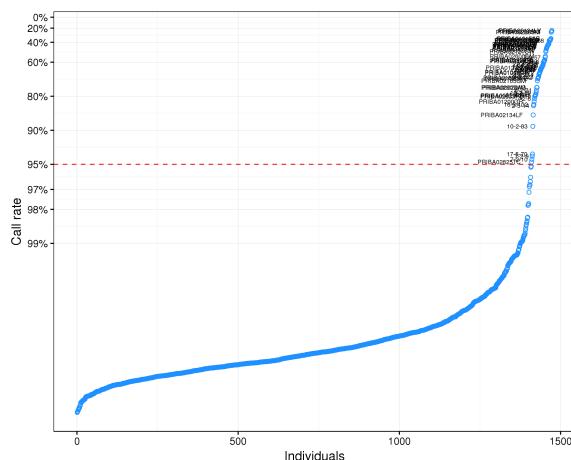
##### a.1 Contrôle qualité des échantillons

- Concordance du genre entre le génotype et le phénotype (Figure 12)

En mesurant le taux d'homoygotie au niveau des gonoosomes pour chaque individu, il est possible de déterminer le sexe à partir du chromosome X. Ainsi, pour les femmes, qui présentent deux chromosomes X et peuvent donc avoir deux allèles différents pour chaque variant, le taux d'homoygotie doit être inférieur à 0,2. Pour les hommes, qui ne présentent qu'un seul chromosome X, le taux d'homoygotie attendu est de 1, ou au moins supérieur à 0,8 (seuil de tolérance). Ce filtre a deux objectifs : vérifier l'information du phénotype, et fournir une information quant à la qualité du génotypage.



**FIGURE 12.** Le taux d'homozygotie (estimé à l'aide des variants du chromosome X) est représenté en fonction de la proportion de génotypes manquants par échantillon. Le taux d'homozygotie attendu pour les hommes est de 1 et inférieur à 0,2 pour les femmes. Les points rouges représentent les échantillons pour lesquels les informations sur le sexe sont discordantes ou manquantes.



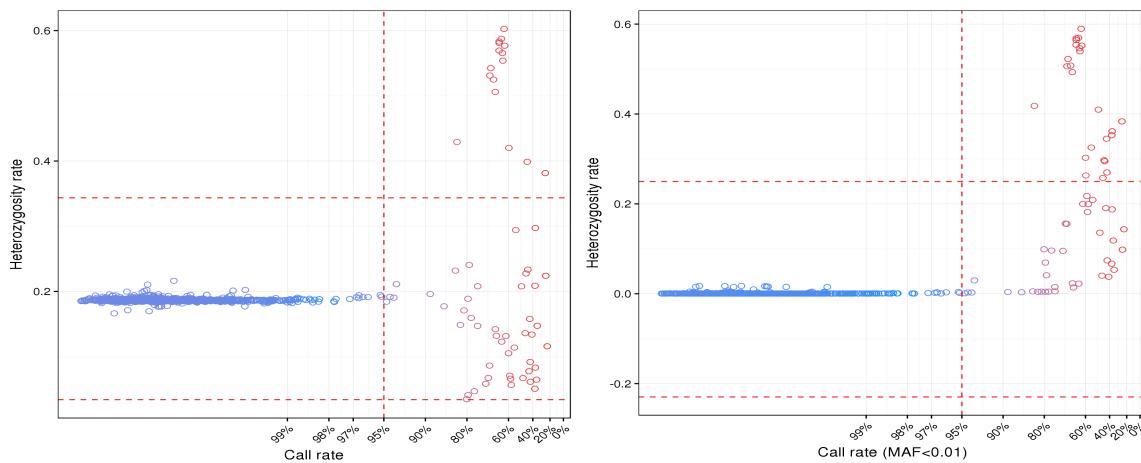
**FIGURE 13.** Distribution du taux de génotype manquant par échantillon.

- Taux de génotypage ou taux de génotype manquants (Figure 13)

Cette vérification permet également d'identifier les individus pour lesquels un problème est survenu lors du génotypage ou de l'extraction d'ADN, en particulier un défaut de qualité de l'ADN. Les individus présentant plus de 5 % de génotypes manquants sont généralement exclus à ce stade.

- Taux d'hétérozygotie (autosomes) (Figure 14)

L'objectif de cette vérification est de s'assurer d'une qualité homogène des génotypes des individus. La distribution du taux d'hétérozygotie (hors chromosomes sexuels) chez tous les individus doit être inspectée pour identifier les individus ayant une proportion excessive ou réduite de génotypes hétérozygotes. En effet, cela peut respectivement indiquer une contamination (p. ex. mélange de deux ADN), ou une consanguinité au sein des échantillons d'ADN. Les individus présentant un taux d'hétérozygotie extrême par rapport à la

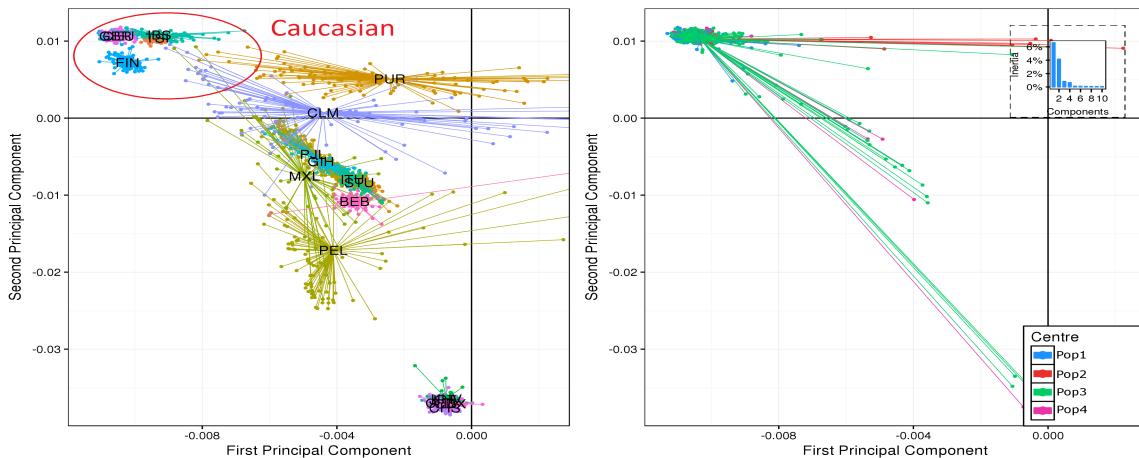


**FIGURE 14.** Taux d'hétérozygotie par échantillon par rapport au taux de génotypage. À gauche, le taux d'hétérozygotie pour l'ensemble des variants, à droite, le taux d'hétérozygotie des variants dont la fréquence allélique est inférieure à 1 %. Les lignes rouges horizontales représentent l'intervalle à plus ou moins quatre fois l'écart-type du taux moyen d'hétérozygotie. La ligne rouge verticale indique le seuil du taux de génotypage.

distribution de celui-ci sur l'ensemble des individus sont exclus, généralement sur la base d'un écart à la moyenne de trois à quatre fois l'écart-type. Le taux d'hétérozygotie (donnée par  $\frac{(n-h)}{n}$ , où  $n$  est le nombre de génotypes total observé et  $h$  est le nombre de génotypes homozygotes observé pour un individu donné) différera selon les populations étudiées et selon l'ensemble des SNPs ciblés par une puce donnée.

- *Degré d'apparentement (étude non-familiale)*

Dans le contexte des études populationnelles d'association pangénomiques, les individus étudiés sont sélectionnés pour satisfaire un critère de non-apparentement en plus de critères purement liés aux hypothèses auxquelles l'étude doit répondre. En effet, la présence d'individus apparentés (c.-à-d. du second degré ou plus proche, ou ayant plus de 20 % de leur génome parfaitement identique), voire d'individus en doublons, peuvent introduire un biais dans l'étude par la surreprésentation de génotypes spécifiques à quelques familles, et conséquemment modifier les fréquences alléliques qui ne seront alors plus représentatives de la population étudiée. Pour éviter ce biais, le degré d'apparentement de chaque paire d'individus est mesuré à partir de la proportion de leurs génomes partagés avec un ancêtre commun (identité par descente ou IBD). De cette façon, les individus en doublons ou jumeaux (monozygotes) présenteront un  $IBD \simeq 1$ , un  $IBD \simeq 0,5$  pour les individus ayant un lien du premier degré (p. ex. parents, enfants, frères et sœurs) et un  $IBD \simeq 0,25$  pour un lien du second degré (p. ex. oncles, tantes, grand-parents, etc.). En raison d'erreur de génotypage, de stratification cachée dans l'échantillon (p. ex. due à différentes origines ethniques) ou de déséquilibre de liaison, ces valeurs d'IBD théoriques peuvent varier avec des données réelles et des intervalles de valeurs sont alors tolérés : par exemple,  $[0,20 ; 0,30]$  pour le premier degré, ou  $[0,40 ; 0,60]$  pour le second degré.

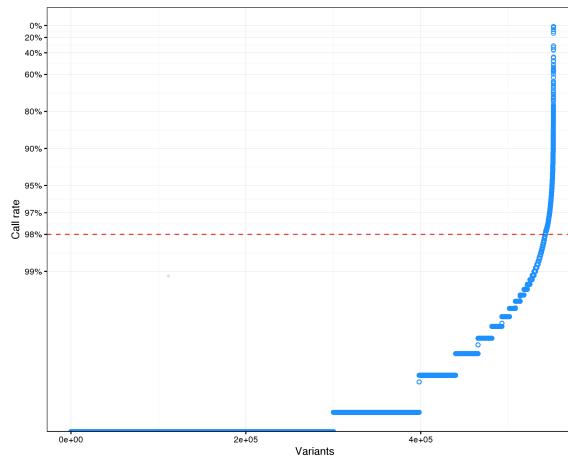


**FIGURE 15.** Premier plan factoriel de l'analyse en composante principale du jeu de données combinant la population d'étude et celle de référence (1 000 génomes). Avec à gauche la population de référence et à droite la population d'étude.

- *Stratification de la population* (Figure 15)

Comme évoqué précédemment, une stratification peut exister au sein de la population d'étude, créée par des individus d'origines ethniques différentes ou de zones géographiques différentes, et peut induire un biais dans les résultats lors de l'analyse [Clayton et al., 2005; Cardon and Palmer, 2003], en particulier si cette stratification n'est pas la même entre les sous-groupes formés des cas et des témoins. L'approche la plus courante pour identifier une stratification demeure l'analyse en composantes principales ou ACP [Caussinus, 1986; Patterson et al., 2006; Price et al., 2006]. L'ACP est une méthode statistique multivariée qui, à partir d'une matrice contenant l'ensemble des observations (dans notre cas, les individus génotypés) sur un nombre  $N$  de variables potentiellement corrélées (c.-à-d. les SNPs), vise à obtenir un nombre réduit  $n < N$  de composantes principales non corrélées et orthogonales. Les composantes sont calculées de sorte que la part de variabilité qu'elles peuvent expliquer décroisse de la première à la dernière composante. Afin d'évaluer une potentielle stratification d'origine ethnique, la matrice des génotypes est augmentée des génotypes provenant d'une base de référence [Gibbs et al., 2003; Siva, 2008; The 1000 Genomes Project Consortium, 2015]. Ces bases de références contiennent des individus dont l'origine ethnique a été vérifiée par génotypage ou séquençage. L'ACP est alors réalisée sur un jeu de données comportant les populations de références et l'échantillon étudié. En raison de la grande diversité génétique observée entre individus d'origines caucasiennes, africaines et asiatiques, les deux premières composantes sont généralement suffisantes pour identifier une stratification ethnique dans l'échantillon.

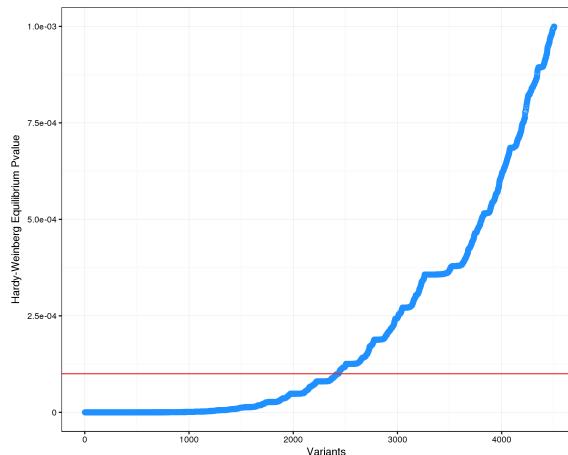
### a.2 Contrôle qualité des SNPs



**FIGURE 16.** Taux de génotypage par variant. La ligne rouge indique le seuil de 98 %.

- *Taux de génotypage ou taux de génotypes manquants* (Figure 16)

Sur le même principe que le taux de génotypage pour un individu, le taux de génotypage d'un SNP est examiné. Un taux de succès, généralement fixé à 95 %, est toléré, seuil en dessous duquel le SNP sera exclus des analyses.

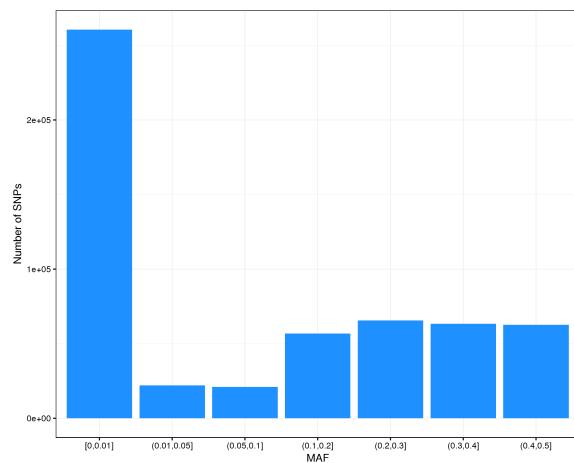


**FIGURE 17.** Distribution des valeurs-p du test des variants à l'équilibre de Hardy-Weinberg (HWE). La ligne rouge horizontale indique le seuil de significativité pour  $\alpha = 0,0001$ .

- *Équilibre de Hardy-Weinberg* (Figure 17)

L'équilibre de Hardy-Weinberg (HWE) constitue l'un des principes fondamentaux de la génétique des populations. Pour une population suffisamment grande, non apparentée (c.-à-d. population panmictique où les accouplements se font au hasard ou de façon équiprobable), sans pression de sélection, et lorsque les générations d'individus successives sont discrètes et séparées, cet équilibre prédit que les proportions génotypiques d'un variant donné restent constantes d'une génération à la suivante et s'écrivent simplement comme le produit mathématique des fréquences alléliques de cette population. Une forte déviation par

rapport à l'HWE est un motif d'exclusion d'un SNP dans les études associations pangénomiques, car peu probable et sans doute révélatrice d'une erreur de génotypage. Mais un écart important par rapport à l'HWE peut également indiquer un effet de sélection, c'est-à-dire que les cas (dans une étude cas/témoin) peuvent montrer une déviation à l'HWE pour des loci associés à la maladie étudiée : exclure ces loci reviendrait donc à exclure ce qui est précisément l'objet de l'étude [Wittke-Thompson et al., 2005]. Ceci explique pourquoi le seuil de significativité du test d'écart à l'HWE varie d'une étude à l'autre, bien que ce test ne soit effectué que dans le groupe témoin [Meyre et al., 2009; Sladek et al., 2007; Burton et al., 2007].



**FIGURE 18.** Répartition du nombre de SNP par classe de fréquence des allèles mineurs (MAF).

- Fréquence allélique mineure (en anglais, maf) (Figure 18)

Un filtre sur la fréquence allélique est appliqué pour ne conserver dans les analyses statistiques que les polymorphismes dont la fréquence de l'allèle mineur est supérieure à 5 % (par définition, cette fréquence est comprise entre 0 et 50 %). Dans certaines études, pour conserver des SNPs considérés comme "rares", c.-à-d. dont la maf est inférieure à 5 %, il est possible d'augmenter le seuil du taux de génotypage par SNP [Burton et al., 2007]. Cependant, les résultats des tests d'associations observés pour ces SNPs rares sont moins robustes, malgré un taux de génotypage plus élevé (p. ex. 99 %), principalement parce que ces résultats peuvent être produits par les génotypes rares de quelques individus seulement. En effet, Morris and Zeggini [2010] ont montré que la puissance statistique pour détecter des associations pour des SNPs rares était faible, particulièrement avec des approches dites "*simple SNP*", c.-à-d. un SNP à la fois. En réalité, leur exclusion n'aurait qu'un impact modéré sur les résultats de l'étude.

En conclusion, même après avoir appliqué ces différents filtres de contrôle-qualité, aussi bien au niveau des individus qu'au niveau des SNPs, des erreurs de génotypage peuvent subsister, d'où la nécessité de répliquer les associations détectées dans d'autres échantillons.

**b Transcriptomique**

Les données de transcriptomique provenant de la lecture et de la quantification de la fluorescence d'une puce à ADN nécessitent également un prétraitement afin de garantir la validité et la fiabilité de celles-ci. Avant la réalisation d'une étude transcriptomique, une considération particulière doit être prise quant à la conception du plan d'expérience pour réduire les biais techniques, p. ex. en équilibrant les échantillons sur les puces et plaques, en réalisant l'expérience en un minimum de temps, ou en limitant le nombre d'expérimentateurs [Quackenbush, 2002]), et ainsi permettre un contrôle qualité plus efficace, particulièrement lors de la normalisation des données.

Les plateformes qui permettent de quantifier l'expression des gènes via la quantification d'ADN complémentaire (cDNA) à partir de séquences d'ARN (mRNA, microRNA) ne font pas appel aux mêmes techniques. Plusieurs outils ont été développés permettant l'importation et le prétraitement des données brutes directement depuis le logiciel statistique R [Smyth et al., 2017; Lopez-Romero, 2016].

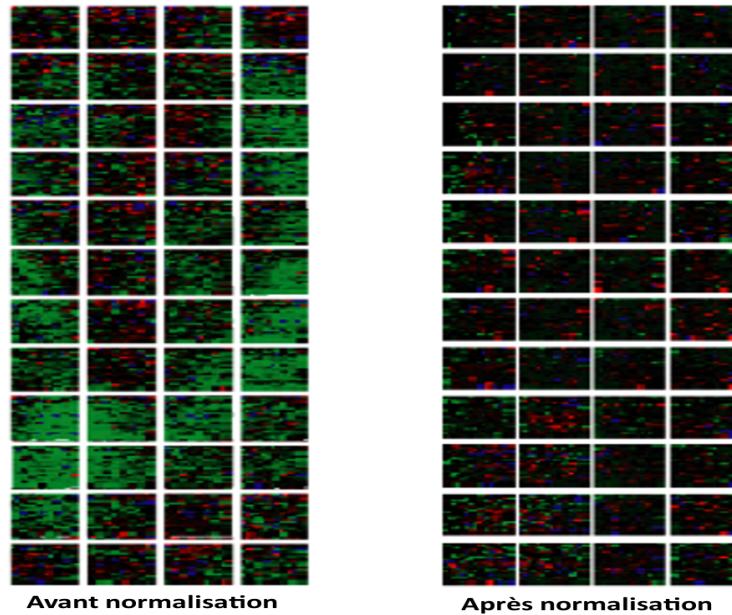
- *Valeur-p de détection*

Selon la plateforme utilisée (principalement chez Illumina), la mesure brute d'expression peut être accompagnée d'une valeur-p de détection calculée à partir de la mesure d'intensité de sondes contrôles, permettant d'évaluer si le signal observé est statistiquement différent de l'intensité (artéfactuelle) observée au niveau des sondes contrôles. Cette mesure peut alors être utilisée en tant que filtre en amont des étapes de normalisation pour exclure les sondes non-détectées (à partir d'un seuil, p. ex. valeur-p < 0,05) sur un nombre suffisant déterminé par l'expérimentateur ou analyste (p. ex. sonde détectée sur 95 % des échantillons).

- *Correction du "bruit de fond"*

En effet, après que les puces à ADN aient été scannées pour évaluer la fluorescence des sondes permettant la quantification indirecte de l'ARN, deux types de mesures sont disponibles : l'intensité de fluorescence au niveau d'un puits (une sonde par puits) et l'intensité de fluorescence ambiante (au voisinage des puits). Cette seconde information, appelée "bruit de fond", doit être prise en compte. Plusieurs approches sont disponibles pour réaliser cette correction de l'intensité des sondes (signal) par l'intensité ambiante, dont l'approche la plus classique consiste à soustraire l'intensité ambiante au signal. Cependant, cette correction produit des effets indésirables, puisqu'elle peut générer des valeurs négatives lorsque l'intensité ambiante est plus forte que le signal ce qui, lors du passage au logarithme ou logarithme-ratios, aboutissent à la génération de données manquantes, rendant de ce fait inexploitable ces mesures. L'extension R *limma* [Smyth et al., 2017] propose plusieurs méthodes pour cette correction du "bruit de fond", dont une approche basée sur un modèle de convolution normale + exponentielle. Le modèle suppose que les intensités observées

sont la somme de l'intensité ambiante et du signal, l'intensité ambiante suivant une distribution normale lorsque le signal suit une distribution exponentielle [Irizarry, 2003; Silver et al., 2009; Ritchie et al., 2007]. Cette méthode permet de garantir que le signal de l'ensemble des sondes est strictement positif, et ainsi permet le passage au logarithme sans perte de données.



**FIGURE 19.** Données d'expression avant et après normalisation des intensités.

- *Normalisation inter-puces*

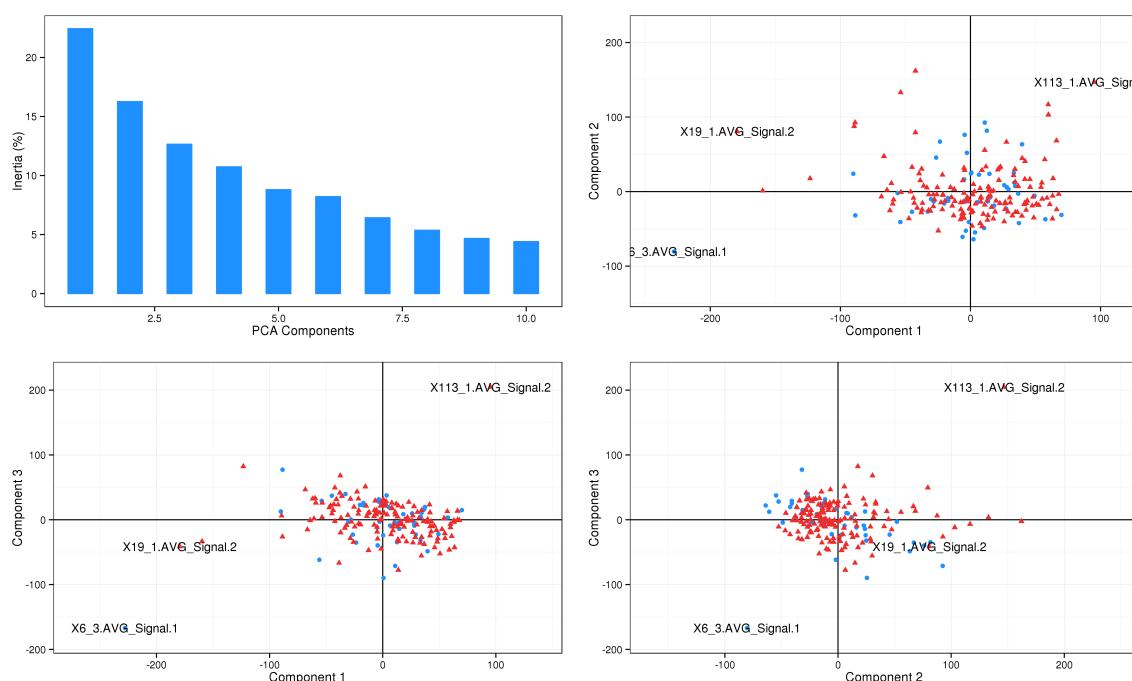
Une étape de normalisation est réalisée afin de repérer et corriger un effet, ou biais, systématique sur l'ensemble des mesures d'expression pouvant correspondre à un effet puce, c'est-à-dire qu'il peut exister une différence dans l'application des protocoles d'une puce à l'autre induisant des mesures systématiquement plus élevées sur une puce en particulier. Ces différences techniques peuvent être le résultat d'une mauvaise hybridation des cDNA sur une puce, ou d'une faible quantité de cDNA aboutissant à des pics de fluorescence plus faibles dans ces cas-là (Figure 19). La principale méthode de normalisation utilisée dans les études transcriptomiques est la normalisation quantile, permettant de rendre similaire, d'un point de vue statistique, deux ou plusieurs distributions. Autrement dit, la distribution des mesures d'expression d'une puce est prise en référence, et la distribution des mesures obtenues sur une seconde puce est normalisée pour que celle-ci soit comparable à la première. Dans le même temps, la normalité des mesures d'expression n'étant pas toujours vérifiée, une transformation logarithmique (base 2) est préalablement appliquée sur les données brutes, ou sur les ratios des mesures obtenues : pour une sonde dans une condition donnée, sa mesure est divisée par la mesure obtenue pour la même sonde dans une condition contrôle, comme cela est fait dans les expériences de quantification par RT-PCR, permettant ainsi la quantification (fluorescence) d'un mRNA spécifique via la transcription inverse suivie d'une amplification des fragments de cDNA.

- *Filtre des sondes*

Une fois les données normalisées, une étape additionnelle peut être appliquée pour filtrer les sondes, par exemple, pour ne conserver que les sondes exprimées (à partir d'un seuil défini au préalable, selon un gène de ménage servant de référence, c'est-à-dire un gène dont le niveau d'expression est constant dans l'ensemble des tissus) sur un certain nombre d'échantillons d'une condition (p. ex. 95 % des individus contrôles).

- *Profil extrême global* (Figure 20)

Enfin, un dernier contrôle consiste à la vérification et à l'identification d'individus présentant un profil transcriptomique extrême, par exemple, au moyen d'une ACP. D'une part, l'ACP permettra de pouvoir identifier une éventuelle stratification, au sein des individus, associée ou non aux conditions expérimentales ; d'autre part, elle permettra d'identifier des individus extrêmes par rapport à l'ensemble des conditions ou d'une condition expérimentale donnée, qui pourrait être liée à la quantité de cDNA ou à la qualité d'ARN.



**FIGURE 20.** Identification de profil extrême à partir des premières composantes de l'analyse en composante principale.

Les puces à ADN utilisées en transcriptomique n'exploitent pas toutes les mêmes techniques expérimentales (p. ex. puce monochrome ou bi-couleur), nécessitant de ce fait d'adapter les étapes de prétraitement et de contrôle qualité décrites précédemment.

### c Méthylomique

Comme les précédentes techniques omiques, les données de méthylomique doivent faire l'objet d'un prétraitement et d'un contrôle qualité. Les principes et techniques évoqués pour la génomique et la transcriptomique peuvent et sont employés également en méthylomique. Néanmoins, d'autres vérifications et certaines adaptations méthodologiques sont nécessaires. En génomique, les valeurs sont discrètes et codées 0, 1, 2 (modèle additif) ; en transcriptomique, les valeurs (brutes) sont continues et définies sur l'intervalle  $[0, +\infty)$ , tandis qu'en méthylomique, les données sont bornées entre 0 et 1.

Les étapes décrites ci-après, quoique non-exclusives ou exhaustives, concernent principalement la puce Illumina HumanMethylation450 [Bibikova et al., 2011], puce qui a été utilisée dans l'article du Chapitre 3. La puce Illumina HumanMethylation450 permet l'identification de la méthylation sur plus de 450 000 sites CpG localisés sur l'ensemble du génome, et se caractérise par l'utilisation de deux processus d'analyses chimiques différents (Infinium I et II). L'Infinium I utilise un système de marquage fluorescent monocouleur, tandis que l'Infinium II exploite un système bi-couleur pour quantifier la méthylation. À cela, s'ajoute un second niveau de différences, puisque l'Infinium I dispose de deux types de sondes pour identifier les allèles méthylés et les allèles non-méthylés. L'Infinium II n'utilise qu'un seul type de sonde qui, lors de l'hybridation des fragments d'ADN sur la puce, rend accessible ou non la base nucléotidique complémentaire à celles marquées (T et G pour les sites CpG respectivement non-méthylés et méthylés) (Figure 21). Dedeurwaerder et al. [2011] ont montré que ces différentes techniques (c.-à-d. réactions chimiques et types de sonde) impactaient directement les résultats obtenus. Les extensions R et les algorithmes de prétraitement des données de méthylation se sont fortement développés et multipliés ces dernières années, imposant à l'utilisateur la délicate tâche de sélectionner les méthodes les plus efficaces et les plus adaptées à ces données.

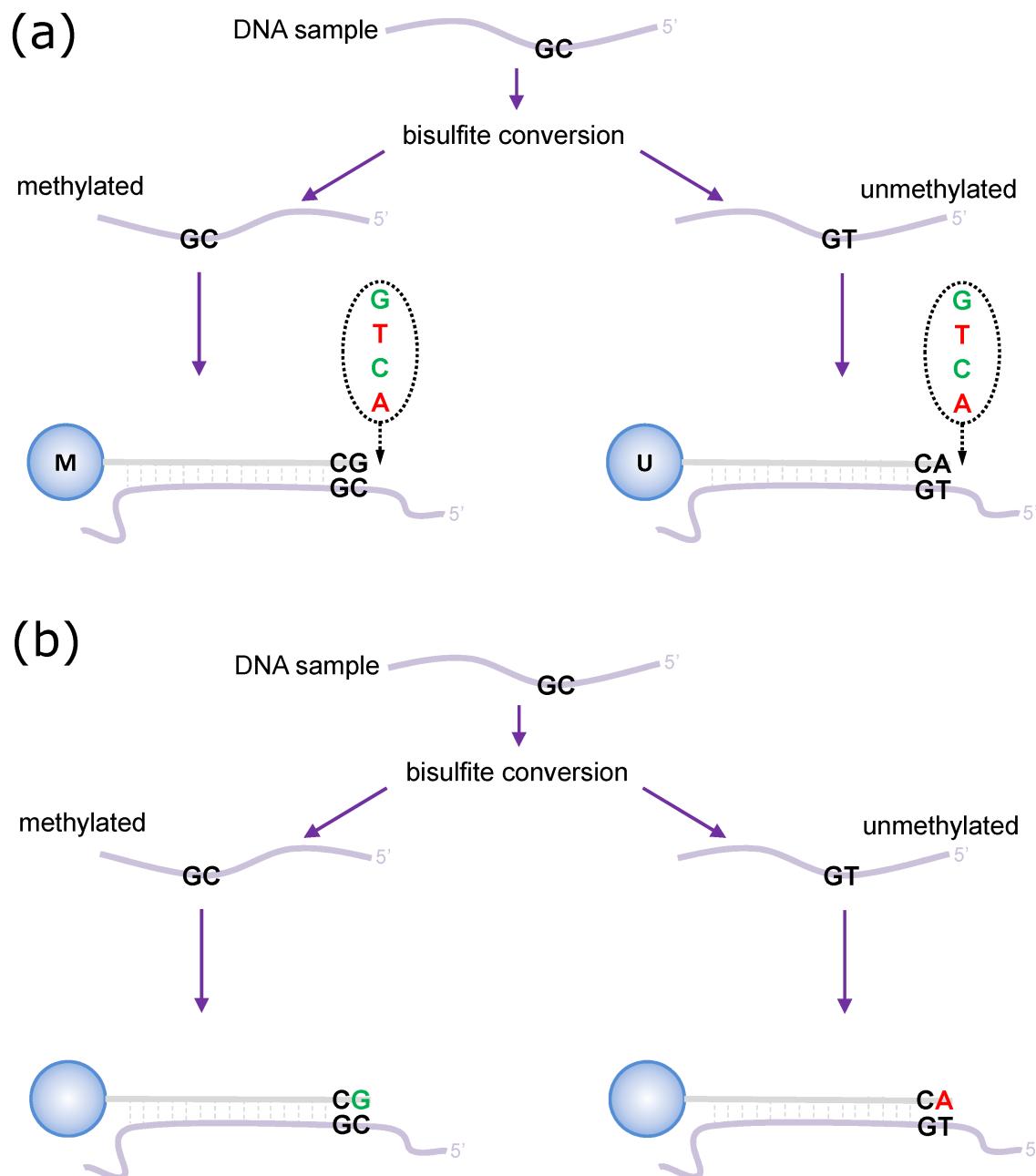
#### c.1 Contrôle qualité des sites CpG

- *Filtre des sites*

En amont, un premier filtre est réalisé pour exclure les sites de méthylation pouvant présenter des résultats étranges, principalement pour des raisons techniques :

- Sondes trans-réactives

La transformation des cytosines non-méthylées en thymine lors de la bisulfitation entraîne un changement dans la distribution des quatre bases nucléotidiques dans le génome, et par conséquent augmente la probabilité que les sondes Infinium puissent s'hybrider sur d'autres portions du génome que celles ciblées initialement. La méthylation observée sur ces sites peut donc être le résultat d'un mélange de la méthylation du site cible et de celle d'autres sites. Des annotations des sites supposés



**FIGURE 21.** Représentation des techniques de marquage utilisées sur la puce Illumina Infinium HumanMethylation450 [Maksimovic et al., 2012]. a) Infinium I. Chaque CpG est interrogé à l'aide de deux types de sondes : méthylée (M) et non-méthylée (U). Les deux types de sondes incorporent le même nucléotide marqué pour un site CpG donné, produisant ainsi la même fluorescence. Le nucléotide qui est ajouté est déterminé par la base en aval du "C" du CpG ciblé. Le pourcentage de méthylation peut être calculé en comparant les intensités des deux sondes (U et M) de la même couleur. b) Infinium II. Chaque CpG est interrogé à l'aide d'un seul type de sonde. L'état de méthylation est détecté par la base complémentaire unique à la position du "C" du CpG ciblé, ce qui entraîne l'ajout d'un nucléotide marqué "G" ou "A", complémentaire à la cytosine (méthylé) ou à la thymine (non-méthylé), respectivement. Le pourcentage de méthylation du site CpG ciblé est déterminé en comparant les intensités des deux couleurs émises par les nucléotides marqués.

ou vérifiés comme non-spécifiques ont été générées pour permettre de les identifier et/ou de les exclure [Price et al., 2013; Zhang et al., 2012; Chen et al., 2013].

- Sondes incluant un SNP

Il peut être nécessaire d'exclure les sondes comportant un SNP, puisque la quantification de la méthylation est basée sur le génotypage (quantitatif) de C/T (après conversion bisulfite) [Price et al., 2013; Chen et al., 2013]. En effet, des polymorphisme C/T peuvent être présents naturellement chez un individu, et ainsi être considérés comme un résultat de la conversion bisulfite. La séquence d'ADN est alors confondue avec la méthylation. Par exemple, un individu homozygote C/C aura une méthylation proche de 100 %, pendant qu'un individu homozygote T/T aura quant à lui une méthylation de 0 %. Enfin, un individu hétérozygote C/T sera à 50 % méthylique sur ce site. En l'absence de données de génomique conjointement aux données de méthylomique, il est préférable d'exclure ces sondes et les sites correspondants avant analyse.

- Valeur-p de détection

Tout comme pour les puces d'expression d'Illumina, les puces de méthylation fournissent, en plus de la quantification de la méthylation, des valeurs-p de détection pour l'ensemble des sondes/sites basées sur des sondes contrôles. Par exemple, dans l'étude présentée au Chapitre 3), une méthode dérivée du contrôle qualité appliquée en génomique a été utilisée. Ainsi, un site est exclu dès lors que la valeur-p de détection est supérieure à  $10^{-6}$  pour au moins 5 % des échantillons. Un filtre équivalent est appliqué sur les échantillons, à savoir qu'un échantillon doit présenter des valeurs-p de détection inférieures à  $10^{-6}$  pour plus de 75 % des sites pour être conservé.

- *Normalisation*

- Intra-puce : Infinium I/II

Une première normalisation est nécessaire pour corriger les différences de distribution des niveaux de méthylation entre Infinium I et Infinium II. Cette étape de normalisation Infinium I/II est indispensable; cependant, il convient de choisir la bonne méthode après avoir examiné soigneusement la distribution des valeurs de méthylation [Marabita et al., 2013; Yousefi et al., 2013].

Plusieurs méthodes ont été développées :

\* “Peak Based Correction” (PBC) [Dedeurwaerder et al., 2011] : la distribution bimodale des valeurs de méthylation, avec un pic pour les sites non-méthylés et un pic pour les sites méthylés, de l'Infinium I est utilisée comme référence pour fixer les deux modes des valeurs de méthylation de l'Infinium II.

\* “Subset quantile Within-Array Normalisation” (SWAN) [Maksimovic et al., 2012] : une normalisa-

tion quantile est appliquée à une sélection aléatoire de sondes Infinium I et Infinium II (dont la composition en nombre de sites CpG est similaire et par type de sondes de l'Infinium II). Les distributions (similaires) ainsi obtenues sont ensuite utilisées pour ajuster (interpolation linéaire) les valeurs des sondes restantes. Similaire à l'approche SWAN, l'approche "Subset Quantile Normalisation" (SQN) [Touleimat and Tost, 2012] ne diffère que par la classification des sondes, effectuée selon leur position par rapport aux îlots CpG.

\* "Beta-Mixture Quantile Normalisation" (BMQ) [Teschendorff et al., 2013] : la densité de distribution des valeurs de méthylation des Infinium I et II est chacune décomposée en un mélange de trois distributions Bêta, où chaque distribution Bêta correspond à un état de méthylation : non-méthylé (proche de 0 %), hémi-méthylé (proche de 50 %), et méthylé (proche de 100 %). Une normalisation quantile est ensuite appliquée entre chaque distribution Bêta de l'Infinium II et la distribution Bêta correspondante de l'Infinium I (Figure 22).

#### – Intra-puce : "bruit de fond"

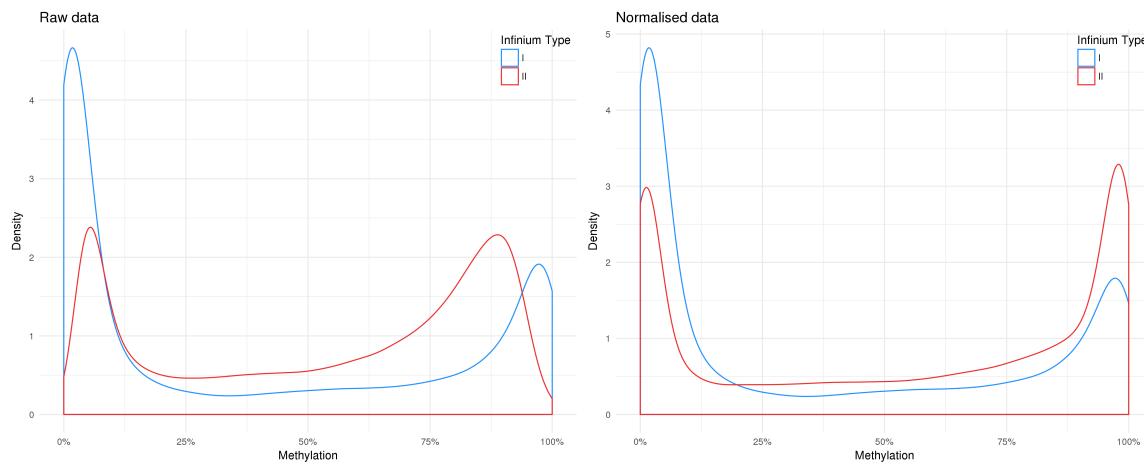
Une correction du "bruit de fond" suivant les mêmes stratégies que celles appliquées pour les puces d'expression (p. ex. modèle par convolution Normale + Exponentielle) [Xu et al., 2015; Triche et al., 2013].

#### – Inter-puce

Une étape supplémentaire de normalisation peut être appliquée pour réduire autant que possible les effets "plaques" et "puces". Des méthodes ont été développées dans cet objectif, comme la méthode *ComBat* [Johnson et al., 2007; Leek et al., 2012]), ayant déjà montré son efficacité sur des données de méthylation [Sun et al., 2011; Leek et al., 2010]. Cependant, il convient de faire attention à l'utilisation de ces méthodes, notamment lorsqu'elles permettent d'inclure des informations sur les conditions biologiques testées (p. ex. groupes cas et contrôle) [Nygaard et al., 2015]. Il est également important de garder à l'esprit que la meilleure façon d'éviter les problèmes liés aux effets "plaques" et "puces" est d'avoir réalisé un plan d'expérience ex ante, c'est-à-dire d'avoir prévu la répartition des échantillons et conditions de façon homogène entre les puces.

### c.2 Contrôle qualité des échantillons

Un premier contrôle des échantillons est réalisé lors de l'application du filtre sur les valeurs-p de détection, permettant ainsi d'exclure les individus ayant une faible qualité de détection, résultant d'une dégradation de l'ADN, par exemple. De la même façon qu'en génomique et transcriptomique, une ACP peut être réalisée dans



**FIGURE 22.** Distribution des valeurs- $\beta$  de méthylation des sondes Infinium I et II, avant normalisation BMIQ [Teschendorff et al., 2013] (à gauche) et après normalisation BMIQ (à droite).

le but d'identifier des échantillons dont le profil de méthylation est extrême, et identifier une stratification des données qui pourrait être liée à un biais technique non pris en compte.

### 3.3 Analyses omique et multi-omique

#### 3.3.1 Génomique

##### a Avant l'ère des études d'association pangénomiques

En l'absence de données omiques, et plus particulièrement de données génomiques, les études d'*agrégation*, d'*héritabilité* et/ou de *ségrégation* ont permis de mettre en avant la contribution de la génétique à de nombreuses pathologies, en se basant sur l'apparentement d'individus provenant d'une même famille. Ces études sont réalisées dans un certain ordre, consistant à évaluer en premier lieu, le caractère héritable d'un trait (binaire ou quantitatif), et en second lieu, le mode de transmission génétique d'une génération à la suivante.

- Les études d'*agrégation* (pour traits binaires, par exemple, le statut diabétique [Bijanzadeh, 2017; Weijnen et al., 2002]) ont pour but d'évaluer le risque relatif d'un trait entre les membres d'une famille [Laird et al., 2000]. Dans ces études, un individu porteur de la pathologie est dit "proband". Même si ces études permettent de montrer qu'un certain trait peut se retrouver de façon plus importante dans une famille plutôt que dans une autre (c.-à-d. *agrégation*), elles ne permettent pas d'éliminer un potentiel effet de confusion produit par l'environnement, puisque celui-ci peut être identique pour tous les individus d'une même famille.
- Les études d'*héritabilité* (pour traits quantitatifs, par exemple, la glycémie à jeun) ressemblent aux études d'*agrégation*, et cherchent à évaluer l'effet de la génétique dans la variabilité d'un trait quantitatif. On parle

de cet effet comme l'héritabilité d'un trait, laquelle donne la proportion de la variabilité totale du trait qui est expliquée par l'ensemble des variations liées à la génétique.

- Les études de ségrégation ont pour objectif d'identifier le mode de transmission génétique d'un trait, à savoir si le gène de susceptibilité suit un modèle génétique de type récessif/dominant, additif ou de codominance. Ces différents modes de transmission mendélienne s'illustrent bien en utilisant le groupe sanguin ABO (phénotype), où les allèles sont définis par A, B et O. L'allèle O est l'allèle récessif par rapport aux allèles A et B : ainsi, un génotype A/O donnera le groupe sanguin A, tandis que le génotype B/O donnera le groupe sanguin B. Le groupe sanguin AB est issu de la codominance des allèles A et B. Enfin, dû au caractère récessif de l'allèle O, un individu doit être porteur de deux allèles O (génotype O/O) pour présenter un groupe sanguin O.

Ces différentes études ont initialement été développées lorsque les techniques de génotypage étaient onéreuses, et ont servi en quelque sorte d'études préliminaires aux études d'association pangénomiques.

### **b Les études d'association pangénomiques**

Les études d'association génétiques se sont développées au cours des 20 dernières années, et viennent compléter les études de liaison qui quantifie l'excès d'allèles transmis par descendance (selon les principes mendéliens) entre des individus apparentés (principalement, du premier et second degré). Nous distinguerons deux types d'études d'association : d'une part, lorsque l'analyse porte sur des familles, et d'autre part, lorsque l'analyse porte sur des individus non apparentés. Dans le cas des études familiales, en particulier des trios correspondant à un enfant atteint/porteur et de ses deux parents, le principal test employé est le test du déséquilibre de transmission (TDT). Ce test se base sur une statistique calculée par  $(t - u)^2 / (t + u)$ , où  $t$  le nombre d'allèles transmis et  $u$  le nombre d'allèles non transmis. Sous l'hypothèse nulle d'indépendance des distributions de  $t$  et  $u$ , cette statistique suit une distribution du  $\chi^2$  à 1 degré de liberté. Dans le contexte des études familiales, ce test confère une puissance statistique plus importante que les tests d'association classiques avec design cas/témoins, et limite l'effet lié à une stratification de la population d'étude, notamment en faisant l'hypothèse d'une homogénéité phénotypique intra-famille plutôt qu'inter-famille.

#### **b.1 L'approche dite "classique"**

Les études d'association pangénomiques (GWAS), dont la toute première a été réalisée en 2005 [Klein, 2005], et la première dans le diabète de type 2 en 2007 [Sladek et al., 2007], ont été le principal moteur dans la découverte de nombreux loci de susceptibilité à diverses maladies complexes, et sont répertoriées au sein de la base de données "GWAS Catalog" [MacArthur et al., 2017]. Deux raisons principales expliquent cet essor : la première

est la limite de détection du TDT dans les maladies complexes (non monogéniques), où l'excès de transmission de chaque allèle contribuant à cette maladie est modéré voire faible. La seconde raison concerne les études de puissance réalisées au début des années 2000 [Risch and Merikangas, 1996; Sham et al., 2000] qui suggéraient un gain de puissance statistique pour détecter des associations dans les études d'association comparativement aux études de liaison. Cependant, ce gain de puissance demeure relatif, puisque celui-ci dépend de la fréquence allélique ou encore du déséquilibre de liaison entre les loci étudiés [Tu and Whittemore, 1999].

Une étude GWA consiste généralement en l'application d'un modèle de régression logistique ou linéaire généralisé, selon la nature du trait étudié, à l'ensemble des polymorphismes identifiés via une puce ADN. Le nombre de SNPs est passé de près de 300 000 à plusieurs millions, en particulier grâce à des techniques d'imputation du génome [Howie et al., 2011] à l'aide de génome de référence [The 1000 Genomes Project Consortium, 2015], permettant de ce fait de pallier aux spécificités des différentes puces, et d'accroître d'autant la quantité d'information génétique disponible pour chaque individu.

$$g[E(\mathbf{Y})] = \beta_0 + \beta_1 \mathbf{X} + \beta \mathbf{Z} \quad (1)$$

Le modèle classique de régression donné par l'Équation (1) consiste à expliquer un trait  $Y$ , par exemple le statut DT2 (binaire) ou la glycémie à jeun (quantitatif), par des SNPs avec en général une hypothèse d'additivité des allèles. Ainsi le génotype  $X$  est codé 0 pour le génotype homozygote majeur (c.-à-d. homozygote pour l'allèle avec la plus forte fréquence allélique), 1 pour le génotype hétérozygote et 2 pour le génotype homozygote mineur (Tableau 5).

La fonction  $g(.)$  est la fonction de lien, par exemple, la fonction *logit* ( $\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right)$ ) dans le cas où  $Y$  est dichotomique (régression logistique), ou encore la fonction *identité* lorsque  $Y$  est quantitatif (régression linéaire).

Ces approches sont préférées aux approches basées sur la construction de table de contingence (p. ex. test exact de Fisher), en particulier, dans les études cas/témoins, puisque ces dernières permettent l'ajustement à des covariables ( $Z$ ) cliniques ou démographiques, par exemple. En outre, l'usage d'un modèle de régression logistique permet de fournir des odds ratios à partir de la mesure de l'effet  $\beta_1$  du génotype  $X$ .

**TABLEAU 5.** Codage du génotype selon le modèle génétique.

Génotype	Codominant	Additif	Recessif	Dominant
AA	$X = (01)$	$X = 2$	$X = 1$	$X = 1$
Aa	$X = (10)$	$X = 1$	$X = 0$	$X = 1$
aa	$X = (00)$	$X = 0$	$X = 0$	$X = 0$

Des covariables d'ajustements ( $Z$ ) sont également ajoutées au modèle, particulièrement lorsque le plan d'expérience n'a pas permis de prendre en compte celles-ci. Malgré un plan d'expérience prenant en compte différents paramètres essentiellement techniques, pouvant être sources de confusion dans l'analyse, il n'est pas toujours possible de tous les contrôler dans un même plan d'expérience. Ainsi, ces variables techniques, démographiques et cliniques peuvent être incluses dans le modèle en tant que covariable, par exemple, l'âge, le sexe ou encore l'IMC. D'autres éventuelles variables peuvent être incluses, comme les premières composantes (deux à cinq) de l'ACP réalisée lors du contrôle qualité, permettant de prendre en compte une stratification ou mélange (p. ex. ethnique ou géographique) au sein de la population d'étude [Novembre et al., 2008; Clayton et al., 2005; Bouaziz et al., 2011]. L'ajustement pour ces covariables est effectué pour deux raisons principales : soit pour éviter de confondre la relation entre le génotype et le phénotype, soit pour réduire la variance résiduelle et ainsi augmenter la précision des estimations.

#### b.2 Les approches "longitudinales"

Les études GWA classiques se concentrent sur une seule mesure d'un trait quantitatif par individu pour identifier des variants génétiques, même lorsque des données longitudinales étaient disponibles, principalement à cause de la complexité des modèles permettant d'analyser ces données et en particulier, les temps élevés de calcul. L'utilisation des modèles linéaires mixtes (LMM) [Laird and Ware, 1982; Liang and Zeger, 1986], des équations estimantes généralisantes (GEE) [Ziegler et al., 1998], et d'autres approches pour prendre en compte les mesures répétées, est devenue plus fréquente dans les études GWA, avec notamment des groupes de travail (Genetic Analysis Workshop 18) dont la thématique portait sur les méthodes permettant l'analyse des données longitudinales dans un contexte génétique [Almasy et al., 2014; Beyene and Hamid, 2014; Wu and Briollais, 2014]. Les méthodes LMM et GEE permettent de prendre en compte différentes structures de données telles que les données longitudinales et familiales. Cependant, ce type de données comporte habituellement de nombreuses données manquantes, et nécessite de vérifier certaines hypothèses quant à leur distribution [Graham, 2009] :

- MCAR ("missing completely at random") : les données sont manquantes indépendamment des données observées et non observées;
- MAR ("missing at random") : conditionnellement aux données observées, les données manquantes sont indépendantes des données non observées;
- MNAR ("missing not at random") : les données manquantes sont dépendantes de variables non observées.

Dans le cas des LMM, les données manquantes doivent être distribuées selon un processus MAR ou MCAR pour obtenir une inférence statistique valide; pour les GEE (inférence non-basée sur la maximisation de la

vraisemblance), elles doivent être distribuées selon un processus MCAR [Robins et al., 1994]. De plus, l'utilisation de données longitudinales implique des hypothèses quant à la structure de corrélation entre les individus et entre les mesures de chaque individu. Une mauvaise spécification de cette structure de corrélation ou l'omission de ces hypothèses peuvent conduire à un biais dans les estimations [Lu et al., 2009]. L'une des raisons derrière le besoin d'exploiter les données longitudinales, lorsque disponibles, reposent sur le gain de puissance statistique qui peut être obtenu à partir de ces données [Costanza et al., 2012; Hossain and Beyene, 2014; Hu et al., 2014; Lee et al., 2014; Wang et al., 2014a; Xu et al., 2014; Zhao et al., 2014]. Ce gain de puissance est généralement obtenu dans les études GWA classiques par l'augmentation du nombre d'individus de façon directe ou indirecte par méta-analyse. Pour remédier et contourner le problème de complexité algorithmique induit par le volume des données génomiques, des méthodes dérivées et approchées des LMM [Sikorska et al., 2013; Verbeke et al., 2001] et des GEE [Robins et al., 1994; Sitlani et al., 2015], ainsi que des approches en "deux-étapes" [Hossain and Beyene, 2014; Houwing-Duistermaat et al., 2014; Musolf et al., 2014; Roslin et al., 2009; Sikorska et al., 2015, 2013; Wang et al., 2014a] ont été développées [Beyene and Hamid, 2014; Wu and Briollais, 2014; Kerner et al., 2009].

Les LMM ont été introduits par Laird and Ware [1982], et leur forme générale est donnée par l'équation

$$\mathbf{Y}_i = \beta \mathbf{X}_i + b_i \mathbf{Z}_i + \epsilon_i \quad (2)$$

, où  $\mathbf{Y}_i$  représentent les mesures de l'individu  $i$  ( $i = 1, \dots, n_i$ ),  $\mathbf{X}_i$  et  $\mathbf{Z}_i$  désignent les matrices respectives des effets fixes et aléatoires. Ces matrices sont de dimensions respectives  $n_i \times p$  et  $n_i \times q$ , où  $p$  et  $q$  donnent le nombre de covariables définies en effet fixe et aléatoire. Les paramètres  $\beta$  et  $b_i$  désignent les effets fixes et aléatoires pour l'individu  $i$  et  $\epsilon_i$  dénote un terme d'erreur supposé normalement distribué. La partie des effets aléatoires peut comporter une ordonnée à l'origine aléatoire, c'est-à-dire que la valeur d'origine de chaque individu peut varier selon l'individu, et inclure une pente aléatoire, ces pentes pouvant varier d'un individu à l'autre. Lorsque seule l'ordonnée à l'origine est incluse dans les effets aléatoires, la structure de corrélation est dite "compound symmetry", où les corrélations entre les mesures pour un même individu sont constantes dans le temps. Cette hypothèse est peu réaliste, en particulier dans les données cliniques et épidémiologiques; conséquemment, une pente est généralement incluse dans les effets aléatoires, permettant ainsi d'avoir une structure de corrélation plus complexe, tel qu'un processus autorégressif d'ordre 1 (AR1), par exemple. Dans un contexte d'étude GWA, le modèle général (en omettant la matrice des covariables) s'écrit sous la forme suivante :

$$Y_{ij} = \beta_0 + b_{0i} + \beta_1 t_{ij} + b_{1i} t_{ij} + \beta_2 G_i + \beta_3 G_i t_{ij} + \epsilon_{ij} \quad (3)$$

où  $G_i$ , le génotype d'un SNP pour l'individu  $i$ , et  $t_{ij}$  désigne le temps de la mesure  $j$  de l'individu  $i$ . Néanmoins, en raison de la complexité computationnelle, ces modèles sont habituellement appliqués sur des ensembles réduits de SNPs (p. ex. chromosome, gènes ou SNPs candidats), et parfois en omettant le terme représentant la pente aléatoire [Hu et al., 2014; Wu and Briollais, 2014; Mei et al., 2012; Lee et al., 2014; Liu et al., 2014; Xu et al., 2014; Smith et al., 2010].

Une approche dérivée des LMM a également été proposée sous le nom de LMM conditionnel (cLMM) [Verbeke et al., 2001]. Ce modèle, en plus d'estimer les effets longitudinaux (c.-à-d. les effets sur l'ensemble des mesures d'un individu), garantit une plus grande robustesse relativement à une mauvaise spécification des caractéristiques de la mesure à l'origine (c.-à-d. la première mesure d'un individu). Le cLMM peut s'écrire, à partir de l'équation (2) :

$$\mathbf{Y}_i = \beta^1 \mathbf{X}_i^1 + \beta^2 \mathbf{X}_i^2 + b_i^1 \mathbf{Z}_i^1 + b_i^2 \mathbf{Z}_i^2 + \epsilon_i \quad (4)$$

Dans ce modèle, les effets fixes ( $\mathbf{X}_i$ ) et aléatoires ( $\mathbf{Z}_i$ ) sont décomposés respectivement en :

- $\mathbf{X}_i = (\mathbf{X}_i^1 | \mathbf{X}_i^2)$ , où  $\mathbf{X}_i^1$  représente la matrice des covariables indépendantes du temps ( $n_i \times p_1$ ) et  $\mathbf{X}_i^2$  la matrice des covariables dépendantes du temps ( $n_i \times p_2$ );
- $\mathbf{Z}_i = (\mathbf{Z}_i^1 | \mathbf{Z}_i^2)$ , où  $\mathbf{Z}_i^1 = \mathbf{1}_{(n_i)}$  et  $\mathbf{Z}_i^2$  représente la matrice des covariables dépendantes du temps ( $n_i \times (q - 1)$ ).

L'évolution temporelle des traits cliniques mesurés aux différentes visites peut être modélisée par une droite, ce qui a motivé les approches dites en "deux étapes". Elles consistent à utiliser un modèle "simplifié", c'est-à-dire sans la variable d'intérêt (SNP testé) en premier lieu, puis utiliser l'un des paramètres estimés dans un second modèle en incluant la variable d'intérêt, par exemple, et en prenant comme variable réponse la pente [Sikorska et al., 2013], l'ordonnée à l'origine [Wang et al., 2014b] ou les résidus [Hossain and Beyene, 2014]. Le modèle s'écrit alors :

$$Y_{ij} = \beta_{0i}^\Delta + \beta_{1i}^\Delta t_{ij} + \epsilon_{ij}^\Delta \quad (5)$$

Enfin, certaines approches utilisent des modèles à classes latentes en "deux étapes" [Roslin et al., 2009; Musolf et al., 2014]. L'idée principale consiste à réduire l'information du trait quantitatif à un trait qualitatif reposant, par exemple, sur les probabilités bayésiennes a posteriori (probabilité de chaque individu d'appartenir à un groupe ou sous-groupe), et d'inclure ces probabilités comme variables réponses dans un second modèle. Cette méthode possède l'avantage de réduire le temps de calcul en diminuant leur complexité.

Une alternative aux LMM est l'approche GEE [Liang and Zeger, 1986]. Il s'agit d'une méthode semi-paramétrique dont l'objectif est l'inférence de l'effet moyen sur la population d'une variable d'intérêt. Cette méthode nécessite de vérifier d'abord la nature de la distribution des données manquantes, ainsi que de définir la "bonne" structure de corrélation des mesures intra-individuelles. Elle peut s'écrire, dans sa formulation la plus simple, comme :

$$E(\mathbf{Y}_{it}) = \beta_0 + \beta_1 \mathbf{X}_{it} + \gamma \mathbf{Z}_{it} \quad (6)$$

Contrairement aux approches LMM, les GEE nécessitent que les données manquantes soit distribuées selon un processus MCAR. Des améliorations ont été proposées pour réduire l'impact du non-respect de cette hypothèse, et rendre de ce fait les GEE plus robustes aux distributions des données manquantes [Robins et al., 1995, 1994]. Cependant, dans le contexte des études GWA, la violation de cette hypothèse pourrait ne pas être un problème, puisque les données manquantes ne peuvent s'expliquer par un seul variant génétique avec effet faible [Sitolani et al., 2015].

Les données longitudinales offrent, en plus de pouvoir modéliser l'évolution de différents traits quantitatifs, la possibilité d'étudier la survenue d'un événement, en particulier le développement d'une pathologie comme le diabète de type 2. Dans ce contexte, les études GWA ont dans un premier temps réalisé des tests d'associations au moyen d'une régression logistique; cependant, cette approche ne permet pas de prendre en compte la composante longitudinale et se limite à une seule mesure du trait. Les modèles de survie, tel le modèle de Cox, se présentent comme des alternatives au modèle de régression logistique. Ces approches commencent à se développer et à être optimisées pour application sur données réelles en génétique [Syed et al., 2016a,b].

La modélisation conjointe permet de modéliser d'une part, la composante longitudinale (c.-à-d. la trajectoire de la variable étudiée) et d'autre part, la composante de survie (c.-à-d. la survenue de l'événement étudié). La composante longitudinale consiste typiquement à l'application d'un modèle linéaire mixte :

$$Y_{ij} = X_{ij} + \epsilon_{ij} \quad (7)$$

où  $Y_{ij}$  est la valeur observée et  $X_{ij}$  la vraie (non observée) valeur de la variable longitudinale. Le terme  $\epsilon_{ij}$  est le terme d'erreur aléatoire supposé distribué selon la loi Normale :

$$\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \quad (8)$$

La quantité  $X_{ij}$  représente la fonction de la trajectoire et est définie usuellement comme une fonction linéaire

(ou quadratique) dépendante du temps. Des covariables peuvent aussi être incluses dans cette fonction, comme l'âge, le sexe ou l'IMC. Par exemple, si  $Y_{ij}$  représente les valeurs mesurées de la glycémie à jeun au temps  $t_{ij}$ ,  $Z_i$  désigne le génotype du SNP analysé pour l'individu  $i$ , et  $W_i$  désigne les covariables, le modèle s'énonce :

$$Y_{ij} = X_{ij} + \epsilon_{ij} = \theta_{0i} + \theta_{1i} \times t_{ij} + \gamma \times Z_i + \delta \times W_i + \epsilon_{ij} \quad (9)$$

Pour simplifier l'écriture dans ce qui suit, le terme  $\delta \times W_i$  sera omis. Les paramètres  $\theta_{0i}$  et  $\theta_{1i}$  sont réputés distribués selon une distribution Normale bivariée :

$$\theta \sim \mathcal{N}_2(\mu, \Sigma) \quad (10)$$

Le paramètre  $\gamma$  évalue l'effet de  $Z_i$  (p. ex. effet additif du SNP) sur la fonction de la trajectoire. Pour tenir compte éventuellement de pentes différentes entre les génotypes, un terme d'interaction entre  $Z_i$  et le temps peut être inclus dans la fonction de la trajectoire.

La composante de survie (p. ex. survenue du DT2) se compose généralement d'un modèle paramétrique (p. ex. Exponentielle ou Weibull) ou semi-paramétrique (p. ex. risques proportionnels de Cox) avec :

$$h_i(t) = h_0(t) \exp(\beta X_i(t) + \alpha Z_i) \quad (11)$$

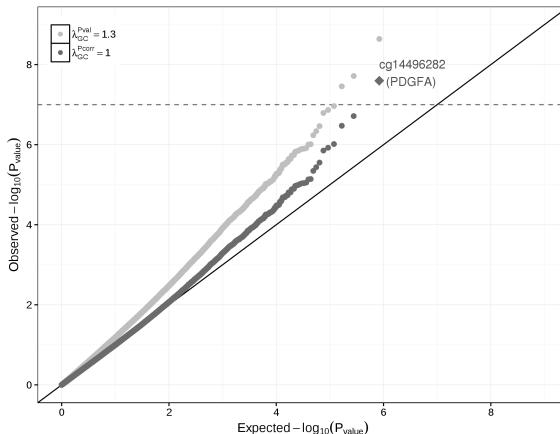
où  $h_i(t)$  est la fonction de risque au temps  $t$  pour l'individu  $i$ , et  $h_0(t)$  est la fonction de risque de base non spécifiée (Tableau 6). Le coefficient  $\alpha$  mesure l'effet de  $Z_i$  sur le temps de survenue de l'événement, alors que le coefficient  $\beta$  mesure l'association entre la trajectoire  $Y_i$  et le temps de survenue.

Le choix d'un modèle joint est généralement motivé par la volonté de modéliser le lien entre les variables d'intérêt de chaque composante (c.-à-d. variable longitudinale et variable événement), d'évaluer le mécanisme sous-jacent aux données manquantes, et de prendre en compte une covariable dépendante du temps [Sudell et al., 2016]. Les modèles principalement employés dans cette approche sont les modèles linéaires mixtes et les modèles de Cox, ceux-ci étant déjà implémentés au sein d'extensions du logiciel R [Rizopoulos, 2016a; Philipson et al., 2017; Rizopoulos, 2016b].

**TABLEAU 6.** Caractéristiques des distributions exponentielle, de Weibull et de Comprtetz sous le modèle de Cox avec  $h(t) = h_0(t) \exp(\beta X)$ .

Caractéristique	Exponentielle	Weibull	Comprtetz
Paramètre d'échelle	$\lambda > 0$	$\lambda > 0$	$\lambda > 0$
Paramètre de forme		$\nu > 0$	$-\inf < \alpha < \inf$
Fonction de risque	$h_0(t) = \lambda$	$h_0(t) = \lambda \nu t^{(\nu-1)}$	$h_0(t) = \lambda \exp(\alpha t)$
Fonction de risque cumulée	$H_0(t) = \lambda t$	$H_0(t) = \lambda t^\nu$	$H_0(t) = \frac{\lambda}{\alpha} (\exp(\alpha t) - 1)$
Fonction inverse de risque cumulée	$H_0^{-1}(t) = \lambda^{-1}t$	$H_0^{-1}(t) = (\lambda^{-1}t)^{(1/\nu)}$	$H_0^{-1} = \frac{1}{\alpha} \log(\frac{\alpha}{\lambda}t + 1)$
Temps d'événement	$T = -\frac{\log(u)}{\lambda \exp(\beta X)}$ $(u \sim Uniforme(0,1))$	$T = \left(-\frac{\log(u)}{\lambda \exp(\beta X)}\right)^{(1/\nu)}$	$T = \frac{1}{\alpha} \left(1 - \frac{\alpha \log(u)}{\lambda \exp(\beta X)}\right)$

### b.3 Correction post-analyse



**FIGURE 23.** Graphique quantile-quantile des valeurs-p de l'étude d'association épigénétique en cas/contrôle sur le diabète de type 2 (Chapitre 3), avant et après normalisation sur le facteur d'inflation  $\lambda$ . La ligne horizontale en pointillé représente le seuil de significativité corrigé selon la méthode de Bonferroni.

Une fois l'analyse réalisée au moyen de logiciels tels que PLINK [Chang et al., 2015; Purcell and Chang, 2015], SNPTEST [Burton et al., 2007; Clark and Li, 2007] ou d'extensions du logiciel R [Huber et al., 2015], une correction appelée contrôle génomique (“Genomic control”) peut être appliquée, avec comme principale motivation la correction d'une éventuelle stratification ou mélange au sein de la population d'étude. Cette inflation est mesurée par le paramètre de sur-dispersion  $\lambda$  sur l'ensemble des statistiques de test observées [Devlin and Roeder, 1999]. Il est estimé à partir de la distribution observée des m statistiques de test de chi-deux comparée

à celle attendue sous l'hypothèse nulle. Pour se prémunir des valeurs extrêmes, la valeur médiane (et non la moyenne) est utilisée  $\lambda = \frac{\text{median}(\chi_1^2, \dots, \chi_m^2)}{0,4549}$ . La distribution de la statistique de test observée dans l'étude est ensuite corrigée par ce facteur  $\lambda$  (Figure 23). Il est à noter que même si le paramètre  $\lambda$  n'est pas directement relié à la fréquence allélique, cette correction peut toutefois engendrer une perte de puissance [Georgiopoulos and Evangelou, 2016]

Les études GWA consistant en la réalisation d'un grand nombre de répétitions d'un même test statistique, il convient d'appliquer une correction pour test multiples afin de diminuer le taux de faux-positifs global de l'étude [Pearson, 2008]. Un test statistique est dit significatif, et son hypothèse nulle rejetée, si la valeur-p est inférieure à un seuil  $\alpha$  déterminé en amont de l'analyse (communément fixé à 0,05). En d'autres mots, on admettra que l'hypothèse nulle est rejetée à tort dans 5 % des répétitions (taux de faux-positifs). Ce seuil de 5 % est applicable sur un seul test statistique; pour un grand nombre de tests, la probabilité cumulée d'observer un ou plusieurs faux-positifs augmente avec le nombre de tests. Depuis la première étude GWA, la méthode de correction du seuil  $\alpha$ , choisie préférentiellement en raison de sa simplicité, est la méthode de Bonferroni [Dunn, 2012]. Cette méthode suppose que les tests effectués sont indépendants entre eux, et consiste à diviser le seuil  $\alpha$  par le nombre de tests. Par exemple, pour 500 000 tests à un seuil de 5 %, on obtient  $\alpha_{cor} = 0,05/(500\,000) = 1 \times 10^{-7}$ . Cependant, l'hypothèse d'indépendance n'est pas vérifiée en raison de la présence connue de déséquilibre de liaison entre les SNPs étudiés. D'autres méthodes de correction existent et ont été utilisées, tels le FDR ("False Discovery Rate") qui consiste à estimer le taux de faux-positifs parmi tous les résultats significatifs à un seuil pré-défini  $\alpha$  [Hochberg and Benjamini, 1990; van den Oord, 2008], ou encore les méthodes de permutation comme celles développées au sein du logiciel PLINK [Chang et al., 2015; Purcell and Chang, 2015]. À l'heure actuelle, le seuil nominal communément admis de significativité est  $\alpha_{cor} = 5 \times 10^{-8}$ . Ce seuil de significativité pangénomique ("genome-wide significance"), a été établi en considérant une étude avec individus d'origine caucasienne [Dudbridge and Gusnanto, 2008] et réalisée à une échelle pangénomique d'environ 1 million de SNPs indépendants.

Enfin, les résultats d'une étude GWA, après la découverte de nouveaux loci, peuvent faire l'objet d'une validation ou d'une réPLICATION, c'est-à-dire que le même modèle (ou suffisamment proche) est appliqué dans une population indépendante à celle de l'étude initiale, mais dont les caractéristiques populationnelles sont similaires. Pour qu'un résultat soit validé, certaines caractéristiques doivent être obtenues dans l'étude de réPLICATION, telles :

- une taille d'échantillon suffisamment grande avec puissance statistique estimée à au moins 80 %, permettant ainsi de limiter le taux de faux-négatifs et d'augmenter la réPLICATION des "vrais-positifs" de l'étude initiale;

- un effet dans la même direction (p. ex. effet positif), présentant une valeur-p inférieure au seuil nominal  $\alpha$  de 5 %.

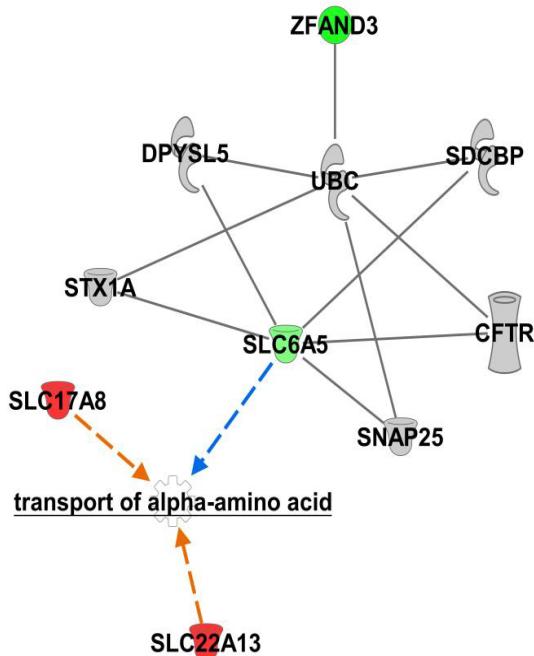
Ces études de réPLICATION, même si elles ne sont pas obligatoires, permettent tout de même de réduire le nombre de faux-positifs rapportés dans les résultats. Une autre approche de validation concerne les méta-analyses [Zeggini and Ioannidis, 2009; Evangelou and Ioannidis, 2013]. Ces méthodes permettent, soit en agrégeant les valeurs-p (méthode de Fisher) de plusieurs études (l'étude initiale n'étant pas incluse pour éviter d'orienter les résultats), soit en utilisant les effets estimés et leur erreurs-types (permettant de pondérer les effets selon la taille de l'échantillon de l'étude), de proposer une valeur-p globale et transversale à plusieurs études, indiquant par le fait même si le résultat de l'étude initiale est validé ou non. L'hétérogénéité des études incluses dans la méta-analyse doit être prise en compte afin d'éviter de tirer des conclusions erronées créées principalement par les différences entre celles-ci.

### 3.3.2 Transcriptomique

Une fois complété le pré-traitement et le contrôle qualité des mesures de fluorescence adaptés à la plateforme utilisée (p. ex. Affymetrix, Illumina ou Agilent), l'approche classique consiste à identifier des gènes différentiellement exprimés selon une ou plusieurs conditions expérimentales. En effet, l'objectif premier des puces d'expression est d'identifier des processus biologiques ou voies biologiques permettant de discriminer deux ou plusieurs groupes. Les différentes puces permettent de quantifier l'expression (quantité de mRNA) pour l'ensemble des gènes sur le génome et les différents transcrits de ces gènes. Cependant, d'une plateforme à une autre, les informations sur les transcrits diffèrent aussi bien en termes de quantification que d'annotation des sondes, c.-à-d. la localisation des sondes par rapport aux transcrits. Il est donc difficile de comparer les mesures d'expression entre des plateformes différentes.

La détection de gènes différentiellement exprimés s'effectue généralement au moyen d'un modèle de régression linéaire généralisé appliqué à chaque gène individuellement. En plus de proposer un cadre commun d'analyse des données issues des différentes plateformes [Smyth et al., 2017], l'extension R *limma* propose une amélioration du modèle en incluant, d'une part, une composante hiérarchique permettant de prendre en compte la variabilité inter-puce, et d'autre part, une modification de la statistique de test en régulant l'erreur-type de l'estimateur par un paramètre de variance de chaque gène sur l'ensemble des puces et des échantillons (paramètre estimé par une approche bayésienne empirique). Cette amélioration accroît la stabilité de l'inférence comparée à celle de l'approche classique obtenue par régression linéaire [Smyth, 2004; Phipson et al., 2016]. Afin d'identifier les gènes différentiellement exprimés, le seuil de significativité  $\alpha$  doit être défini en

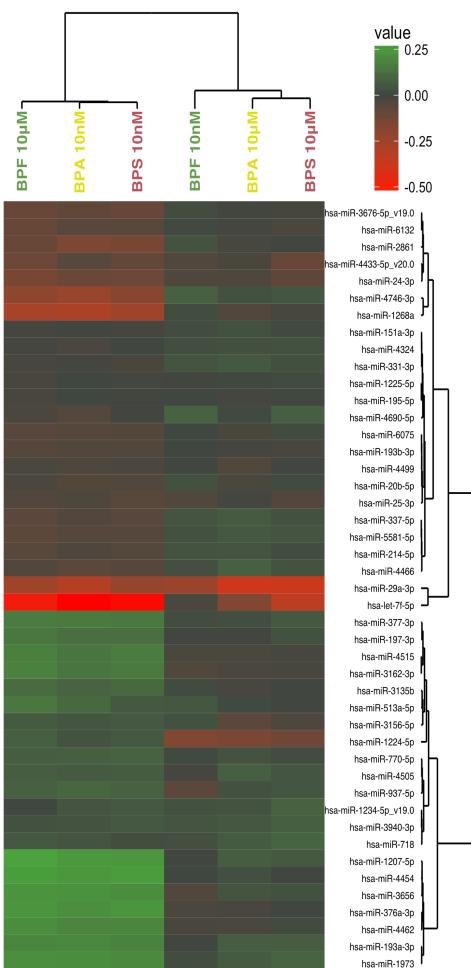
considérant le nombre de tests réalisées, par exemple, au moyen d'une correction sur le FDR (option par défaut dans (option par défaut dans *limma*).



**FIGURE 24.** Identification par IPA d'un réseau de gènes associé à la diminution de l'expression de *ZFAND3* (Chapitre 2).

À l'issue de l'analyse, un grand nombre de gènes peut alors être différemment exprimé. A priori, il est difficile d'identifier parmi ces gènes, une fonction ou une voie métabolique commune qui pourrait éclairer sur l'aspect biologique de ces résultats. Afin de réduire l'information et identifier des groupes de gènes impliqués dans une fonction particulière, une approche consiste à effectuer des tests d'enrichissement, c'est-à-dire à évaluer si l'ensemble des gènes présente en excès un groupe de gènes particuliers liés à une fonction ou à une localisation dans un tissu particulier. Ces études d'enrichissement peuvent se faire à partir de différentes bases de données telles que "Gene Ontology" [Ashburner et al., 2000], KEGG ("Kyoto Encyclopedia of Genes and Genomes") [Kanehisa et al., 2017] ou encore depuis des logiciels d'analyse ayant accès à leurs propres bases de connaissance comme "Ingenuity Pathway Analysis" (IPA) (Figure 24) et "Gene Set Enrichment Analysis" (GSEA) [Subramanian et al., 2005].

Une autre approche consiste à utiliser des méthodes de classification hiérarchique représentées sous la forme de "heatmap" comportant deux dendrogrammes, l'un représentant les dissimilarités entre les échantillons, et l'autre celles entre les gènes/transcrits (Figure 25). Cette méthode permet de visualiser rapidement et simplement les groupes de gènes différemment exprimés entre plusieurs conditions expérimentales.



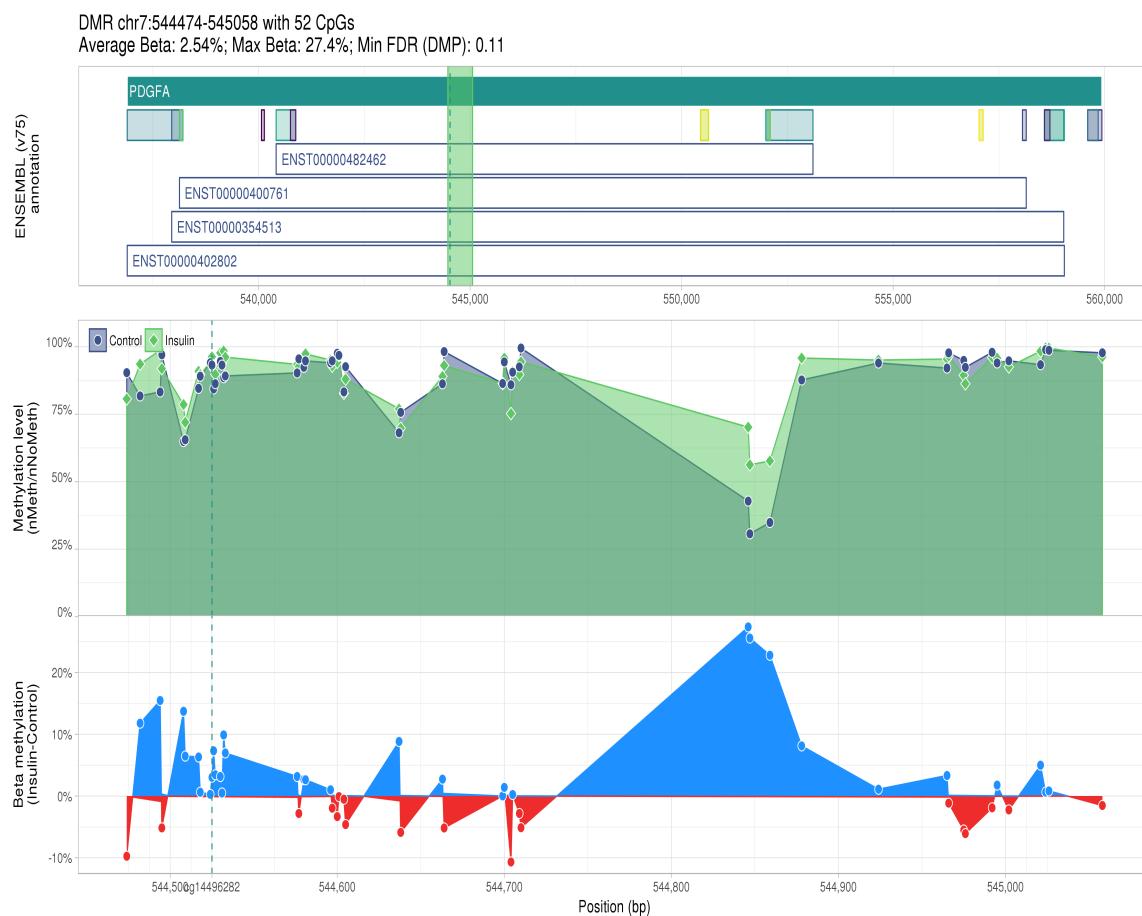
**FIGURE 25.** Représentation “heatmap” des micro ARN différemment exprimés entre 10 nM et 10  $\mu$ M, dans les 3 bisphénols (A, F et S). Les valeurs représentent les  $\log_2$  des “Fold Change” (c.-à-d. le ratio de l’expression d’un miRNA, pour un bisphénol donné, sur la condition contrôle, sans exposition bisphénol) [Verbanck et al., 2017].

### 3.3.3 Méthylomique

Après le pré-traitement des données provenant des puces de méthylation (c.-à-d. après filtrage des sondes problématiques et normalisation des différents biais techniques), il est possible d’analyser ces données généralement obtenues selon un plan d’expérience de type cas/témoins.

Lorsque le plan d’expérience le permet, c'est-à-dire lorsque les individus sont appariés entre cas et témoins, un test-t ou un test non-paramétrique de Mann-Whitney peut être appliqué à chaque sonde. Dans le cas contraire, les mêmes covariables d’ajustement que dans les études GWA sont incluses dans un modèle de régression (logistique ou linéaire), à savoir l’âge, le sexe, l’IMC, voire des composantes principales. Une sonde est une position différemment méthylée (“Differentially Methylated Position”, DMP) lorsque la valeur-p du test est inférieure au seuil nominal de  $\alpha = 0,05$  ou corrigé pour les tests multiples (c.-à-d. Bonferroni, FDR, etc.). Les données extraites depuis le logiciel GenomeStudio représentent les niveaux de méthylation mesurés à

chaque sonde, et rapportés en tant que valeur- $\beta$  (suit une distribution Bêta sous l'hypothèse que les intensités sont distribuées selon une loi Gamma) définie comme  $\beta = \frac{M}{(M+U+\phi)}$ , où  $M$  est l'intensité de l'allèle méthylé,  $U$  est l'intensité de l'allèle non-méthylé et  $\phi$  une constante pour compenser des intensités sont faibles. Les valeurs- $\beta$  varient de zéro (complètement non-méthylé) à un (complètement méthylé). Les caractéristiques de la distribution de ces valeurs n'en font pas une bonne variable pour les tests statistiques (p. ex. test-t et régression linéaire) qui s'appuient sur l'hypothèse d'homoscédasticité de la variable (variance constante). La transformation des valeurs- $\beta$  en valeurs- $M$  ( $M = \log_2 \left( \frac{M+\phi}{U+\phi} \right)$ ) a été proposée et étudiée pour contourner cet écueil [Du et al., 2010]. En raison de l'interprétation biologique simple des valeurs- $\beta$ , lesquelles sont exprimées en pourcentages, mais de la "qualité" statistique des valeurs- $M$ , les analyses statistiques sont réalisées sur les valeurs- $M$ , et ce sont les valeurs- $\beta$  qui sont reportées.

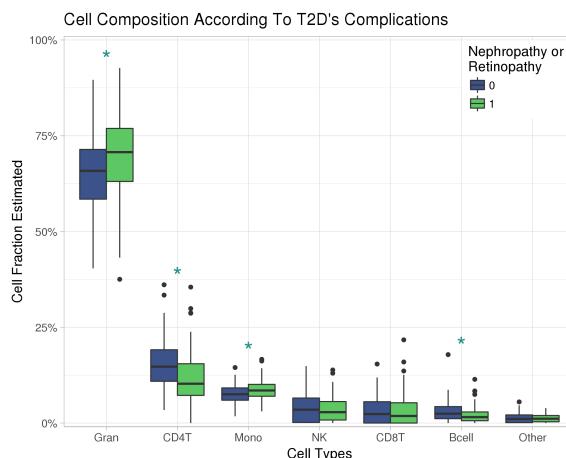


**FIGURE 26.** Représentation de la méthylation du gène *PDGFA* et du locus CpG cg14496282 en présence d'insuline (Chapitre 3).

Une seconde approche peut être réalisée en considérant non plus chaque sonde de façon individuelle, mais en considérant des régions ou blocs, permettant de ce fait d'identifier des régions différentiellement méthylées ("Differentially Methylated Region", DMR) [Peters et al., 2015; Hansen et al., 2011, 2012] (Figure 26). Cette approche se base sur l'hypothèse que des sondes dans un même voisinage auraient le même comportement

en termes de méthylation : par exemple, une région hypométhylée ou méthylée au niveau du promoteur de la transcription d'un gène. Cependant, en raison de la couverture relativement faible et non-homogène du génome par la puce Illumina HumanMethylation450 [Bibikova et al., 2011], cette approche ne permet d'étudier que partiellement ces régions, et est destiné principalement aux techniques de séquençage (p. ex. séquençage bisulfite) ou aux puces bénéficiant d'une plus grande couverture, comme c'est le cas dans les dernières puces Illumina HumanMethylation850 [Moran et al., 2016].

Comme dans les études GWA, une validation ou une réPLICATION des résultats est préconisée d'autant plus que les biais techniques (p. ex. types de sonde, effet plaques, etc.) représentent des facteurs majeurs menant à une inflation du taux de faux-positifs. Une première approche consiste à réaliser une réPLICATION avec la même technologie sur une cohorte indépendante présentant des caractéristiques similaires. Une seconde approche consiste à réaliser une validation d'un DMP ou DMR à l'aide d'une autre technologie (p. ex. mesurer la méthylation d'un DMP/DMR via une technique ciblée comme le pyroséquençage) sur la même population que l'étude initiale [Kurdyukov and Bullock, 2016]. Cette dernière présente l'avantage de pouvoir valider ce qui a été observé dans une population donnée avec une technologie donnée, et permet en conséquence de réduire le risque de faux-positifs imputable à un facteur technique, mais ne permet toutefois pas d'éliminer une potentielle spécificité de la population étudiée (p. ex. biais de sélection). En revanche, la première approche permet de contrôler à la fois ces deux phénomènes, lorsque la population de réPLICATION est adaptée.



**FIGURE 27.** Estimation des différences de composition cellulaire (selon une base de référence) entre le groupe diabétique et le groupe contrôle.

Il est à noter que la méthylation diffère entre les tissus, et plus précisément entre les types cellulaires. Les échantillons provenant généralement de prélèvements sanguins, voire de biopsies, la composition cellulaire du prélèvement, c.-à-d. le nombre et la proportion des types cellulaires qui composent le tissu, peut ne pas être homogène d'un individu à l'autre, et par conséquent, d'un groupe à l'autre. Ceci peut représenter un facteur de confusion avec le statut cas/témoin d'une étude. Il convient alors de considérer l'étude des valeurs- $\beta$  à l'aide

d'une base de référence [Houseman et al., 2012; Teschendorff et al., 2017; Cardenas et al., 2016] (Figure 27) ou de façon algorithmique, comme s'il s'agissait d'un mélange de distributions, par exemple, ou en effectuant une décomposition de celles-ci sur le même principe qu'une ACP [Houseman et al., 2015, 2016, 2014]. Cette différence potentielle de composition cellulaire peut également être corrigée de la même façon que dans les études GWA, au regard de la stratification de la population, avec notamment l'application d'une correction post-analyse comme le contrôle génomique, simultanément ou en plus de sa prise en compte en tant que covariable d'ajustement dans le test d'association.

### 3.3.4 Multi-omique

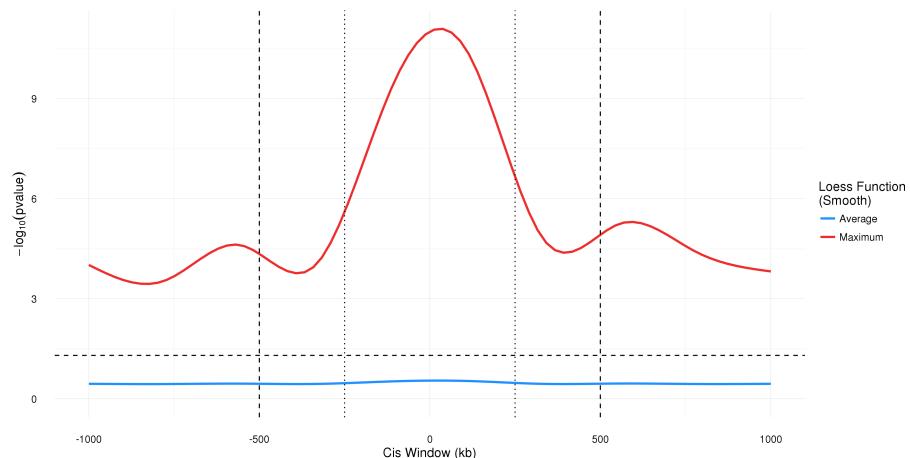
Lorsque les données de différentes omiques sont disponibles pour un ensemble d'individus, il peut être intéressant d'intégrer ces différentes sources d'information afin d'améliorer et de parfaire la compréhension d'une association mise en avant lors des analyses décrites précédemment. Lorsqu'un gène/SNP/CpG est identifié comme candidat pour la susceptibilité au trait d'intérêt, des analyses complémentaires et ciblées peuvent être réalisées et pourront être étudiées de façon conjointe. La méthode employée le plus souvent consiste alors soit à mesurer les corrélations entre les différentes données omiques, soit à inclure cette information additionnelle en tant que covariable dans le modèle de régression. Lorsque des puces ADN sont employées et qu'aucune hypothèse ou sélection *a priori* n'est réalisée, le volume de données devient très important et difficile à intégrer pour l'ensemble des trois omiques discutées dans ce manuscrit (Tableau 7). Quelques méthodes exploitant deux omiques différentes seront rapidement détaillées dans la suite. Heureusement, des approches de type "machine learning" commencent à voir le jour pour analyser tous les types de données omiques simultanément [Lin and Lane, 2017], données provenant de puces ou issues de méthodes de séquençage à haut-débit dit de nouvelle génération (*Next Generation Sequencing*).

**TABLEAU 7.** Plateformes omiques (puce-à-ADN) utilisées dans les Chapitres 1, 3 et 4.

Plateforme	Nombre de sondes/marqueurs	Omique
Illumina HumanHT-12	~50 000	Transcriptomique
Agilent SurePrint G3 Human mRNA/lncRNA Microarray	~50 000	Transcriptomique
Agilent SurePrint G3 Human miRNA Microarray	~2 500	Transcriptomique
Illumina Cardio-Metabochip	~200 000	Génomique
Illumina HumanMethylation450	~480 000	Methylomique

### a eQTL: “expression Quantitative Trait Loci”

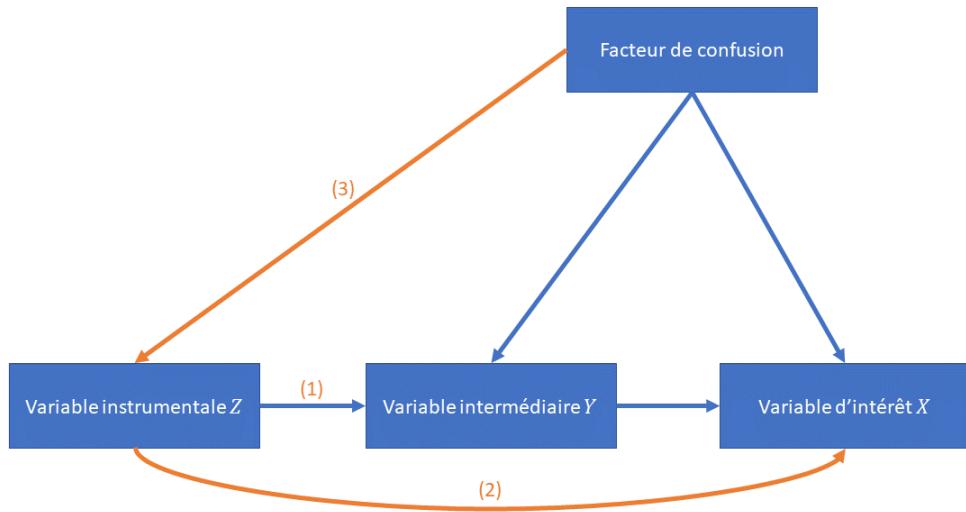
Quoique les études GWA aient permis d'identifier des loci de susceptibilité, le mécanisme reliant ces SNPs à la maladie n'est pas évident. Les analyses dites eQTL pour “expression Quantitative Trait Loci” [Rockman and Kruglyak, 2006; Gibson and Weir, 2005] proposent d'intégrer l'information d'expression des gènes à ces loci de susceptibilité de façon à identifier une relation de causalité entre génomique et transcriptomique [Rockman and Kruglyak, 2006; Gibson and Weir, 2005]. Ces analyses eQTL considèrent la mesure d'expression d'un gène/transcrit en tant que trait quantitatif sur lequel le génotype d'un SNP, ainsi que des covariables, seront régressés pour évaluer la relation entre le génotype et l'expression. Les combinaisons de SNPs (environ 1 000 000) et de gènes (environ 20 000), pouvant aboutir à un très grand nombre de configurations possibles ( $\simeq 10^{10}$ ), ces analyses soulèvent notamment des problèmes computationnels [Mackay et al., 2009].



**FIGURE 28.** Distribution des valeurs-p d'association entre les SNPs et l'expression des gènes à moins de un mégabase du promoteur du gène.

Les tests peuvent être effectués en *cis*, c'est-à-dire qu'un gène est testé pour un SNP si celui-ci est localisé à moins d'une certaine distance en paires de bases (en général, entre 100 kb et 1 Mb) du gène, ou en *trans* lorsque le SNP excède la distance définie [Rockman and Kruglyak, 2006; Cheung and Spielman, 2009]. Une façon de définir la distance est de réaliser l'analyse sur un sous-ensemble (aléatoire) des combinaisons à tester et d'étudier la répartition du signal selon la distance en paire de base (Figure 28). Les SNPs en *trans* ne sont, en raison de contraintes computationnelles ou d'interprétations difficiles, généralement pas analysés. En effet, l'hypothèse courante suppose qu'un SNP a une probabilité plus élevée de produire un effet sur l'expression d'un gène s'il se trouve dans le corps du gène, ou en amont, c'est-à-dire proche des promoteurs de la transcription (“Transcription Start Site” ou TSS). Par extension, cette approche eQTL est applicable aux données d'épigénomique et connue sous le nom de meQTL (“methylation Quantitative Trait Loci”) ou mQTL (utilisé parfois comme “metabolomic Quantitative Trait Loci”), avec en lieu et place de l'expression du gène, le niveau de méthylation d'un site CpG.

### b Causalité: la randomisation mendélienne



**FIGURE 29.** Représentation schématique de la randomisation mendélienne. La randomisation mendélienne peut être utilisée pour tester l'hypothèse selon laquelle  $Y$  est causale de  $X$  selon que les conditions (1), (2) et (3) soient remplies par la variable instrumentale  $Z$ , où (1)  $Z$  est associée à  $Y$ , (2)  $Z$  n'est pas associée à  $X$ , et (3)  $Z$  n'est pas associé à des facteurs de confusions mesurés ou non.

En combinant les résultats des études GWA et eQTL, il est possible d'évaluer la causalité d'un SNP sur une pathologie en considérant que, si un SNP est associé au risque de développement d'une pathologie et que ce même SNP est également associé à des changements de l'expression d'un gène (*cis*), alors il est probable que le SNP soit causal de la pathologie [Nica et al., 2010; Schadt et al., 2005; Zhu et al., 2007; Chang et al., 2016]. Cette idée sous-tend la randomisation mendélienne [Lawlor et al., 2008], où l'expression d'un gène peut être considérée comme un phénotype intermédiaire ( $Y$ ) entre le génotype (variable instrumentale,  $Z$ ) et le phénotype d'intérêt ( $X$ ) (p. ex. statut diabétique, glycémie, insulinémie, etc.) (Figure 29).

En général, même s'il est possible de tester l'association entre deux variables ( $X$  et  $Y$ ), il n'est pas possible de répondre quant à la direction de cette relation : cause ou conséquence ? En outre, l'association entre ces deux variables peut être le résultat de facteurs de confusion, mesurés ou non, ayant un effet sur une ou les deux variables, causant un biais dans l'estimation de cette association. Pour évaluer si la variable intermédiaire  $Y$  est causale de la variable d'intérêt  $X$ , la randomisation mendélienne consiste à utiliser une variable instrumentale  $Z$ , répondant aux conditions suivantes : 1)  $Z$  est associée à la variable intermédiaire  $Y$ , 2)  $Z$  ne présente pas d'association (directe) avec la variable d'intérêt  $X$  et 3)  $Z$  ne présente pas d'association avec des facteurs de confusion (mesurés ou non) (Figure 29). Le génotype d'un SNP ou un score agrégé de génotypes (p. ex. score de

risque génétique) remplit ces conditions et permet de déduire la direction de la causalité, puisque la génétique exerce un effet partiel ou total sur les traits biologiques, et non l'inverse. De plus, la mesure du génotype n'est soumise qu'à peu de variabilité et de biais. Grâce à la structure de LD entre les variants, il n'est pas obligatoire d'utiliser le SNP causal du trait; un variant en fort LD avec celui-ci peut être utilisé de façon équivalente.

---

## ***Objectifs & Plan***

---

### **Objectifs & Contexte**

Ce travail de thèse a été conduit au sein de l'unité de “Génomique Intégrative et Modélisation des Maladies Métaboliques” UMR 8199 (CNRS / Université de Lille 2 / Institut Pasteur de Lille) sous la direction du Pr. Philippe Froguel et du Dr. Ghislain Rocheleau. Le laboratoire de “Génomique Intégrative et Modélisation des Maladies Métaboliques”, membre de la fédération de recherche “European Genomic Institute for Diabetes” (EGID), est un acteur majeur dans l'étude de la génétique du diabète et de l'obésité, notamment par ces nombreuses publications (245 publications entre 2007 et 2017, référencées via l'outil SAMPRA de l'Université de Lille) et la publication de la première étude d'association pangénomique sur le diabète de type 2 en 2007 [Sladek et al., 2007]. L'unité a développé une expertise en génomique, transcriptomique, études fonctionnelles (modèles animaux et cellulaires), et depuis quelques années en épigénomique, le volume et la variété des données générées par l'unité ont donc augmenté, nécessitant un développement méthodologique en amont (p. ex. plan d'étude/expérience, calcul de puissance statistique, calcul du nombre d'échantillons nécessaire, etc.) et en aval (p. ex. nouveau modèle statistique, correction des biais expérimentaux et technologiques, etc.) de la génération des données.

Ce travail de recherche s'inscrit dans un esprit pluridisciplinaire et transversal, se situant à l'interface entre la biologie, la génétique et la statistique, en intégrant différents types de données “-omiques”. Les objectifs de cette thèse consistaient à apporter un support en statistique ainsi qu'une veille méthodologique, afin d'améliorer la compréhension des mécanismes biologiques au moyen d'outils d'analyse statistique adaptés et d'outils de visualisation des données “-omiques”. Tous les résultats d'analyses de ces données, conduites en collaboration avec des chercheurs de l'unité et des chercheurs à l'international, ont tenté de répondre au questionnement biologique inhérent à l'étiologie du diabète de type 2, en accord avec les besoins identifiés par les chercheurs au sein de l'unité

## Plan

Différentes méthodes statistiques et types de données seront abordés au travers de quatre chapitres, chacun correspondant à un article publié (Chapitres 2 et 4), soumis (Chapitre 1 et 3) dans des revues internationales à comité de lecture.

Le premier chapitre porte sur le développement et l'application d'un modèle joint permettant de modéliser conjointement deux processus stochastiques : d'une part, la modélisation de la trajectoire de la glycémie à jeun chez des individus issus d'une population générale, et d'autre part, l'évolution du risque de développement d'un diabète de type 2, conditionnellement à la trajectoire de la glycémie à jeun. Nous nous intéressons particulièrement à l'effet simultané des polymorphismes (SNPs) sur ces deux processus. Le principal objectif est d'évaluer ce modèle du point de vue de la puissance statistique, de l'erreur de type 1 et du temps de calcul, dans un contexte d'application à la génomique, c'est-à-dire avec un volume très élevé de données. Cette évaluation est faite, en premier lieu, sur des données simulées, puis en second lieu, sur un jeu de données réelles générées par l'unité.

- ARTICLE 1 : Variants génétiques associés à la trajectoire de la glycémie à jeun et à l'incidence du diabète de type 2 : Une approche par modèle joint (Soumis à *Genetic Epidemiology*)

Le second chapitre vise à étudier l'expression des gènes de susceptibilité au diabète de type 2, et notamment la contribution de ces gènes dans la sécrétion d'insuline au niveau des cellules  $\beta$  du pancréas.

Deux objectifs ont été remplis : dans un premier temps, identifier les gènes (parmi 104 candidats) étant exprimés dans l'organe clé de la sécrétion d'insuline, soit les cellules  $\beta$  (24 tissus et types cellulaires considérés, dont du tissu pancréatique), dans un second temps, évaluer l'impact de ces gènes sur la sécrétion d'insuline dans un modèle humain de cellules  $\beta$ . Enfin, le séquençage de l'ARN a été effectué pour les gènes présentant un effet sur la sécrétion d'insuline, suivi d'une étude dans un modèle murin dont la fonction pancréatique a été altérée.

- ARTICLE 2 : L'Expression et l'Évaluation Fonctionnelle des Gènes de Susceptibilité au Diabète de Type 2 Identifiant Quatre Nouveaux Gènes Contribuant à la Sécrétion d'Insuline Humaine (Publié dans *Molecular Metabolism*)

Le troisième chapitre s'intéresse à la fois au transcriptome et au méthylome dans une étude cas/témoins portant sur le diabète de type 2. Dans ce chapitre, le foie est l'organe étudié, notamment pour son implication dans la production de glucose, l'insulinorésistance hépatique, et dans les complications souvent associées au diabète de type 2, comme les NAFLD ("Non-Alcoholic Fatty Liver Disease"). L'étude du méthylome a permis de mettre

en évidence un site CpG (cg14496282) localisé sur le gène *PDGFA* ("Platelet-Derived Growth Factor subunit A"), présentant une hypométhylation chez les diabétiques de type 2. Cette hypométhylation est inversement corrélée avec l'expression de *PDGFA*, l'insulinémie et l'insulinorésistance (évalué par l'indice HOMA-IR). Les résultats d'une étude sur un modèle d'hépatocytes humains et de différents scores de risque génétique (*Genetic Risk Score*) suggèrent une relation causale de l'hyperinsulinémie sur le niveau de méthylation du site CpG identifié, ouvrant ainsi la voie vers une potentielle cible thérapeutique.

- ARTICLE 3 : La Surexpression Hépatique de PDGF-AA Affaiblit la Signalisation de l'Insuline dans le Diabète  
(Soumis à **Nature Communications**)

Le quatrième chapitre propose d'étudier les effets du bisphénol A (BPA) et de ses substituants, soit les bisphénol F (BPF) et bisphénol S (BPS), sur l'expression des gènes (ARN codant et non codant) dans le tissu adipeux, et notamment les adipocytes. Le lien entre le BPA et les désordres métaboliques comme le diabète de type 2 ayant déjà été démontré dans des études antérieures, l'objectif consiste à mesurer au niveau transcriptomique l'effet d'une faible concentration de bisphénol correspondant à celle retrouvée chez l'Homme, et une concentration plus forte qui pourrait résulter d'un relargage massif des adipocytes lors d'une perte de poids, par exemple, ou pendant la lipolyse des adipocytes survenant dans certaines maladies métaboliques comme le diabète de type 2.

- ARTICLE 4 : L'Exposition à Faible Dose aux Bisphénols A, F et S des Adipocytes Primaires Humains Modifie les Profils d'ARN Codant et Non-Codant (Publié dans **PLoS ONE**)

Enfin, une discussion générale clôt cette thèse et tente de replacer ces travaux dans un contexte multi-omique élargi. Elle apporte des perspectives de travail quant à l'évolution de la statistique génétique, tributaire en partie de l'évolution des technologies permettant de générer (encore) plus de données de différentes natures (biologique ou informatique), en regard notamment de la diminution exponentielle des coûts de séquençage de ces dernières années, et le développement des études d'association portant sur les variants rares dans les populations étudiées.



# Chapitre 1

---

*Variants génétiques associés à la trajectoire de la glycémie à jeun et à l'incidence du diabète de type 2 : Une approche par modèle joint*

---

Soumis à ***Genetic Epidemiology***.

**Mickaël Canouil<sup>1,2,3</sup>, Philippe Froguel<sup>1,2,3,4</sup> & Ghislain Rocheleau<sup>1,2,3</sup>**

<sup>1</sup>Université de Lille, UMR 8199 - EGID, F-59000 Lille, France; <sup>2</sup>CNRS, UMR 8199, F-59000 Lille, France; <sup>3</sup>Institut Pasteur de Lille, F-59000 Lille, France; <sup>4</sup>Department of Genomics of Common Disease, Imperial College London, London, United Kingdom.

---

## 1 Introduction

### 1.1 Contexte/objectifs

Dans le but d'optimiser l'utilisation des données phénotypiques existantes, nous proposons une approche statistique par modèle joint (JM) permettant l'identification de marqueurs génétiques simultanément associés à la trajectoire temporelle d'un trait phénotypique et à la survenue d'un événement. Nous illustrons l'application du modèle joint dans un contexte génétique des maladies métaboliques, en exploitant la forte association entre la trajectoire temporelle de la glycémie à jeun et l'incidence du diabète de type 2 (DT2).

### 1.2 Méthodes

Le modèle proposé dans notre étude consiste en un modèle de régression linéaire mixte combiné à un modèle de survie dit de Cox à risque proportionnel. À partir des données de génotypage (Illumina Metabochip DNA arrays) obtenues pour près de 4 500 individus de la cohorte D.E.S.I.R. (Données Épidémiologiques sur le Syndrome d'Insulino-Résistance), nous avons analysé l'ensemble des variants génétiques disponibles (SNPs). Sur la base de simulations faisant varier plusieurs paramètres comme le nombre de mesures, le nombre d'in-

dividus, la fréquence allélique et/ou le taux d'incidence, aboutissant ainsi à 240 scénarios différents (c.-à-d. 240 combinaisons des valeurs possibles pour chaque paramètre) chacun simulés 500 fois, l'erreur de type I, la puissance statistique et les estimations obtenues ont fait l'objet d'une étude comparative entre l'approche par modèle joint et les approches classiques utilisées dans les études d'association pangénomiques (GWAS).

### 1.3 Résultats

Nos résultats démontrent la forte association entre la glycémie à jeun et l'incidence du DT2 (ce qui était attendu selon la définition clinique du DT2), et confirment également l'association entre la glycémie et certains SNPs rapportés dans les études de type GWAS, tels que les SNPs situés dans les gènes *G6PC2* ou encore *TCF7L2*. Les associations relevées ici sont pour la plupart nominales (valeur-p < 0,05), principalement en raison de la faible taille de notre cohorte en comparaison aux tailles d'effectifs rapportée en méta-analyse, et aussi en raison du nombre peu élevé de cas de DT2 incident (environ 5 % sur 9 ans de suivi dans la cohorte D.E.S.I.R.). Notre analyse par modèle joint a révélé que les SNPs se situant près ou dans le gène *MTNR1B* pourraient ne pas avoir d'effet simultané sur l'élévation de la glycémie à jeun et le risque de survenue du DT2.

Notre étude comparative des différents modèles révèle que l'approche JM pourrait être plus puissante, en comparaison des approches transversales (c.-à-d. régression linéaire et logistique), pour détecter des effets de polymorphisme, aussi bien sur le trait longitudinal (paramètre  $\gamma$ ) que sur le risque de survenue d'un événement (paramètre  $\alpha$ ), tout en maintenant l'erreur de type I près du niveau global de 5 %. En outre, nous avons pu observer que l'approche en deux-étapes ("Two-Step" ou TS) présentait une puissance et une erreur de type I similaires à celles obtenues avec l'approche JM.

L'étude du RMSE (*Root Mean Square Error*) montre que l'estimation de  $\alpha$  est impactée par le nombre d'individus et la fréquence du polymorphisme, mais reste similaire entre l'approche JM et l'approche par modèle linéaire mixte. Les valeurs de RMSE divergent selon les méthodes, notamment pour l'estimation de  $\beta$  (effet de la trajectoire sur le risque d'événement) : le modèle de Cox avec covariable dépendante du temps fournit les valeurs de RMSE les plus élevées sur l'ensemble des scénarios. Les valeurs de RMSE de l'approche TS tendent à se rapprocher de celles de l'approche JM, particulièrement lorsque le nombre de mesures longitudinales augmente. Dans le cas du paramètre  $\alpha$ , les valeurs de RMSE se montrent sensibles au nombre d'individus (< 1 000), au faible nombre d'événements (taux d'incidence < 2,5 %), et à la fréquence du polymorphisme (< 5 %). Dans ces scénarios de faible fréquence allélique, faible taux d'incidence ou petit nombre d'individus, l'approche JM présente les valeurs de RMSE les plus faibles comparativement aux approches TS et Cox, ces différences tendant à s'estomper lorsque le nombre d'individus est supérieur à 2 500, ou que la fréquence

allélique est supérieure à 5 %. Enfin, la maximisation de la vraisemblance jointe de l'approche JM se révèle être consommatrice de temps, à hauteur d'un facteur de 30 à 40 fois le temps de calcul requis par l'approche TS.

#### 1.4 Conclusion

L'analyse par modèle joint a montré, d'une part, une grande cohérence avec les résultats des études antérieures de type GWAS, et d'autre part, semble indiquer un gain de puissance statistique pour détecter l'effet d'un SNP sur l'évolution de la glycémie à jeun et/ou sur la survenue du DT2. Cependant, l'approche JM présente un frein important dû au temps de calcul énorme qui ne permet pas l'exploration systématique de tous les SNPs à une échelle pangénomique. L'approche TS ayant montré des caractéristiques (estimation, puissance et erreur de type I) proches de celles de JM, et réalisable dans un temps raisonnable, pourrait être employée comme un filtre sur les polymorphismes. Dans un second temps, un affinage des estimations pourrait s'obtenir au moyen de l'approche JM. Enfin, le résultat obtenu pour le gène *MTNR1B* tend à montrer qu'une modélisation statistique simultanée des deux processus pourrait mener à une identification plus fine des variants génétiques associés à l'homéostasie du glucose sanguin ou à la physiopathologie du diabète.

---

## 2 Article

## Single Nucleotide Polymorphisms Associated with Fasting Plasma Glucose Trajectory and Type 2 Diabetes Incidence: A Joint Modelling Approach

Mickaël Canouil<sup>1,2,3</sup>, Beverley Balkau<sup>4,5,6</sup>, Ronan Rousse<sup>7,8,9</sup>, Philippe Froguel<sup>1,2,3,10</sup> and Ghislain Rocheleau<sup>1,2,3</sup>

<sup>1</sup>Univ. Lille, UMR 8199 - EGID, F-59000 Lille, France.

<sup>2</sup>CNRS, UMR 8199, F-59000 Lille, France.

<sup>3</sup>Institut Pasteur de Lille, F-59000 Lille, France.

<sup>4</sup>CESP Centre for Research in Epidemiology and Population Health, Villejuif, France.

<sup>5</sup>Univ. Paris-Saclay, Univ. Paris Sud, UVSQ, UMRS 1018, F-94807, Villejuif, France.

<sup>6</sup>INSERM U1018, CESP, Renal and Cardiovascular Epidemiology, UVSQ-UPS, Villejuif, France.

<sup>7</sup>INSERM, U1138 (équipe 2: Pathophysiology and Therapeutics of Vascular and Renal Diseases Related to Diabetes, Centre de Recherches des Cordeliers), Paris, France.

<sup>8</sup>Univ. Paris 7 Denis Diderot, Sorbonne Paris Cité, France.

<sup>9</sup>AP-HP, DHU FIRE, Department of Endocrinology, Diabetology, Nutrition, and Metabolic Diseases, Bichat Claude Bernard Hospital, Paris, France.

<sup>10</sup>Department of Genomics of Common Disease, Imperial College London, London, United Kingdom.

### Corresponding authors

Mickaël Canouil, EGID - UMR 8199, Pôle Recherche - 1er étage Aile Ouest, 1 place de Verdun, 59045 Lille CEDEX, France. Phone: +33(0)3-74-00-81-29. E-mail: mickael.canouil@cnrs.fr

Ghislain Rocheleau, Maelstrom Research - The Research Institute of the McGill University Health Centre (RI MUHC), 2155 Guy, 4th Floor, Office 458, Montreal, Quebec, H3H 2R9, Canada. E-mail: grocheleau@maelstrom-research.org

## Abstract

In observational cohorts, longitudinal data are collected with repeated measurements at predetermined time points for many biomarkers, along with other covariates measured at baseline. In these cohorts, time to a certain event of interest occurring is commonly reported and very often, a relationship will be observed between a biomarker repeatedly measured over time and that event. Joint models were designed to efficiently estimate statistical parameters by combining a mixed model for the longitudinal biomarker trajectory and a survival model for the event risk, using a set of random effects to account for the link between the two types of data.

First, we checked model consistency based on different simulation scenarios, varying sample size, minor allele frequency and number of repeated measurements. Second, using genotypes assayed with the Metabochip DNA arrays (Illumina) from close to 4,500 individuals recruited in the French cohort D.E.S.I.R. (*Data from an Epidemiological Study on the Insulin Resistance syndrome*), we assessed the feasibility of implementing the joint modelling approach in a real high-throughput genomic dataset and showed more precise and less biased estimations through a joint modelling approach (e.g. Joint Model, Two-Step approach). Although, the Joint Model showed better estimations, the Two-Step model showed more suitable computation times and thus could be used for screening purposes at genome-wide scale. To the best of our knowledge, joint models have never been applied in a genetic epidemiology context and could help identify novel loci sharing effects on both glycaemic traits and T2D.

## Key words

Diabetes; fasting plasma glucose; genetic association; joint modelling; longitudinal studies

## Introduction

With the increased availability of longitudinal and survival data in cohorts, joint models have emerged to account for both types of data, particularly when dealing with the informative/non-informative dropouts which occur in such cohorts. Joint models have been studied and overviewed in the literature (Chen, Ibrahim, & Chu, 2011; Elashoff, Li, & Li, 2016; Tsiatis & Davidian, 2004; Wulfsohn & Tsiatis, 1997) and implementation has been proposed in different software and platforms (Diggle & Kenward, 1994; Elashoff, Li, & Li, 2008; Proust-Lima, Joly, Dartigues, & Jacqmin-Gadda, 2009; Rizopoulos, 2010; Rizopoulos & Ghosh, 2011; Sun, Sun, & Liu, 2007). The main idea behind the joint modelling is: 1) to model efficiently the survival process with a time-varying covariate, accounting for missing data and measurements errors, and 2) to account for informative dropouts in the longitudinal data. To model the two components of a joint model, a linear mixed effects (LME) model and a Cox proportional hazards model (CoxPH), are classically used to, respectively, fit the longitudinal component, and the survival component. Unlike the CoxPH model, in which the time-varying covariate is assumed to be exogenous, i.e. not modified by the occurrence of an event (Kalbfleisch & Prentice, 2002), the joint modelling framework allows to account for an endogenous time-varying covariate. An example of an endogenous covariate would be the fasting plasma glucose which is irremediably modified due to glucose lowering medication, once T2D is diagnosed.

Two approaches can be used for the estimation and inference of the model parameters: a "naive" two-step (TS) method or a joint likelihood method (JM). In the first method, the random effects of the trajectory are estimated by an LME model, and they are included as a time-varying covariate in a CoxPH model, then parameter estimation uses the partial likelihood of the CoxPH model (Therneau & Grambsch, 2000). The second method is based on a joint likelihood of the two components (longitudinal and survival) at the same time. Comparison of these two approaches showed that the latter offers more consistent and efficient estimators than the former (Albert & Shih, 2010a, 2010b). But JM can be challenging to compute, especially achieving convergence at the Expectation-Maximisation (EM) step. Moreover, depending on the number of time points and/or the sample size, the overall computation time can substantially increase.

In this paper, we conduct a comprehensive simulation study to compare two modelling approaches, JM and TS, for jointly modelling the longitudinal and the survival components. Our main goal is to show whether the JM approach, when compared to TS, might improve statistical power to detect an effect on either, or both, the longitudinal and the survival processes, while resulting in a bias reduction in parameter estimation. We also compared JM with TS and show that in the context where highly demanding computation and convergence issues might arise in JM computation, whether the TS offers a good alternative to JM in a reasonable computation time-span, especially when applied at the genome-scale level. We also investigated and decomposed the computational time required by the R package "JM" (Rizopoulos, 2010, 2016), and by

the TS approach combining the R packages: "survival" (Therneau, 2017) and "nlme" (Pinheiro, Bates, & R-core, 2017).

Finally, we applied these approaches to a real dataset, the D.E.S.I.R. cohort (*Data from and Epidemiological Study on the Insulin Resistance syndrome*), that included 5,212 individuals with extensive phenotypic measures recorded at four three-yearly intervals, spanning a nine-year follow up. Individuals were genotyped using the Illumina Metabochip DNA array of nearly 200,000 SNPs) (Voight et al., 2012). Relying on cross-sectional genome-wide association study design, the D.E.S.I.R. cohort was instrumental in identifying novel loci associated with prevalent type 2 diabetes (T2D) and with fasting plasma glucose (FPG) level in normoglycemic individuals (Bouatia-Naji et al., 2008; Rung et al., 2009; Sladek et al., 2007). We specifically focus on prediabetes conditions, such as IFG (Impaired Fasting Glucose), and on time-to-onset of T2D, in order to possibly identify loci, novel or published, which simultaneously associate with the risk of developing T2D and with increasing FPG. Our results were then compared to the genetic variants as reported in the literature (Vaxillaire et al., 2014; Welter et al., 2014), and to the meta-analyses results published by large consortia, such as, DIAGRAM (Morris et al., 2012) and the MAGIC (Dupuis et al., 2010) consortia.

## Methods

### **Model Formulations**

#### **Joint Likelihood Model (JM)**

The standard formulation of the joint model involves two components: a longitudinal component and a time-to-event component. Let  $n$  denote the sample size, and  $Y_{ij}$  the longitudinal measurements collected for each individual  $i$  at time points  $t_{ij}$ ,  $i = 1, \dots, n$ ,  $j = 1, \dots, m_i$ , where  $m_i$  is the number of measurements on individual  $i$ . The longitudinal component (measurements) typically consists of a (generalised) linear mixed effect (LME) model, whose within-subject correlation matrix is modelled using random-effect parameter vector  $b_i = \begin{pmatrix} \theta_{0i} \\ \theta_{1i} \end{pmatrix}$ .

Under the joint likelihood framework as implemented in "JM" (Rizopoulos, 2010, 2016), within the class of "shared parameter models" (Elashoff et al., 2016; Rizopoulos, 2012), we define

$$Y_{ij} = X_{ij} + \epsilon_{ij} \quad (1)$$

where  $Y_{ij}$  is the observed value and  $X_{ij}$  is the true (unobserved) value of the longitudinal measurement at time  $t_{ij}$  for individual  $i$ . The quantity  $\epsilon_{ij}$  is a random error term usually assumed to be normally distributed:

$$\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \quad (2)$$

The quantity  $X_{ij}$  is typically called the trajectory function and is usually specified as a linear or quadratic function of time  $t_{ij}$ ; for simplicity here, we assume linearity over time. We also define  $Z_i$ , a vector denoting the genotype of individual  $i$ , and  $W_i$ , a set of adjusting covariates:

$$Y_{ij} = X_{ij} + \epsilon_{ij} = \theta_{0i} + \theta_{1i}t_{ij} + \gamma Z_i + \delta W_i + \epsilon_{ij} \quad (3)$$

For simplification here, the term  $\delta W_i$  will be omitted. Random effects  $\theta_{0i}$  (intercept) and  $\theta_{1i}$  (slope) are assumed bivariate Normal:  $\theta \sim \mathcal{N}_2(\mu, \Sigma)$ , and independently distributed from  $\epsilon_{ij}$ . The coefficient  $\gamma$  assesses the genotypic (additive) effect of variable  $Z_i$  in the trajectory function. To account for possible varying slopes, an interaction term between  $Z_i$  and time  $t_{ij}$  could be added into the trajectory function; this term was not considered in this study.

The time-to-event (survival) component usually consists of a parametric (e.g. exponential or Weibull distribution) or semi-parametric (e.g. Cox proportional hazards) model.  $T_i$  denotes the event time for individual  $i$ , and  $C_i$  the right censoring time (end of the follow-up). Let  $\Delta_i$  be the event indicator:  $\Delta_i = 0$ , if  $T_i > C_i$ , and  $\Delta_i = 1$ , if  $T_i \leq C_i$ . Under the Cox proportional hazards model, variable  $T_i$  is specified using the following equation:

$$\lambda_i(t) = \lambda_0(t)\exp(\beta X_i(t) + \alpha Z_i) \quad (4)$$

where  $\lambda_i(t)$  is the hazard function at time  $t$  for individual  $i$  and  $\lambda_0(t)$  is the unspecified baseline hazard function, which we assume piecewise constant with two knots placed at intermediate time points in the follow-up. The coefficient  $\alpha$  measures the effect of  $Z_i$  on the hazard function, while  $\beta$  measures the association between the trajectory function and the hazard function. In this formulation, we suppose that the subject-specific parameters  $b_i = \begin{pmatrix} \theta_{0i} \\ \theta_{1i} \end{pmatrix}$  included in the trajectory  $X_i(t)$  could modify the hazard function, which implies that  $\beta$  is the parameter linking the longitudinal and survival components.

## Two-Step Model (TS)

As an alternative to JM, and based on the work of Tsiatis, DeGruttola, & Wulfsohn (1995), the two-step model estimates parameters of the joint model by first, estimating parameters of the trajectory function  $X_i(t)$  in Equation (3), and second, by substituting this estimated trajectory, say  $X_i^*(t)$ , into (4) before fitting the Cox survival model.

## Simulation Study

Simulation studies were carried out to further examine the sensitivity of the JM estimations under several scenarios. Parameters were set based on values estimated from the strongest SNPs associated with T2D (Table I), that is rs17747324 in gene *TCF7L2* (T2D effect allele: C;  $OR = 1.43$ ;  $p = 8.5 \times 10^{-55}$  (Morris et al., 2012); FPG effect allele: C;  $\beta = 0.025$ ;  $p = 6.47 \times 10^{-08}$  (Dupuis et al., 2010)).

Longitudinal data were simulated according to Equation (3), while event times were generated from an exponential distribution for the CoxPH model (Austin, 2012)

$$\lambda_0(t) = \lambda \quad (5)$$

$$H_i(T_i) = \int_0^{T_i} \lambda \exp(\beta X_i(t) + \alpha Z_i) dt \quad (6)$$

$$T_i = \frac{1}{\beta \theta_{1i}} \log \left( 1 - \frac{\beta \theta_{1i} \times \log(1 - u)}{\lambda \exp(\beta \theta_{0i} + (\beta \gamma + \alpha) Z_i)} \right) \quad (7)$$

where  $\lambda$  was set to achieve the targeted incidence rate in the simulated dataset.

Datasets were simulated by varying the number of longitudinal measurements  $m \in \{2; 3; 4; 5\}$ , the number of individuals studied  $n \in \{500; 1,000; 2,500; 5,000; 10,000\}$ , the allele frequency  $f \in \{0.05; 0.1; 0.25; 0.5\}$  and the incidence rate  $d \in \{0.025; 0.05; 0.1\}$ , thereby leading to 240 different scenarios. Each scenario was simulated 500 times.

The Root-Mean Square Error (RMSE)

$$\text{RMSE}(\hat{\theta}) = \sqrt{E((\hat{\theta} - \theta)^2)} \quad (8)$$

was used to assess precision for the estimation of  $\beta$ ,  $\gamma$  and  $\alpha$ , when testing the association between  $Y_{ij}$ , and  $T_i$ , the  $Z_i$  effect on  $Y_{ij}$  and the  $Z_i$  effect on  $T_i$ , respectively. We compared JM and TS approaches with a linear mixed effect model and the Cox regression model with a time-varying covariate, fasting plasma glucose. In addition, statistical power and type I error were also studied. The computational burden of each approach was also investigated as our goal is to implement these models at a genome-wide scale.

### **Computational times**

Based on our simulations, we provide approximate computational times for four sample sizes with parameters as listed in Table I, using a UNIX system with Intel® Xeon® CPU E7- 4870 @ 2.40GHz (80 such CPUs available computing in parallel). Table II shows computational time for one model, and when extrapolating the total computational time for 100,000 SNPs, which is the approximate number of SNPs on the Metabochip, after data cleaning and the quality-control over common SNPs (minor allele frequency > 0.05).

To investigate further computational time issues, we profiled the execution of the main function "*jointmodel*" from the R package "JM", which implements the joint likelihood modelling approach as described in this paper. In the "JM" package, the linear mixed effect sub-model is handled by the function "*lme*" from the "nlme" package. One may argue that using a faster approach, e.g. as implemented in the R package "lme4", the computational time might be decreased.

### Real Data

SNP genotyping was performed with Metabochip DNA arrays (Voight et al., 2012) using Illumina HiScan technology and GenomeStudio software (Illumina, San Diego, USA) in 5,212 individuals from the French cohort D.E.S.I.R. (Balkau, 1996). They have been followed for nine years, and extensive phenotypic data has been recorded at four different three-yearly times during that follow-up. Quality control was performed using PLINK 1.90 beta version (Chang et al., 2015; Purcell & Chang, 2015). SNPs with call rate of at least 95%, with no significant deviation from Hardy-Weinberg equilibrium at  $p > 1 \times 10^{-5}$ , and with minor allele frequency (MAF) over 5% were kept for analysis, resulting in 101,305 SNPs. Due to missing phenotypes which did not allow to confirm T2D status, 232 individuals were removed. An additional 554 individuals were excluded due to individual call rate lower than 95%, leaving 4,426 individuals for analysis after these quality control steps (Supplementary Figure S1).

Principal component analysis was performed in a combined dataset with the 4,426 D.E.S.I.R. participants, and individuals from the publicly available 1,000 Genomes database (The 1000 Genomes Project Consortium, 2015). SNPs retained for analysis were restricted to those common to both samples. The first two components were sufficient to discriminate ethnic origin. Non-Caucasians (62) were excluded from the analysis. A further 12 prevalent cases of T2D at baseline were also removed.

The final dataset included 4,352 individuals, of whom 167 were diagnosed as T2D incident cases. Type 2 diabetes was defined using one of the following criteria: use of glucose lowering medication, fasting plasma glucose [ $FPG$ ]  $\geq 7$  mmol/L, or glycated hemoglobin A1c [ $HbA1c$ ]  $\geq 6.5\%$  (48 mmol/mol).

Using the joint modelling approach implemented in the package JM (Rizopoulos, 2010, 2016) within the R software version 3.3.3 (R Core Team, 2015), all 101,305 SNPs were tested for joint association with FPG and T2D. Based on the joint modelling formulation, let  $Y_{ij}$  denote the observed values of FPG, and let  $Z_i$  represent the genotype of individual  $i$  at each SNP, along with  $W_i$  covariates such as age, sex and BMI (Figure 1). Finally, let  $T_i$  is the time at which an individual is diagnosed with T2D.

In the joint modelling framework, the trajectory of FPG is viewed as a dropout process, since all FPG values become missing after T2D diagnosis, as a result of individuals with diabetes being placed under treatment to lower and regulate the glucose level in their blood. In this case, FPG is considered as an endogenous covariate, because the dropout process is not independent from the measured glucose values prior to T2D diagnosis.

## Results

### **Comparison of estimation accuracy**

Due to the complexity of the estimating algorithm within JM, convergence could not be obtained (4.53 ± 5.81% of convergence issues on average per scenario) for the whole set of 500 simulations (i.e. algorithm "piecewise-PH-aGH" for a time-dependent relative risk model with a piecewise constant baseline risk function, using the adaptive Gauss-Hermite quadrature rule to approximate integrals within the Expectation-Maximisation (EM) step (Rizopoulos, 2010, 2016)).

RMSE for parameter  $\gamma$  (Figure 2) showed similar performance for JM and TS, which was expected given the formulation of the joint model within the "Shared Parameter Models" framework, in which  $Y_i$  (mean of  $Y_{ij}$ ) modelled within LME according to Equation (3)) links the longitudinal data to the time of event.

RMSE for parameter  $\beta$  (Figure 3) and for parameter  $\alpha$  (Figure 4) were smaller within the joint modelling framework (either JM or TS) than in the more classical CoxPH model with time varying fasting plasma glucose. While RMSE for  $\beta$  was the same in the CoxPH model across all scenarios, under JM or TS it decreased with increases in the sample size, incidence rate or allele frequency. Differences in RMSE for parameter  $\alpha$  were less than for parameter  $\beta$ , and TS and CoxPH with time-dependent covariate, performed equally probably because partial likelihood inferences were used in both approaches. JM estimations were less biased in almost all scenarios when the sample size was greater than 2,500.

Overall, our simulations showed that JM is less biased than when separate approaches are used to model the effect of  $Z_i$  on the longitudinal  $Y_i$ , and on the time-to-event  $T_i$ . While separate approaches performed well for parameters  $\gamma$  and  $\alpha$ , the bias for  $\beta$  was the greatest observed across all scenarios.

For the default simulation settings (Table I), the type 1 error and statistical power showed similar results between JM and TS (Table III). Nevertheless, the simulations highlighted convergence issues that might occur within the joint likelihood approach (19.4% of the power simulation study).

### **Computational time**

Computational times are reported in Table II. The time required to complete JM or TS algorithms increased linearly with sample size in our simulations. However, these times are very optimistic since our simulations did not include any covariate or more complex random parameters.

The main issue is within the "*jointmodel*" function which took over 95 % of the global computation time (Supplementary Figure S2). After examination of the call tree diagram, we can see that the more time-consuming task within the "*jointmodel*" function is the optimisation of the EM algorithm (described in

Rizopoulos (2012), Appendix B), despite the use of a calculation trick (i.e. adaptive Gauss-Hermite quadrature for numerical integration).

### **Application in real data**

Applying the R package JM to our D.E.S.I.R. cleaned dataset, 265 SNPs (Figure 5) were associated (with p-value<0.05) with FPG and T2D events through their respective parameters  $\gamma$  and  $\alpha$ . Amongst these 265 SNPs (163 unique genes), we identified 17 genes (Supplementary Table I) which had already been reported to be associated with FPG and/or T2D risk. Parameter  $\beta$  was highly significant (below the genome-wide threshold of  $5 \times 10^{-8}$ ) for all these SNPs, which was expected considering the fact that  $\beta$  estimates the association between FPG trajectory and T2D risk, therefore one of the criteria used to define T2D.

In Figure 6, we specifically focused on parameters  $\gamma$  and  $\alpha$ . After Bonferroni correction (nominal p-value  $\simeq 5 \times 10^{-7}$ ), no genetic variants showed a highly significant association with both parameters  $\gamma$  and  $\alpha$  simultaneously; only SNPs in the following genes (or within a 100 kb window) remained significant when testing for  $\gamma$ : *G6PC2/ABCB11*, *GCK/YKT6*, *GCKR* and *MTNR1B*, with effect per risk allele of increasing FPG from 0.10 mmol/L to 0.047 mmol/L. Zooming in on simultaneous associations with the longitudinal and survival processes revealed well known genes, such as *TCF7L2*, which has been shown in many meta-analyses to be associated with elevated FPG and an increased risk of T2D (Table IV). *MTNR1B* was also found to be associated (34 SNPs within 30kb) with  $\alpha = -0.44$  (p – value =  $9.37 \times 10^{-4}$ ) and  $\gamma = 0.099$  (p – value =  $1.33 \times 10^{-23}$ ) for SNP rs10830963, the SNP usually reported.

To better compare JM and TS, we repeated the analysis on the whole dataset using TS. As shown in Figure 7, p-values can differ, especially for parameter  $\alpha$ ; for parameter  $\gamma$ , approximations were quite close to the p-values provided via the joint likelihood framework.

## **Discussion**

With the ever-increasing availability of genomic data generated by genotyping arrays and next generation sequencing, the need to develop and implement efficient models is important to ensure that statistical analysis will be achieved in a reasonable time frame. In this paper, we propose a comparison of two approaches, namely the joint model (JM) and the two-step model (TS), to estimate parameters accounting simultaneously the SNP effect on longitudinal and on survival processes without omitting information about missing values and dropouts for the status of the longitudinal variable of interest. In our real data application, FPG is the longitudinal trait, whereas T2D diagnosis is the survival time of interest, both being linked together by the fact that an upper threshold on FPG actually defines T2D onset (currently,  $FPG \geq 7$  mmol/L), along with glucose lowering medication. Through simulations over different scenarios, we showed that joint models are less biased than classical separate approaches, could provide more insight regarding

the event of interest, and could better assess the potential impact of a SNP on incident T2D than current methods.

By looking at statistical measures, such as RMSE for accuracy in the model estimators, and by estimating computational time using the available R implementation of joint models, our study showed that the use of an approximate method, such as TS, at a genome-wide scale, might be a good trade-off between accuracy and computational time. TS could be used to overcome the computational burden of current joint likelihood methods by exploiting available software performing the two steps, LME and CoxPH, and could help filter out SNPs with low or undetectable associations during a first preliminary scan. However, depending on the parameters of the data set (sample size, incidence rate, number of measures), a joint likelihood method is highly preferred to obtain accurate estimation of parameters  $\gamma$  and  $\alpha$ , describing the SNP effect on the trajectory of FPG and time-to-onset of T2D. Although, we computed the theoretical statistical power to detect a genetic joint effect  $\beta\gamma + \alpha$  based (Chen et al., 2011), we did not test this effect at genome-wide scale due to its computational burden. Joint effect of the SNP can be tested using a likelihood ratio test to compare the full joint model (i.e. with SNP in both submodels) to the joint model without SNP in the survival submodel, as implemented in the package JM (Rizopoulos, 2010, 2016). Finally, using parallel and grid computing approaches will reduce the computational time to a more suitable time frame when applied at a genome-wide level (i.e. with millions of SNPs).

In our real data application, results for rs17747324 showed consistent results, with the DIAGRAM meta-analysis for both  $\alpha$  and  $\gamma$  (Table IV), and for rs10830963 showed an opposite effect on T2D compared to the effect reported in MAGIC for FPG ( $\alpha = 0.104$ ,  $p - \text{value} = 7.3 \times 10^{-7}$ ). The results observed for *MTNR1B* (rs10830963) in the French cohort D.E.S.I.R., even if they seemed inconsistent with previous studies, may uncover some interesting peculiarities pertaining to T2D incident cases in this population. In the literature, SNPs in *MTNR1B* were reported as being associated with higher FPG and T2D risk, but meta-analyses were performed on populations with different genetic backgrounds, and the two traits were never co-analysed jointly. However, we recognise that *MTNR1B* associations identified in our study need to be confirmed and replicated in other cohorts, as they might be cohort-specific associations. In addition, a major limitation of our study is the low number of incident T2D cases in the D.E.S.I.R. cohort (167 incident T2D cases in 4,352 individuals followed over 9 years).

## Acknowledgments

This study was supported by grants for funding of scientific research conducted in France and within the European Union: "Centre National de la Recherche Scientifique", "Université de Lille 2", "Institut Pasteur de Lille", "Société Francophone du Diabète", "Lilly", "Contrat de Plan Etat-Région", "Agence Nationale de la Recherche", ANR-10-LABX-46, ANR EQUIPEX Ligan MP: ANR-10-EQPX-07-01, European Research Council GEPIDIAB - 294785.

The D.E.S.I.R. study has been funded by INSERM contracts with Caisse nationale de l'assurance maladie des travailleurs salariés (CNAMTS), Lilly, Novartis Pharma, and sanofi-aventis; INSERM (Réseaux en Santé Publique, Interactions entre les déterminants de la santé, Cohortes Santé TGIR 2008); the Association Diabète Risque Vasculaire; the Fédération Française de Cardiologie; La Fondation de France; Association de Langue Française pour l'Etude du Diabète et des Maladies Métaboliques (ALFEDIAM)/Société Francophone de Diabétologie (SFD); l'Office national interprofessionnel des vins (ONIVINS); Ardix Medical; Bayer Diagnostics; Becton Dickinson; Cardionics; Merck Santé; Novo Nordisk; Pierre Fabre; Roche; Topcon.

The D.E.S.I.R. Study Group. INSERM U1018: B. Balkau, P. Ducimetière, E. Eschwège; INSERM U367: F. Alhenc-Gelas; CHU D'Angers: Y Gallois, A. Girault; Centre de Recherche des Cordeliers, INSERM U1138, Bichat Hospital: F. Fumeron, M. Marre, R Roussel; CHU de Rennes: F. Bonnet; CNRS UMR8090, Lille: A. Bonnefond, S. Cauchi, P. Froguel; Centres d'Examens de Santé: Alençon, Angers, Blois, Caen, Chateauroux, Chartres, Cholet, Le Mans, Orléans, Tours; Institute de Recherche Médecine Générale: J. Cogneau; General practitioners of the region; Institute inter-Regional pour la Santé: C. Born, E. Caces, M. Cailleau, O Lantieri, J.G. Moreau, F. Rakotozafy, J. Tichet, S. Vol.

## Conflict of interest disclosure

The authors declare that they have no conflict of interest.

## Tables

**Table I. Parameters and numerical values used for sensitivity analysis and simulations, based on results from rs17747324 within gene TCF7L2.**

Parameters	Values
Number of participants ( $n$ )	4,352
Number of measures ( $m$ )	4
Diabetes incidence rate ( $d$ )	0.0384
Minor allele frequency ( $f$ )	0.244
Random effects ( $\theta$ )	$\sim \mathcal{N}_2 \left( \begin{bmatrix} 4.55 \\ 0.0108 \end{bmatrix}, \begin{bmatrix} 0.143 & -0.00109 \\ -0.00109 & 6.8 \times 10^{-4} \end{bmatrix} \right)$
SNP effect on $Y_{ij}$ ( $\gamma$ )	0.0229
SNP effect on $T_i$ ( $\alpha$ )	0.265
Association between $Y_{ij}$ and $T_i$ ( $\beta$ )	3.17
Error term ( $\epsilon$ )	$\sim \mathcal{N}(0, 0.305^2)$

**Table II. Approximate computational times using function *system.time* of R software. System time was computed ten times per sample size (number of individuals). Extrapolation is displayed for 100,000 SNPs**

Sample Size	Joint Model		Two-Step Model	
	mean (sd) per SNP in seconds	100K SNPs in days	mean (sd) per SNP in seconds	100K SNPs in days
500	51 (3.4)	59	0.71 (0.066)	0.82
2,500	100 (11)	120	3.1 (0.092)	3.6
5,000	180 (25)	210	6.3 (0.17)	7.3
10,000	340 (34)	400	9 (0.22)	10

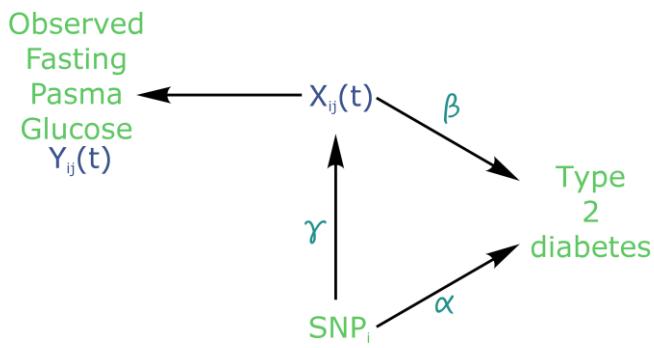
**Table III. RMSE, Type 1 Error and Power from default simulation settings for rs17747324 (TCF7L2).**

Parameter	Joint Model			Two-Step Model		
	RMSE	Type 1 Error	Power	RMSE	Type 1 Error	Power
$\alpha$	0.137	0.036	45.4%	0.139	0.051	48.5%
$\gamma$	0.010	0.051	61.8%	0.010	0.050	58.8%

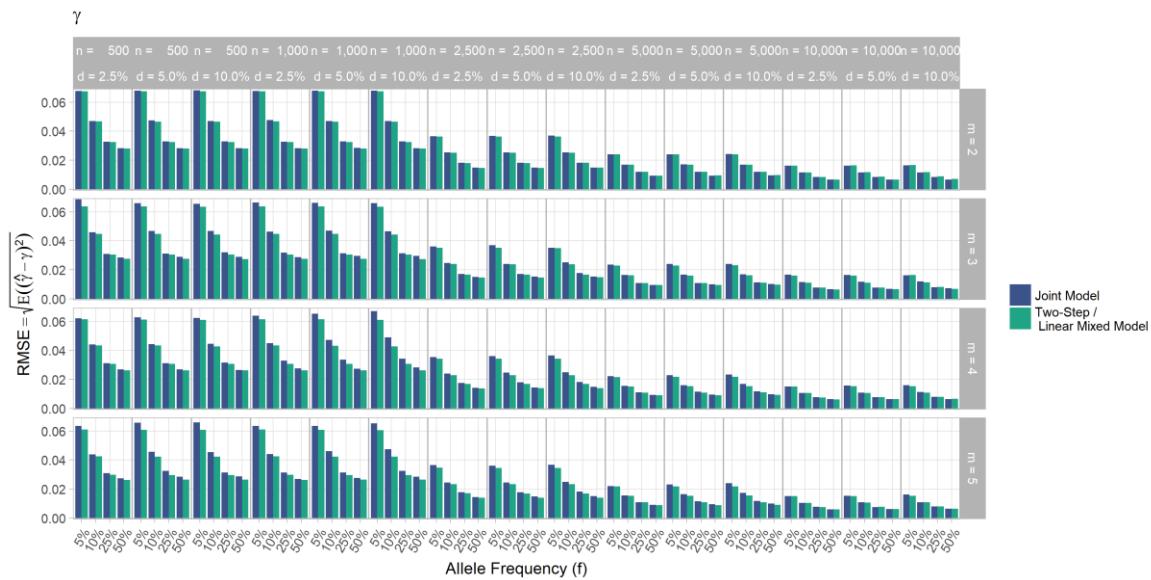
**Table IV. Effect sizes on FPG and T2D risk estimated using JM (Joint Model) and TS (Two-Step). Comparison is shown with effect sizes as reported by consortia meta-analyses in genes MTNR1B and TCF7L2.**

SNP (gene)	$\alpha$ (p-value)			$\gamma$ (p-value)			$\beta$ (p-value)		
	JM (D.E.S.I.R.)	TS (D.E.S.I.R.)	DIAGRAM	JM (D.E.S.I.R.)	TS (D.E.S.I.R.)	MAGIC	JM (D.E.S.I.R.)	TS (D.E.S.I.R.)	TS (D.E.S.I.R.)
rs10830963_C (MTNR1B)	-0.440 (9.4 × 10 <sup>-04</sup> )	-0.443 (5.0 × 10 <sup>-04</sup> )	0.104 (7.3 × 10 <sup>-07</sup> )	0.0991 (1.3 × 10 <sup>-23</sup> )	0.0992 (8.9 × 10 <sup>-24</sup> )	0.079 (1.3 × 10 <sup>-68</sup> )	0.079 (3.6 × 10 <sup>-42</sup> )	3.25 (3.6 × 10 <sup>-42</sup> )	3.52 (2.7 × 10 <sup>-54</sup> )
rs17747324_C (TCF7L2)	0.265 (4.1 × 10 <sup>-02</sup> )	0.284 (2.2 × 10 <sup>-02</sup> )	0.358 (8.5 × 10 <sup>-55</sup> )	0.0229 (3.0 × 10 <sup>-02</sup> )	0.0218 (3.8 × 10 <sup>-02</sup> )	0.025 (6.5 × 10 <sup>-08</sup> )	0.025 (8.9 × 10 <sup>-42</sup> )	3.17 (8.9 × 10 <sup>-42</sup> )	3.39 (2.2 × 10 <sup>-52</sup> )

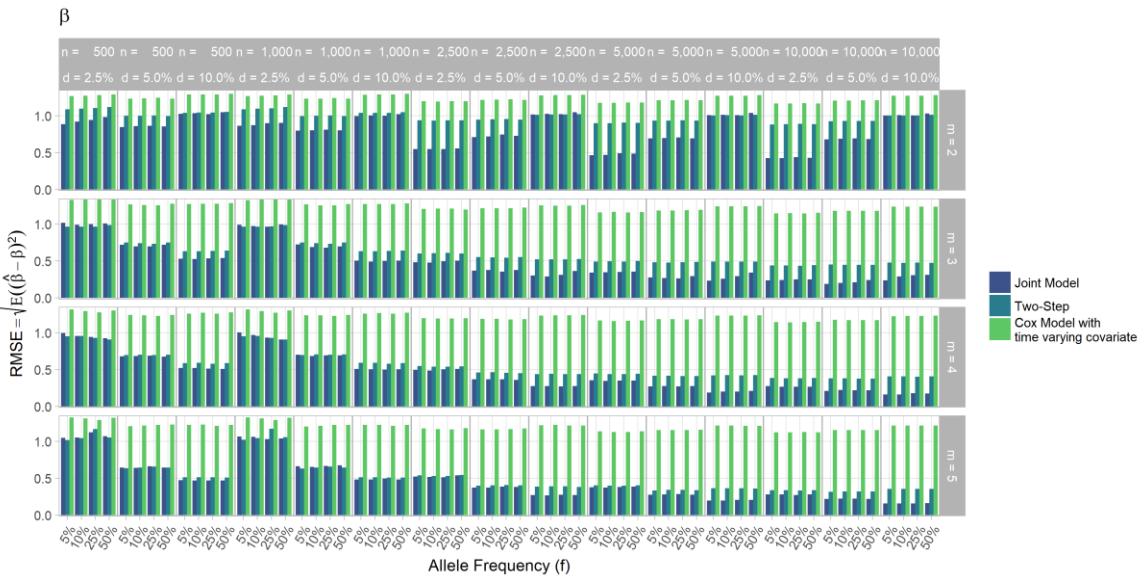
## Figures



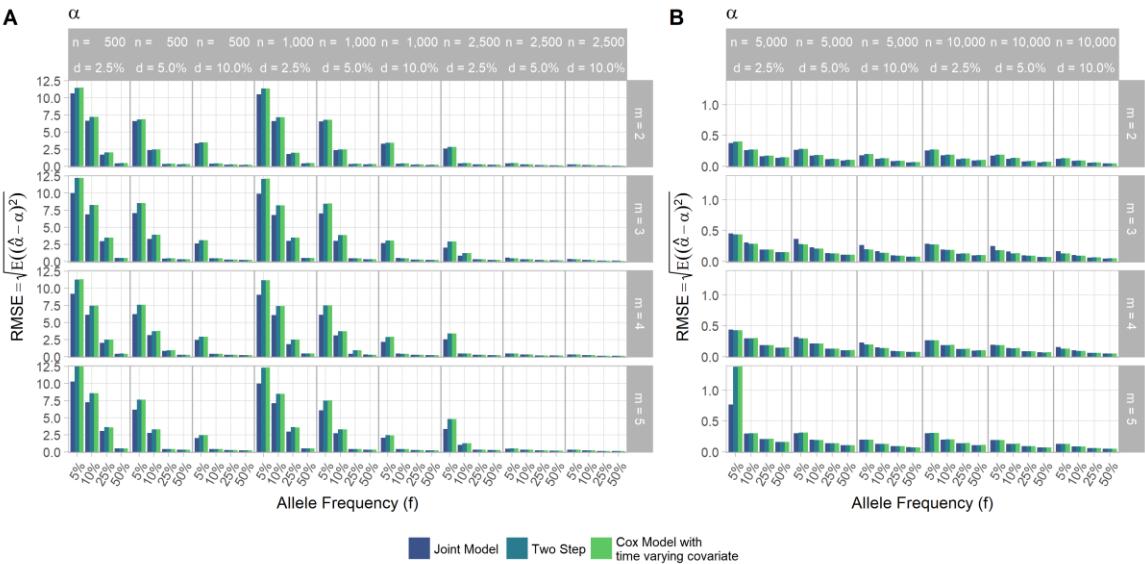
**Figure 1.** Causal diagram for joint modelling applied to fasting plasma glucose (FPG) and type 2 diabetes (T2D) (adapted from Ibrahim, Chu, & Chen (2010)). SNP: Single Nucleotide Polymorphism.



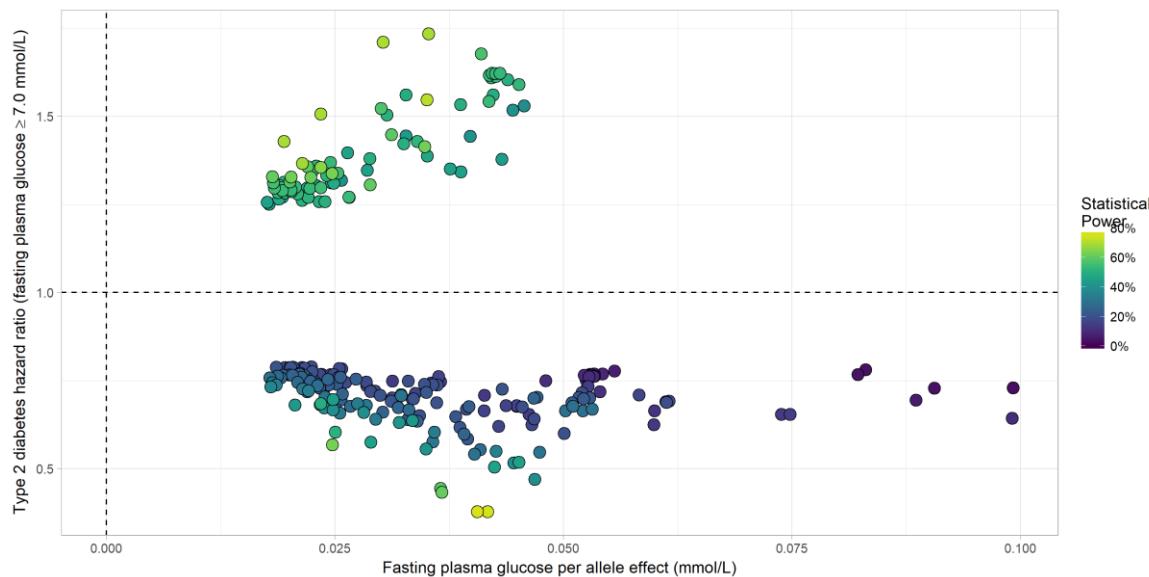
**Figure 2.** Simulation study for accuracy of estimator  $\hat{\gamma}$  provided by the joint model (JM package) and by the two-step linear mixed effect model (nlme package).  $m$ : number of measures;  $n$ : number of individuals;  $d$ : diabetes incidence rate.



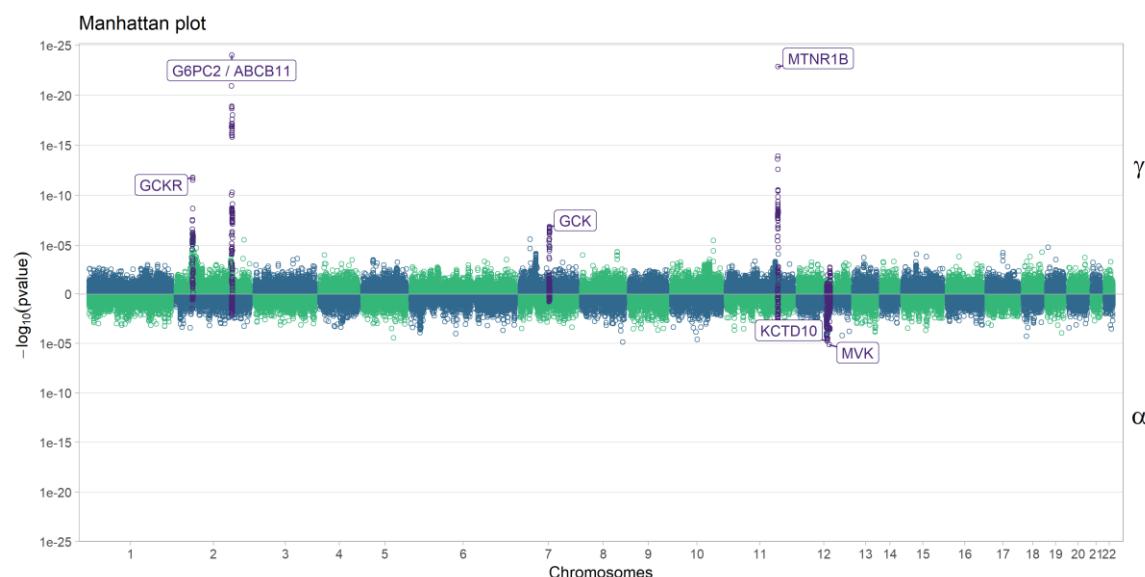
**Figure 3. Simulation study for accuracy of estimator  $\hat{\beta}$  provided by the joint model (JM package), by the two-step linear mixed effect model (nlme package) and by the Cox model with time-varying covariate. m: number of measures; n: number of individuals; d: diabetes incidence rate.**



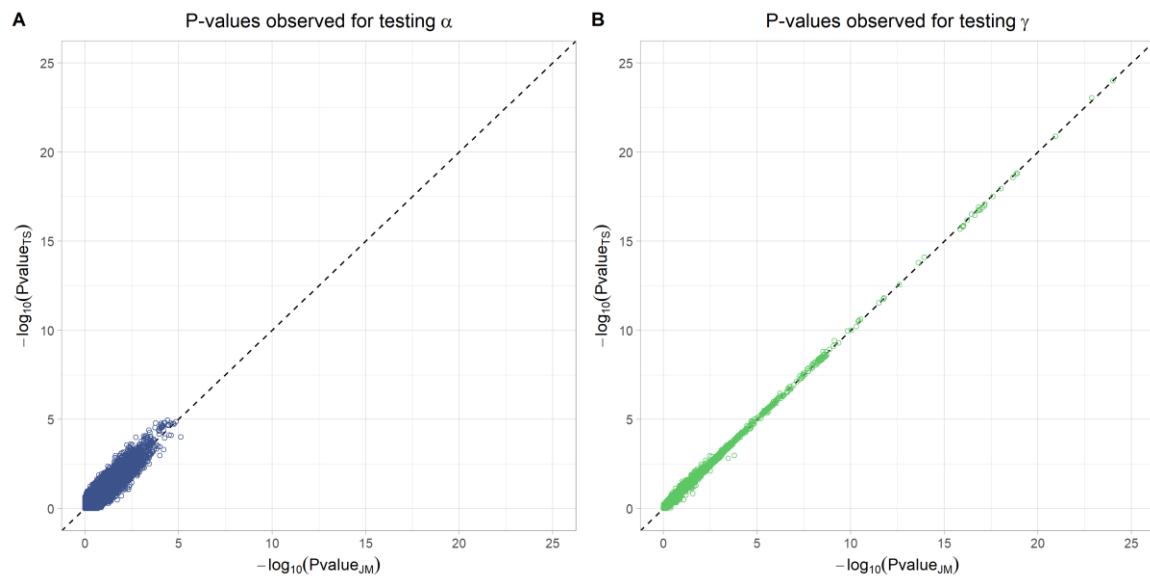
**Figure 4. Simulation study for accuracy of estimator  $\hat{\alpha}$  provided by the joint model (JM package), by the two-step linear mixed effect model (nlme package) and by the Cox model with time-varying covariate. m: number of measures; n: number of individuals; d: diabetes incidence rate.**



**Figure 5.** Results from statistical analysis using JM (Rizopoulos, 2010, 2016). Estimated effects of  $\gamma$  are displayed on the x-axis, with corresponding estimated odds ratio  $\exp(\alpha)$  on the y-axis. Statistical power reported is the theoretical (retrospective) power to detect a genetic joint effect  $\beta\gamma + \alpha$  based on estimated model parameters (Chen et al., 2011).



**Figure 6.** Manhattan plot for estimated effects of  $\gamma$  and  $\alpha$  using JM. Results are presented for the cleaned set of 101,305 SNPs.



**Figure 7. Testing for  $\alpha$  (SNP effect on onset of T2D) and  $\gamma$  (SNP effect on the trajectory of FPC) using Two-Step approach compared to Joint Model approach. On the x-axis,  $-\log_{10}(p\text{-value})$  from the Joint Model and on the y-axis the corresponding  $-\log_{10}(p\text{-value})$  from the approximate Two-Step approach.**

## References

- Albert, P. S., & Shih, J. H. (2010a). An approach for jointly modeling multivariate longitudinal measurements and discrete time-to-event data. *The Annals of Applied Statistics*, 4(3), 1517–1532.
- Albert, P. S., & Shih, J. H. (2010b). On Estimating the Relationship between Longitudinal Measurements and Time-to-Event Data Using a Simple Two-Stage Procedure. *Biometrics*, 66(3), 983–987.
- Austin, P. C. (2012). Generating survival times to simulate Cox proportional hazards models with time-varying covariates. *Statistics in Medicine*, 31(29), 3946–3958.
- Balkau, B. (1996). An epidemiologic survey from a network of French Health Examination Centres, (D.E.S.I.R.): epidemiologic data on the insulin resistance syndrome. *Revue D'épidémiologie et de Santé Publique*, 44(4), 373–375.
- Bouatia-Naji, N., Rocheleau, G., Van Lommel, L., Lemaire, K., Schuit, F., Cavalcanti-Proença, C., ... Froguel, P. (2008). A polymorphism within the G6PC2 gene is associated with fasting plasma glucose levels. *Science (New York, N.Y.)*, 320(5879), 1085–1088.
- Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, 4, 7.
- Chen, L. M., Ibrahim, J. G., & Chu, H. (2011). Sample size and power determination in joint modeling of longitudinal and survival data. *Statistics in Medicine*, 30(18), 2295–2309.
- Diggle, P., & Kenward, M. G. (1994). Informative Drop-Out in Longitudinal Data Analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 43(1), 49–93.
- Dupuis, J., Langenberg, C., Prokopenko, I., Saxena, R., Soranzo, N., Jackson, A. U., ... Barroso, I. (2010). New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nature Genetics*, 42(2), 105–116.
- Elashoff, R. M., Li, G., & Li, N. (2008). A Joint Model for Longitudinal Measurements and Survival Data in the Presence of Multiple Failure Types. *Biometrics*, 64(3), 762–771.
- Elashoff, R., Li, G., & Li, N. (2016). *Joint Modeling of Longitudinal and Time-to-Event Data* (1st ed.). Chapman and Hall/CRC.
- Ibrahim, J. G., Chu, H., & Chen, L. M. (2010). Basic Concepts and Methods for Joint Models of Longitudinal and Survival Data. *Journal of Clinical Oncology*, 28(16), 2796.
- Kalbfleisch, J. D., & Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. Hoboken, NJ, USA: John Wiley & Sons, Inc.
- Morris, A. P., Voight, B. F., Teslovich, T. M., Ferreira, T., Segrè, A. V., Steinhorsdottir, V., ... McCarthy, M. I. (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genetics*, 44(9), 981–990.

Pinheiro, J., Bates, D., & R-core. (2017). *Nlme: Linear and nonlinear mixed effects models*. Retrieved from

<https://CRAN.R-project.org/package=nlme>

Proust-Lima, C., Joly, P., Dartigues, J.-F., & Jacqmin-Gadda, H. (2009). Joint modelling of multivariate longitudinal outcomes and a time-to-event: A nonlinear latent class approach. *Computational Statistics & Data Analysis*, 53(4), 1142–1154.

Purcell, S., & Chang, C. (2015). PLINK v1.9ob3.36.

R Core Team. (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rizopoulos, D. (2010). JM: An R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software*, 35(9), 1–33.

Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data: With applications in R*. CRC Press.

Rizopoulos, D. (2016). *JM: Joint modeling of longitudinal and survival data*. Retrieved from <https://CRAN.R-project.org/package=JM>

Rizopoulos, D., & Ghosh, P. (2011). A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Statistics in Medicine*, 30(12), 1366–1380.

Rung, J., Cauchi, S., Albrechtsen, A., Shen, L., Rocheleau, G., Cavalcanti-Proen  a, C., ... Sladek, R. (2009). Genetic variant near IRS1 is associated with type 2 diabetes, insulin resistance and hyperinsulinemia. *Nature Genetics*, 41(10), 1110–1115.

Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., ... Froguel, P. (2007). A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 445(7130), 881–885.

Sun, J., Sun, L., & Liu, D. (2007). Regression Analysis of Longitudinal Data in the Presence of Informative Observation and Censoring Times. *Journal of the American Statistical Association*, 102(480), 1397–1406.

The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, 526(7571), 68–74.

Therneau, T. M. (2017). *Survival: Survival analysis*. Retrieved from <https://CRAN.R-project.org/package=survival>

Therneau, T. M., & Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. (K. Dietz, M. Gail, K. Krickeberg, J. Samet, & A. Tsiatis, Eds.). New York, NY: Springer New York.

Tsiatis, A. A., & Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview. *Statistica Sinica*, 14, 809–834.

Tsiatis, A. A., DeGruttola, V., & Wulfsohn, M. S. (1995). Modeling the Relationship of Survival to Longitudinal Data Measured with Error. Applications to Survival and CD4 Counts in Patients with AIDS. *Journal of the American Statistical Association*, 90(429), 27–37.

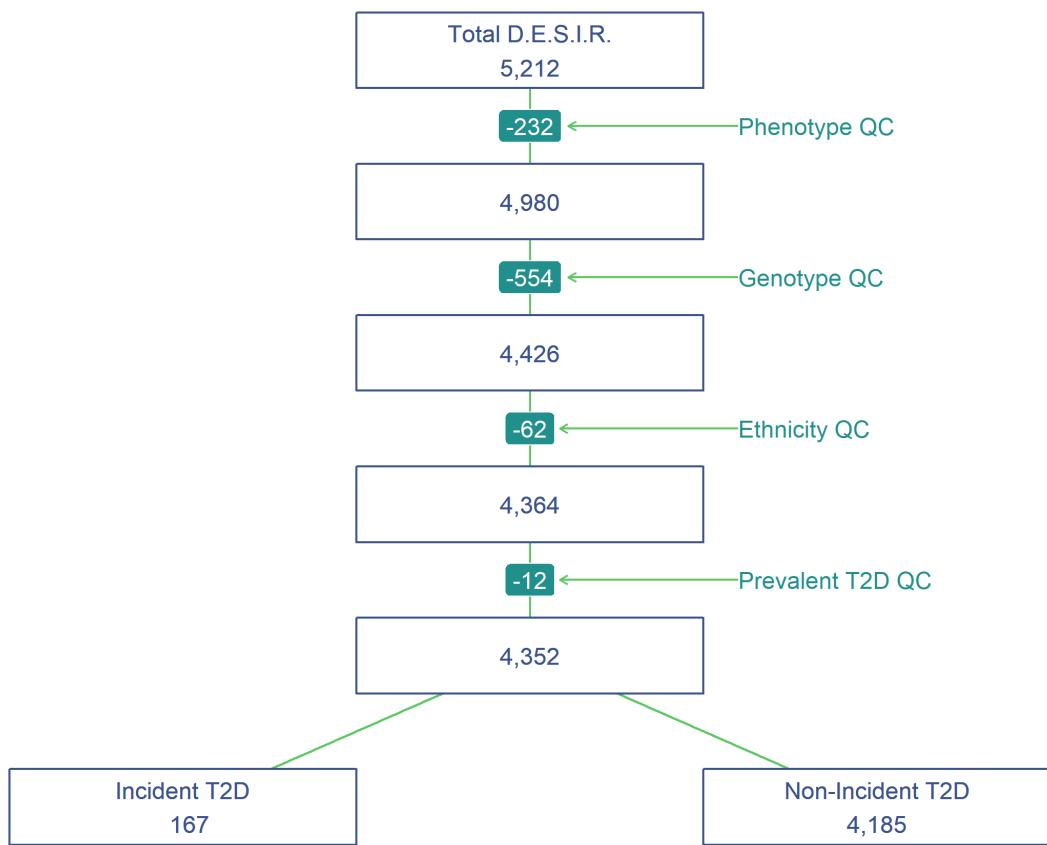
- Vaxillaire, M., Yengo, L., Lobbens, S., Rocheleau, G., Eury, E., Lantieri, O., ... Froguel, P. (2014). Type 2 diabetes-related genetic risk scores associated with variations in fasting plasma glucose and development of impaired glucose homeostasis in the prospective DESIR study. *Diabetologia*, 57(8), 1601–1610.
- Voight, B. F., Kang, H. M., Ding, J., Palmer, C. D., Sidore, C., Chines, P. S., ... Boehnke, M. (2012). The Metabochip, a Custom Genotyping Array for Genetic Studies of Metabolic, Cardiovascular, and Anthropometric Traits. *PLoS Genetics*, 8(8), e1002793.
- Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., ... Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, 42(D1), D1001–D1006.
- Wulfsohn, M. S., & Tsiatis, A. A. (1997). A Joint Model for Survival and Longitudinal Data Measured with Error. *Biometrics*, 53(1), 330.

## Supplementary

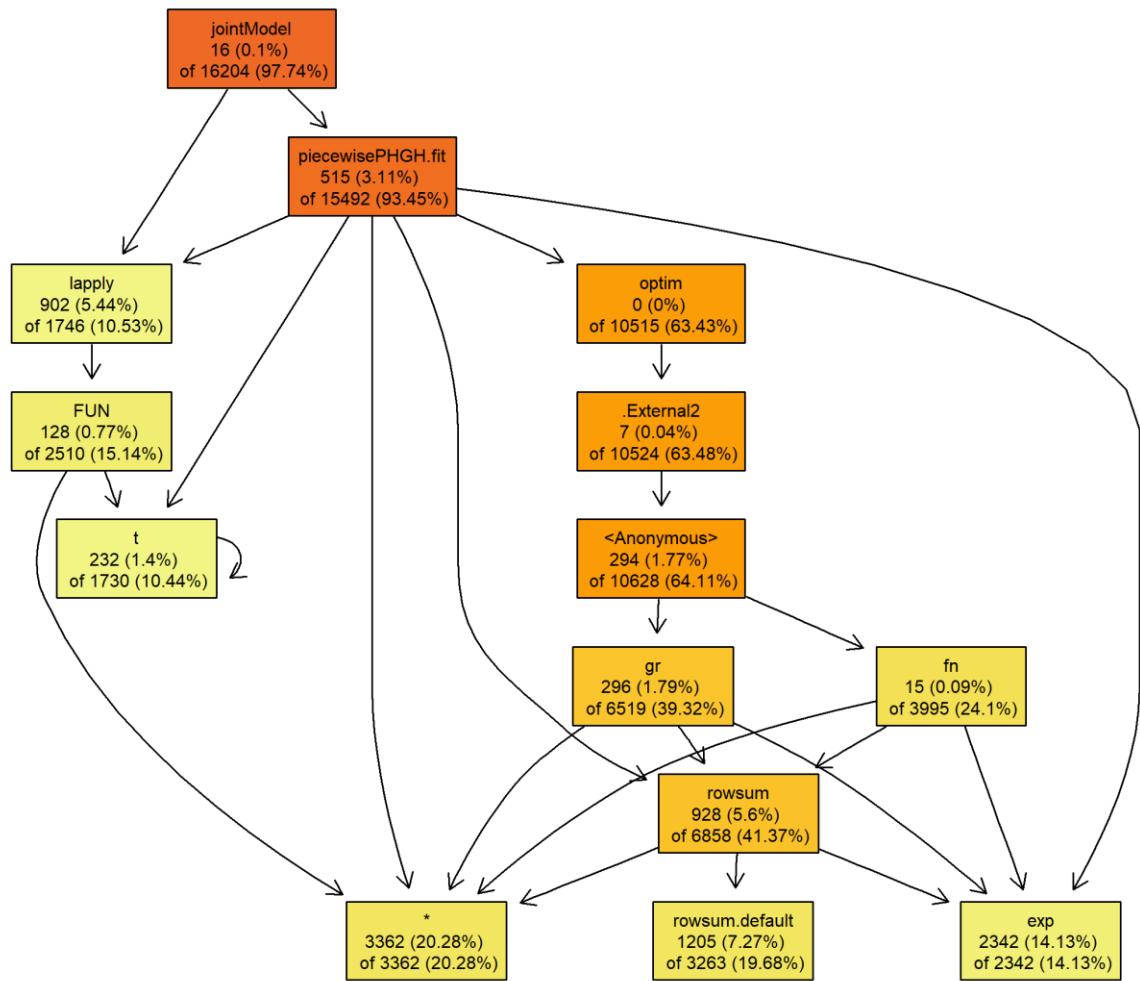
**Supplementary Table I. List of loci found to be associated within the joint modelling framework with both FPG and T2D, previously shown as associated with FPG and/or T2D in the NHGRI GWAS Catalogue (Welter et al., 2014).**

SNP (gene)	$\alpha$ (p-value)	$\gamma$ (p-value)	$\beta$ (p-value)	Power ( $\beta\gamma + \alpha$ )	Risk Allele Frequency
rs6945660_G (ETV1)	0.550 ( $3.7 \times 10^{-02}$ )	0.035 ( $2.5 \times 10^{-02}$ )	3.480 ( $9.6 \times 10^{-45}$ )	69.7%	0.91
rs1942873_C (MC4R)	0.410 ( $1.3 \times 10^{-02}$ )	0.023 ( $3.7 \times 10^{-02}$ )	3.140 ( $1.9 \times 10^{-41}$ )	69.6%	0.81
rs55899248_G (TCF7L2)	0.292 ( $2.7 \times 10^{-02}$ )	0.025 ( $1.7 \times 10^{-02}$ )	3.490 ( $1.7 \times 10^{-44}$ )	55.3%	0.24
rs17301514_A (ST6GAL1)	-0.657 ( $4.4 \times 10^{-03}$ )	0.045 ( $3.4 \times 10^{-03}$ )	3.650 ( $2.9 \times 10^{-45}$ )	45.8%	0.09
rs833425_C (PTPRD)	0.321 ( $5.0 \times 10^{-02}$ )	0.043 ( $4.2 \times 10^{-03}$ )	3.510 ( $1.3 \times 10^{-43}$ )	44.2%	0.1
rs7072870_A (C10orf35)	-0.404 ( $7.5 \times 10^{-03}$ )	0.025 ( $2.2 \times 10^{-02}$ )	3.580 ( $1.7 \times 10^{-45}$ )	39.6%	0.22
rs61871514_A (KCNQ1)	0.425 ( $4.7 \times 10^{-02}$ )	0.046 ( $2.0 \times 10^{-02}$ )	3.180 ( $8.5 \times 10^{-42}$ )	39.4%	0.06
rs9883865_A (ADAMTS9)	-0.598 ( $7.5 \times 10^{-04}$ )	0.043 ( $1.2 \times 10^{-02}$ )	3.200 ( $5.9 \times 10^{-42}$ )	34.9%	0.92
rs853787_T (ABCB11)	-0.247 ( $4.3 \times 10^{-02}$ )	0.083 ( $9.3 \times 10^{-19}$ )	3.210 ( $1.7 \times 10^{-42}$ )	3.3%	0.65
rs114508985_C (HLA)	-0.294 ( $2.1 \times 10^{-02}$ )	0.021 ( $3.0 \times 10^{-02}$ )	3.220 ( $8.2 \times 10^{-43}$ )	27.1%	0.31
rs560887_C (G6PC2)	-0.315 ( $1.2 \times 10^{-02}$ )	0.099 ( $9.6 \times 10^{-25}$ )	3.210 ( $1.3 \times 10^{-42}$ )	2.6%	0.7
rs10814856_T (GLIS3)	-0.265 ( $4.0 \times 10^{-02}$ )	0.025 ( $1.5 \times 10^{-02}$ )	3.200 ( $1.5 \times 10^{-42}$ )	18.5%	0.73
rs73025532_C (SLC22A1)	-0.377 ( $4.8 \times 10^{-02}$ )	0.032 ( $3.6 \times 10^{-02}$ )	3.580 ( $1.3 \times 10^{-45}$ )	17.3%	0.9
rs11769484_C (JAZF1)	-0.254 ( $4.8 \times 10^{-02}$ )	0.022 ( $3.6 \times 10^{-02}$ )	3.210 ( $2.1 \times 10^{-42}$ )	16.9%	0.77
rs6450176_G (ARL15)	-0.291 ( $1.8 \times 10^{-02}$ )	0.036 ( $3.0 \times 10^{-04}$ )	3.540 ( $2.2 \times 10^{-45}$ )	15.2%	0.73
rs4712580_C (CDKAL1)	-0.289 ( $4.2 \times 10^{-02}$ )	0.031 ( $7.4 \times 10^{-03}$ )	3.570 ( $1.2 \times 10^{-45}$ )	14.0%	0.82
rs10830963_G (MTNR1B)	-0.440 ( $9.4 \times 10^{-04}$ )	0.099 ( $1.3 \times 10^{-23}$ )	3.250 ( $3.6 \times 10^{-42}$ )	10.2%	0.29





**Supplementary Figure 1.** Study flowchart of people from the French cohort D.E.S.I.R.



**Supplementary Figure 2.** Call tree diagram of the main function `jointmodel` in the R package JM. Call based on a simulated dataset with three longitudinal measures and 5,000 individuals (other parameter values set as in Table I).



# Chapitre 2

---

## *L'Expression et l'Évaluation Fonctionnelle des Gènes de Susceptibilité au Diabète de Type 2 Identifient Quatre Nouveaux Gènes Contribuant à la Sécrétion d'Insuline Humaine*

---

Publié dans **Molecular Metabolism**<sup>1</sup>.

Fatou K Ndiaye<sup>1,\*</sup>, Ana Ortalli<sup>1,\*</sup>, **Mickaël Canouil**<sup>1,\*</sup>, Marlène Huyvaert<sup>1</sup>, Clara Salazar-Cardozo<sup>1</sup>, Cécile Lecoeur<sup>1</sup>, Marie Verbanck<sup>1</sup>, Valérie Pawlowski<sup>1</sup>, Raphaël Boutry<sup>1</sup>, Emmanuelle Durand<sup>1</sup>, Iandry Rabearivelo<sup>1</sup>, Olivier Sand<sup>1</sup>, Lorella Marselli<sup>2</sup>, Julie Kerr-Conte<sup>3</sup>, Vikash Chandra<sup>4</sup>, Raphaël Scharfmann<sup>4</sup>, Odile Poulain-Godefroy<sup>1</sup>, Piero Marchetti<sup>2</sup>, François Pattou<sup>3</sup>, Amar Abderrahmani<sup>1,5</sup>, Philippe Froguel<sup>1,5,†</sup> & Amélie Bonnefond<sup>1,5,†</sup>

<sup>1</sup>CNRS UMR 8199, European Genomic Institute for Diabetes (EGID), Institut Pasteur de Lille, University of Lille, 59000 Lille, France;

<sup>2</sup>Department of Clinical and Experimental Medicine, Islet Cell Laboratory, University of Pisa, 56100 Pisa, Italy; <sup>3</sup>Inserm U1190, EGID, CHU Lille, University of Lille, 59000 Lille, France; <sup>4</sup>Inserm U1016, Institut Cochin, Faculté de Médecine, Paris Descartes University, Sorbonne Paris Cité, 75014 Paris, France; <sup>5</sup>Department of Genomics of Common Disease, Imperial College London, W12 0NN London, United Kingdom.

\*Co-premier auteurs.

†Co-dernier auteurs.

---

1. <http://doi.org/10.1016/j.molmet.2017.03.011>

---

## 1 Introduction

### 1.1 Contexte/objectifs

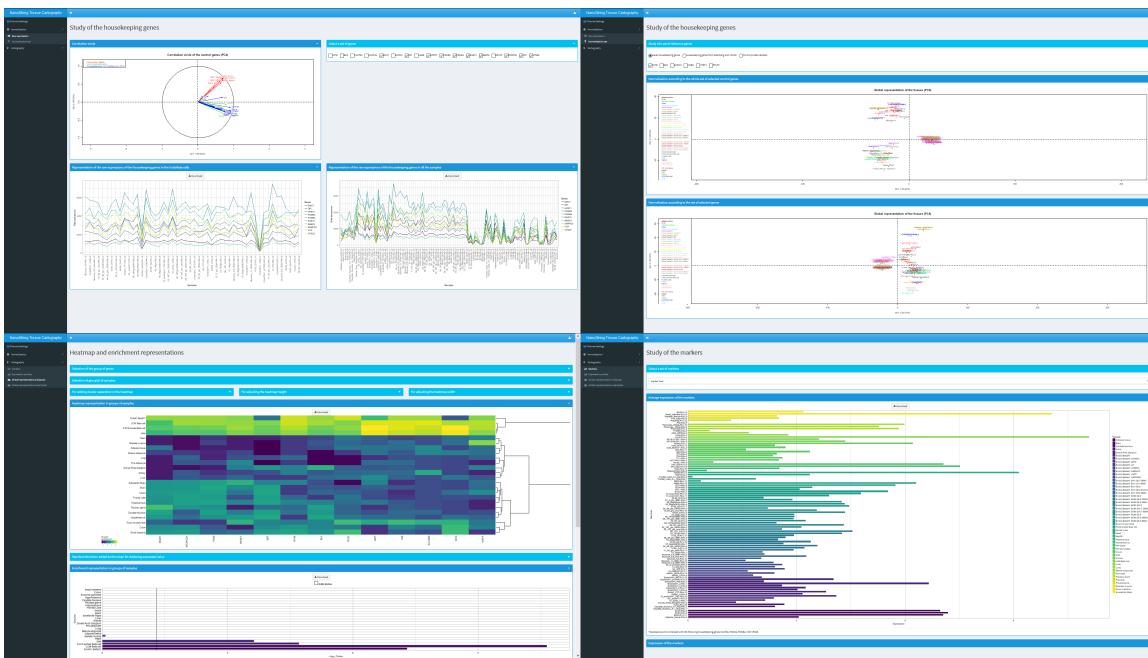
Les études d'association pangénomiques (GWAS) ont permis l'identification de plus de 100 loci associés au risque de diabète de type 2 (DT2) dont la fonction n'a pas encore été élucidée. Notre objectif dans cette étude est de palier à ce manque de connaissances quant au processus liant l'expression des gènes identifiés et la pathophysiologie du DT2, notamment au travers de l'effet de ces gènes sur la sécrétion d'insuline.

### 1.2 Méthodes

#### 1.2.1 Spécificité et enrichissement dans un panel pluritissulaire

Le transcriptome de 104 gènes candidats associés au DT2 en *cis* (c.-à-d. localisés sur le même chromosome) et des SNPs identifiés par GWAS a été étudié dans 24 organes, tissus et types cellulaires différents, incluant des échantillons de foie, muscle squelettique, cerveau, pancréas, cellules  $\beta$  pancréatiques, îlots pancréatiques, pancréas exocrine et adipocytes (primaires et matures). Un ensemble de cinq gènes de ménages a été constitué sur une base d'expression ubiquitaire, c'est-à-dire exprimés de la même façon dans les différents tissus. L'expression des gènes est ensuite mesurée pour 148 cibles (incluant les 5 gènes de ménages et les 104 gènes candidats) au moyen d'une technique sans étape d'amplification PCR (NanoString) qui peut, suite à des erreurs de copie de l'ARN polymérase, engendrer un biais des mesures. Les données transcriptomiques obtenues sont alors normalisées, en prenant la transformation logarithme en base 2, du ratio d'expression du gène d'intérêt par la moyenne d'expression des 5 gènes de ménages et ce, dans chacun des 24 tissus.

Dans un premier temps, une interface web (Figure 30), via l'extension R shiny [Chang et al., 2017], a été développée pour visualiser l'ensemble des données générées, principalement à l'aide de représentations "heatmap" et de dendrogrammes, créées à partir de la classification hiérarchique des mesures d'expression (distance de Ward sur les données centrées et réduites). Cette interface, permet la visualisation et l'identification de groupes de gènes exprimés de façon similaires entre les tissus, notamment dans les échantillons liés au pancréas, siège de la sécrétion d'insuline. Dans un second temps, les gènes ont été regroupés selon leur nature, à savoir les 104 gènes candidats, les gènes spécifiques à chaque organe (p. ex. gènes exprimés uniquement dans le foie), et les gènes identifiés dans les formes monogéniques de DT2. Pour chacun de ces ensembles, une table de contingence a été construite sur la base des comptages de gènes présentant une expression supérieure à celle observée en moyenne dans l'ensemble des tissus ( $Expr_i > \mu_{Tissus} + 1,5 \times \sigma_{Tissus}$ ). L'enrichissement en



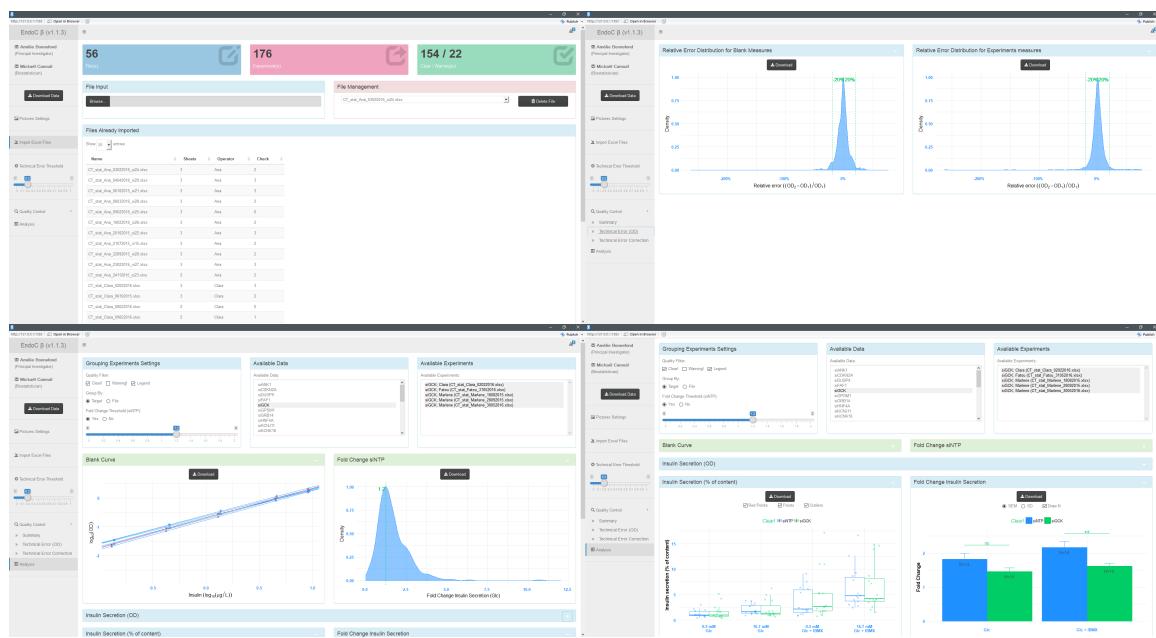
**FIGURE 30.** Application Shiny développée dans le cadre de l'analyse de l'expression de gènes de susceptibilité au diabète de type 2 dans un panel pluritissulaire.

gènes surexprimés dans un ensemble et dans un tissu donné est testé au moyen du test exact de Fisher. Les valeurs-p obtenues ont été présentées sous la forme d'un histogramme au sein de l'interface web, permettant la visualisation simultanée des résultats de l'enrichissement des ensembles de gènes, ainsi que l'homogénéité ou l'hétérogénéité de l'expression ces gènes.

### 1.2.2 Sécrétion d'insuline en réponse au glucose (modèle cellulaire)

Le rôle des gènes candidats a été ensuite étudié en diminuant leur expression au moyen de petits ARN interférents (siRNA), ayant pour fonction de cibler spécifiquement les mRNA et induire leur dégradation, dans un modèle de cellules  $\beta$  humaines (c.-à-d. EndoC  $\beta$ H1).

Le processus d'analyse comprend d'abord une étape de contrôle-qualité des différentes étapes de l'expérimentation, notamment au niveau des mesures de la gamme étalon d'absorbance, permettant d'évaluer la sécrétion d'insuline par les cellules. Cette étape consiste en l'évaluation du biais (erreur relative) entre les deux mesures d'absorbance (duplicates) des triplicats expérimentaux (cellules), et des mesures servant à établir la gamme étalon sur plus de 100 expériences. Un seuil de qualité a été défini graphiquement via la courbe de distribution des erreurs relatives des mesures d'absorbances : les expériences dont l'erreur relative était inférieure à 20 % étaient conservées pour analyse. Les expériences validant les critères du contrôle-qualité sont ensuite analysées pour deux conditions de stimulation : glucose et glucose + IBMX (3-isobutyl-1-methylxanthine). Les mesures de sécrétion d'insuline sont ainsi comparées entre les cellules contrôles (où l'expression du gène n'est pas altérée),



**FIGURE 31.** Application Shiny développée dans le cadre de l'analyse de la sécrétion d'insuline par le modèle cellulaire d'EndoC  $\beta$ H1.

et les cellules d'intérêt (où l'expression du gène est réduite via la transfection d'un siRNA), par une approche de régression linéaire avec ajustement sur les variables d'expérimentation, tels l'expérimentateur et le jour de l'expérience. Les mesures de sécrétion d'insuline sont ensuite exprimées en “Fold Change”.

L'ensemble des étapes de contrôle qualité et des analyses ont été implémentées au sein d'une interface web (Figure 31), via l'extension R shiny [Chang et al., 2017] permettant la visualisation, à la volée, de la qualité et des résultats de chaque expérience, dès leur inclusion dans l'application.

### 1.3 Résultats

#### 1.3.1 Spécificité et enrichissement dans un panel pluritissulaire

L'étude transcriptomique des 104 gènes candidats a montré que ces gènes étaient préférentiellement exprimés dans les cellules  $\beta$  du pancréas :

- cellules  $\beta$  prélevées à l'aide d'une microdissection par capture laser (LCM) : valeur-p =  $5,1 \times 10^{-4}$  (valeur-p du test exact de Fisher);
- cellules  $\beta$  triées (“Fluorescence Activated Cell Sorting” ou FACS) : valeur-p =  $1,6 \times 10^{-3}$ ;
- modèle de cellules  $\beta$  (EndoC  $\beta$ H1) : valeur-p =  $1,6 \times 10^{-3}$ ;

mais aucun enrichissement significatif de ces gènes n'a pu être montré dans les tissus cibles de l'insuline (c.-à-d. foie, muscle squelettique et tissu adipeux).

### 1.3.2 Sécrétion d'insuline en réponse au glucose (modèle cellulaire)

L'étude de la sécrétion de l'insuline par les EndoC  $\beta$ H1, pour les gènes dont la transfection de siRNA a réussi, a permis l'identification et la confirmation de sept gènes (*GCK*, *HNF4A*, *TCF19*, *SLC30A8*, *TBC1D4*, *CDKN2A* et *KNCK16*) connus pour être exprimés ou ayant un rôle dans les cellules  $\beta$ , et particulièrement sur la sécrétion d'insuline. L'approche développée ici a également permis de mettre en lumière quatre gènes candidats additionnels (*PRC1*, *SRR*, *ZFAND3* et *ZFAND6*), pouvant impacter la sécrétion d'insuline, et dont le rôle dans la cellule  $\beta$  en fait de bons candidats pour l'étude des mécanismes liant la cellule  $\beta$  au développement d'un diabète de type 2. L'expression de ces gènes a été validée par immunofluorescence. De plus, une corrélation positive significative a été retrouvée dans les îlots de cellules  $\beta$  pancréatiques de souris entre l'expression de l'insuline et l'expression de ces quatre gènes. Enfin, un séquençage de l'ARN d'EndoC  $\beta$ H1, transfectées avec *siPRC1*, *siSRR*, *siZFAND6*, *siZFAND3* ou *siNTP* (contrôle), a été réalisé afin d'identifier des voies physiopathologiques pouvant expliquer la corrélation avec la sécrétion d'insuline observée (p. ex. réseau de gènes liés à l'apoptose des cellules  $\beta$  pancréatiques, au stress du réticulum endoplasmique, etc.)

## 1.4 Conclusion

Les développements statistiques apportés dans cette étude fonctionnelle, quoique non directement appliquée à l'ensemble des gènes localisés au voisinage des SNPs identifiés par GWAS ou méta-analyses, se révèlent des outils robustes ayant permis de mettre en évidence un aspect plus mécanistique/pathophysiologique des loci identifiés par les approches GWAS, augmentant ainsi la compréhension des maladies complexes.

---

## 2 Article

Article disponible en ligne sur **Molecular Metabolism** (<http://doi.org/10.1016/j.molmet.2017.03.011>)



# Chapitre 3

---

## *La Surexpression Hépatique de PDGF-AA Affaiblit la Signalisation de l'Insuline dans le Diabète*

---

Soumis à **Nature Communications**.

Amar Abderrahmani<sup>1,2\*</sup>, Loïc Yengo<sup>1\*</sup>, Robert Caiazzo<sup>3\*</sup>, **Mickaël Canouil**<sup>1\*</sup>, Stéphane Cauchi<sup>1</sup>, Violeta Raverdy<sup>2</sup>, Valérie Plaisance<sup>1</sup>, Stéphane Lobbens<sup>1</sup>, Julie Maillet<sup>1</sup>, Laure Rolland<sup>1</sup>, Raphael Boutry<sup>1</sup>, Maxime Kwapich<sup>1</sup>, Mathie Tenenbaum<sup>1</sup>, Julien Bricambert<sup>1</sup>, Sophie Saussenthaler<sup>4</sup>, Elodie Anthony<sup>5</sup>, Pooja Jha<sup>6</sup>, Julien Derop<sup>1</sup>, Olivier Sand<sup>1</sup>, Iandry Rabearivelo<sup>1</sup>, Audrey Leloiré<sup>1</sup>, Marie Pigeyre<sup>2</sup>, Martine Daujat-Chavanieu<sup>7</sup>, Sabine Gerbal-Chaloin<sup>7</sup>, Tasnim Dayeh<sup>8</sup>, Guillaume Lassailly<sup>2</sup>, Philippe Mathurin<sup>9</sup>, Bart Staels<sup>10</sup>, Johan Auwerx<sup>5</sup>, Annette Schürmann<sup>4</sup>, Catherine Postic<sup>5</sup>, Clemens Schafmayer<sup>11</sup>, Jochen Hampe<sup>12</sup>, Amélie Bonnefond<sup>1,2</sup>, François Pattou<sup>3†</sup> & Philippe Froguel<sup>1,2†</sup>

<sup>1</sup>Univ. Lille, CNRS, Institut Pasteur de Lille, UMR 8199 - EGID, F-59000 Lille, France; <sup>2</sup>Department of genomics of common disease, Imperial College London, UK; <sup>3</sup>Univ. Lille, Inserm, CHU Lille, U1190 - EGID, F-59000 Lille, France; <sup>4</sup>Department of Experimental Diabetology, German Institute of Human Nutrition Potsdam-Rehbrücke, Nuthetal and German Center for Diabetes Research (DZD), München-Neuherberg, Germany; <sup>5</sup>Inserm, U1016, Institut Cochin, Paris, France CNRS UMR 8104, Paris, France Université Paris Descartes, Sorbonne Paris Cité, Paris, France; <sup>6</sup>Laboratory of Integrative and Systems Physiology, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland; <sup>7</sup>INSERM U1183, Univ. Montpellier, UMR1183, Institute for Regenerative Medicine and Biotherapy, CHU Montpellier, France; <sup>8</sup>Department of clinical science; Skane University Hospital Malmö, Malmö, Sweden; <sup>9</sup>Univ. Lille, Inserm, CHU Lille, U995 - LIRIC - Lille Inflammation Research International Center, F-59000 Lille, France; <sup>10</sup>Univ. Lille, Inserm, CHU Lille, Institut Pasteur de Lille, U1011- EGID, F-59000 Lille, France; <sup>11</sup>Department of Visceral and Thoracic Surgery, University Hospital Schleswig-Holstein, Kiel, Germany; <sup>12</sup>Medical Department1, Technische Universität Dresden (TU Dresden), Dresden, Germany.

\*Co-premier auteurs.

†Co-dernier auteurs.

## 1 Introduction

### 1.1 Contexte/objectifs

Les études d'association pangénomique (GWAS) n'ont pu expliquer qu'environ 15 % de l'hérédité du diabète de type 2 (DT2). Les mécanismes sous-jacents à la pathophysiologie du DT2 et aux complications dérivées de celui-ci, comme les stéatoses hépatiques non-alcooliques (NAFLD : "Non-Alcoholic Fatty Liver Disease"; NASH : "Non-Alcoholic SteatoHepatitis"), restent en grande partie méconnus. Les modifications épigénétiques de l'ADN dans un tissu clé tel que le foie pourraient contribuer à expliquer une partie de cette hérédité manquante dans le DT2 et/ou fournir des indications sur le lien entre le DT2 et ses complications. Une étude du méthylome et du transcriptome du foie de patientes obèses a été réalisée selon une approche cas/témoins (96 DT2 contre 96 normoglycémiques).

### 1.2 Méthodes

#### 1.2.1 Génome

Le génotypage des 192 individus provenant de la cohorte ABOS (Atlas Biologique de l'Obésité Sévère), a été réalisé au moyen de la puce Illumina Metabochip, une puce personnalisée comportant 200 000 SNPs, dont environ 120 000 situés près de 257 loci identifiés par GWAS pour plusieurs traits comme le diabète de type 2 ou la glycémie. Ces données de génotypage ont fait l'objet d'un contrôle qualité visant à exclure les SNPs dont le taux de génotypage était inférieur à 95 % et dont l'équilibre de Hardy-Weinberg n'était pas respecté (valeur-p  $\leq 10^{-4}$ ). Afin de vérifier l'homogénéité ethnique de nos individus, une analyse en composantes principales a été réalisée sur un jeu de données combinant les génotypes (196 470 SNPs) des 192 individus, aux génotypes de 272 individus provenant du projet de génotypage HapMap comportant 87 individus caucasiens (Europe de l'ouest et du nord), 97 individus asiatiques (Chine, Beijing) et 88 individus africains (Nigéria). À partir des SNPs précédemment identifiés par GWAS (disponibles sur la puce Illumina Metabochip) comme étant associés à l'insulinémie à jeun (19 SNPs), à la glycémie à jeun (24 SNPs), au risque de DT2 (65 SNPs) et à l'indice de masse corporelle (97 SNPs), quatre scores de risque génétique (GRS) ont été construits en prenant la somme des allèles à risque portés par chaque individu.

### 1.2.2 Méthylome

L'ensemble des sites de méthylation disponibles au sein de la puce Illumina HumanMethylation450 ont été analysés dans l'objectif d'identifier des marques de méthylation associées au statut diabétique des individus. Une étape de contrôle qualité des données brutes (format IDAT) a été réalisée sur la base de deux critères :

- exclusion des individus avec moins de 75 % des sites de méthylation détectés (valeur-p < 10<sup>-16</sup>) selon le logiciel Illumina GenomeStudio (outil permettant la lecture des images de fluorescence) ;
- exclusion des sites de méthylation lorsque le niveau de méthylation n'a pu être détecté (valeur-p < 10<sup>-16</sup>) dans au moins 95 % des individus, selon le logiciel Illumina GenomeStudio.

Suite à l'application de ces critères, l'ensemble des individus et environ 85 % (416 693) des sites de méthylation ont été conservés pour analyse.

La puce Illumina HumanMethylation450 comporte deux technologies de détection des niveaux de méthylation (valeur- $\beta$ ) : sondes Infinium I et sondes Infinium II. Une étape de normalisation est effectuée, d'une part, pour corriger un éventuel effet plaque, et d'autre part, pour corriger les différences de méthylation entre les deux types de sondes, en particulier au niveau de la distribution des niveaux de méthylation. La méthode BMIQ ("Beta-Mixture Quantile normalisation") a permis de normaliser la distribution des valeurs- $\beta$  des sondes Infinium II par rapport à celles des sondes Infinium I, tout en conservant la variabilité biologique inhérente aux différentes sondes (p. ex. Infinium I principalement employée pour les îlots CpG tandis que Infinium II est employée pour des sites CpG isolés), ainsi que le caractère monotone (rang) des valeurs- $\beta$  pour chaque type de sonde. Cette méthode consiste à modéliser la distribution des valeurs- $\beta$  (Infinium II) sur la base de trois états de méthylation, soit méthylé, semi-méthylé et non-méthylé, dont les paramètres sont estimés au moyen d'un algorithme EM (Espérance-Maximisation). Une transformation quantile des valeurs- $\beta$  (Infinium II) est ensuite appliquée sur la base de la distribution estimée. Une analyse en composantes principales est réalisée sur l'ensemble des données afin d'identifier une potentielle structure (p. ex. plusieurs sous-populations) et de potentiels individus extrêmes en termes de profil de méthylation.

Les sites différentiellement méthylés, selon le statut diabétique, ont été identifiés au moyen d'une régression linéaire avec un ajustement sur le niveau de stéatose (en pourcentage), la NASH (trait binaire) et la fibrose (trait binaire) en plus des covariables classiques tels l'âge et l'indice de masse corporelle (IMC). Les valeurs-p sont ensuite corrigées par un facteur d'inflation selon la méthode du "contrôle génomique" utilisée dans les GWAS pour corriger des effets liés à une éventuelle stratification du groupe cas et du groupe témoin menant à une inflation de l'erreur de type 1. Enfin, une correction de Bonferroni pour tests multiples est appliquée, produisant un seuil de significativité nominal 10<sup>-7</sup>.

### 1.2.3 Transcriptome

Le transcriptome a été étudié sur l'ensemble des individus via la puce Illumina HumanHT-12 v4, qui permet de mesurer l'ARN sur plus de 47 000 sondes, dont moins de 4 000 correspondant à de l'ARN non-codant. Après lecture de la fluorescence et attribution d'une valeur-p par le logiciel Illumina GenomeStudio, les données provenant de deux expérimentations font l'objet d'une normalisation quantile pour corriger les différences de distribution entre les mesures d'expression des deux séries de puces. Les sondes d'expression ne présentant pas une valeur-p de détection inférieure à  $\alpha = 0,05$  pour l'ensemble des individus étaient exclues, aboutissant à 18 412 sondes (13 664 gènes) et 189 individus conservés pour les analyses. À cela s'ajoute l'exclusion de deux individus présentant des profils transcriptomiques extrêmes, et détectés au moyen d'une analyse en composantes principales. L'expression différentielle selon le statut DT2 a été testée par une approche de régression linéaire avec ajustement pour l'âge et l'IMC, et dont la significativité a été évaluée sur la base d'un seuil FDR ("False Discovery Rate") à 5 %.

## 1.3 Résultats

### 1.3.1 Diabète de type 2

Après correction des valeurs-p, le site CpG cg14496282 localisé sur le gène *PDGFA* ("Platelet-Derived Growth Factor subunit A") a été mis en évidence comme associé significativement au risque de DT2 ( $\beta = -15,6\%$ ; valeur-p =  $2,5 \times 10^{-8}$ ), avec une hypométhylation chez les DT2 (41,3 % en moyenne) et une hyperméthylation chez les témoins (60,3 % en moyenne). Cette association persiste lorsqu'on ajuste sur la composition cellulaire du tissu ( $\beta = -14,9\%$ ; valeur-p =  $6,9 \times 10^{-7}$ ), celle-ci étant évaluée au moyen d'une méthode d'estimation de la contribution au méthylome global, d'un nombre donné de types cellulaires déterminés selon une approche "Bootstrap" (méthode de rééchantillonnage). De plus, une étude de réPLICATION dans une cohorte allemande comportant 12 cas et 53 témoins, a montré des résultats cohérents avec notre étude ( $\beta = -14\%$ ; valeur-p = 0,01). L'étude de l'expression du gène *PDGFA* démontre que celle-ci est inversement corrélée au niveau de méthylation du site cg14496282. La méthylation du site cg14496282 a également été montrée dans notre étude (dans le groupe témoin) comme étant associée à une diminution de l'insulinémie à jeun et de l'insulinorésistance (indice HOMA2-IR), pendant que l'expression de *PDGFA* était associée à une augmentation de l'insuline à jeun et une diminution de l'insulinorésistance.

### 1.3.2 Complication : stéatose/fibrose

La méthylation de cg14496282 a également été montrée comme étant associée à une diminution du risque de NASH C chez les individus diabétiques et les individus normoglycémiques, alors que la méthylation était associée à une diminution de fibrose hépatique, et l'expression à une augmentation de fibrose hépatique chez les DT2. Ces résultats sont cohérents avec des études ayant précédemment montré que l'activation du récepteur PDGF stimule les cellules stellaires et accroît ainsi la fibrose du foie. De plus, il a été montré que la surexpression de *Pdgfa* dans le foie de souris engendre une fibrose spontanée du foie.

### 1.3.3 PDCFA et action de l'insuline

L'association négative entre le GRS associé à l'insulinémie et la méthylation de cg14496282 (1,05%; valeur-p =  $4 \times 10^{-3}$ ), demeurant robuste aux ajustements à l'IMC, au cholestérol (HDL) ou aux triglycérides, suggère que l'hyperinsulinémie contribue à la modification (diminution) du niveau de méthylation du site cg14496282 de PDCFA. Dans le même temps, aucune association entre la méthylation de cg14496282 et les GRS associés à la glycémie à jeun, le statut DT2 et l'obésité (IMC), n'a été observée dans notre étude. La relation entre l'expression de PDCFA et la méthylation de cg14496282 dans des conditions hyperinsulinémiques a également été vérifiée *in vitro* (c.-à-d. avec des hépatocytes primaires humains et des hépatocytes humains immortalisés). Un modèle murin a permis l'étude de l'expression de *Pdgfa*, qui était augmentée sous stimulation insuline. Cependant, la méthylation de cg14496282 n'a pu être étudiée dans ce dernier modèle, puisque le site n'est pas conservé entre l'Homme et la souris.

## 1.4 Conclusion

La contribution de l'épigénétique dans la pathophysiologie du DT2, notamment dans la dérégulation des fonctions du foie, reste compliquée à appréhender mais fournit tout de même de nouvelles pistes d'investigation. En effet, l'étude *in vitro* et *in vivo*, en plus de l'étude génétique au moyen des GRS dans notre étude, nous indique que l'hyperinsulinémie pourrait avoir un effet causal sur la méthylation de cg14496282 et sur l'expression de PDCFA. De plus, notre étude fonctionnelle suggère que PDCFA pourrait avoir un effet autocrine sur l'hyperinsulinémie, c'est-à-dire parallèlement à la stimulation de l'expression de PDCFA par la voie de l'insuline, l'expression de PDCFA induit la sécrétion d'insuline via l'activation de PKC (Protéine Kinase C).

Les associations de PDCFA et des altérations du foie (plus précisément, fibrose et stéatose) trouvées dans plusieurs études et la nôtre soutiennent l'hypothèse du rôle fibrotique de PDCFA dans le foie au moyen d'une élévation de l'insulinémie. En outre, les études fonctionnelles réalisées sur des modèles cellulaires, incluant

l'étude de la metformine (traitement principal utilisé dans le diabète de type 2), soulignent la portée de nos découvertes, en particulier, en tant que cible thérapeutique du DT2 et de ses complications. L'étude du méthylome se révèle être un outil efficace dans l'étude de la pathogénèse des maladies communes, particulièrement lorsqu'elle cible un type cellulaire ou un tissu spécifique.

---

## **2 Article**

## Liver overexpression of PDGF-AA impairs insulin signaling in diabetes

Amar Abderrahmani<sup>1,2\*</sup>, Loïc Yengo<sup>1\*</sup>, Robert Caiazzo<sup>3\*</sup>, Mickaël Canouil<sup>1\*</sup>, Stéphane Cauchi<sup>1</sup>, Violeta Raverdy<sup>2</sup>, Valérie Plaisance<sup>1</sup>, Valérie Pawlowski<sup>1</sup>, Stéphane Lobbens<sup>1</sup>, Julie Maillet<sup>1</sup>, Laure Rolland<sup>1</sup>, Raphael Boutry<sup>1</sup>, Maxime Kwapich<sup>1</sup>, Mathie Tenenbaum<sup>1</sup>, Julien Bricambert<sup>1</sup>, Sophie Saussenthaler<sup>4</sup>, Elodie Anthony<sup>5</sup>, Pooja Jha<sup>6</sup>, Julien Derop<sup>1</sup>, Olivier Sand<sup>1</sup>, Landry Rabearivelo<sup>1</sup>, Audrey Leloiré<sup>1</sup>, Marie Pigeyre<sup>2</sup>, Martine Daujat-Chavanieu<sup>7</sup>, Sabine Gerbal-Chaloin<sup>7</sup>, Tasnim Dayeh<sup>8</sup>, Guillaume Lassailly<sup>2</sup>, Philippe Mathurin<sup>9</sup>, Bart Staels<sup>10</sup>, Johan Auwerx<sup>5</sup>, Annette Schürmann<sup>4</sup>, Catherine Postic<sup>5</sup>, Clemens Schafmayer<sup>11</sup>, Jochen Hampe<sup>12</sup>, Amélie Bonnefond<sup>1,2</sup>, François Pattou<sup>3\*</sup>, Philippe Froguel<sup>1,2\*</sup>

<sup>1</sup>Univ. Lille, CNRS, Institut Pasteur de Lille, UMR 8199 - EGID, F-59000 Lille, France; <sup>2</sup>Department of genomics of common disease, Imperial College London, UK; <sup>3</sup>Univ. Lille, Inserm, CHU Lille, U1190 - EGID, F-59000 Lille, France; <sup>4</sup>Department of Experimental Diabetology, German Institute of Human Nutrition Potsdam-Rehbrücke, Nuthetal and German Center for Diabetes Research (DZD), München-Neuherberg, Germany.

<sup>5</sup>Inserm, U1016, Institut Cochin, Paris, France CNRS UMR 8104, Paris, France Université Paris Descartes, Sorbonne Paris Cité, Paris, France. <sup>6</sup>Laboratory of Integrative and Systems Physiology, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland. <sup>7</sup>INSERM U1183, Univ. Montpellier, UMR 1183, Institute for Regenerative Medicine and Biotherapy, CHU Montpellier, France; <sup>8</sup>Department of clinical science; Skane University Hospital Malmö, Malmö, Sweden; <sup>9</sup>Univ. Lille, Inserm, CHU Lille, U995 - LIRIC - Lille Inflammation Research International Center, F-59000 Lille, France; <sup>10</sup>Univ. Lille, Inserm, CHU Lille, Institut Pasteur de Lille, U1011- EGID, F-59000 Lille, France; <sup>11</sup> Department of Visceral and Thoracic Surgery, University Hospital Schleswig-Holstein, Kiel, Germany; <sup>12</sup>Medical Department 1, Technische Universität Dresden (TU Dresden), Dresden, Germany.

\*These authors equally contributed to the study

### Corresponding authors:

Philippe Froguel, MD, PhD, p.froguel@imperial.ac.uk, and Amar Abderrahmani, PhD, amar.abderrahmani@univ-lille2.fr

## Summary

Type 2 diabetes (T2D) is closely linked with non-alcoholic fatty liver disease (NAFLD) and hepatic insulin resistance, but the mechanisms of this interaction are still elusive. In the liver from obese individuals, we found a decreased methylation level of a CpG site in *PDGFA* (encoding platelet derived growth factor alpha known as a hepatic fibrosis marker), which was associated with increased *PDGFA* expression, systemic insulin resistance and steatohepatitis. Both genetic risk score (GRS) analysis and cell modeling using immortalized human hepatocytes pointed to a causative impact of high insulin levels on *PDGFA* hypomethylation and overexpression, and on PDGF-AA increased liver secretion. We found that insulin induced PDGF-AA contributes to insulin resistance through the reduction of IRS1 content and PKC $\theta$  and PKC $\epsilon$  activation. Furthermore, hepatocyte insulin sensitivity was restored by PDGF-AA blocking antibodies, PDGF receptors inhibitors or metformin. Therefore, in T2D, the increased hepatic PDGF-AA signaling caused by hyperinsulinemia aggravates hepatocyte insulin resistance, opening new therapeutic avenues against T2D and NAFLD.

**KEY WORDS:** Type 2 diabetes, Obesity, Liver, DNA methylation, Cell models, Mouse models, Insulin resistance, Epigenetics

## Introduction

Genome-wide association studies (GWAS) for type 2 diabetes (T2D) and related metabolic traits have identified many loci associated with the risk of T2D<sup>1</sup>. However, these loci only explain 15% of T2D inheritance, and they have hardly opened new insights into the pathophysiology of T2D and its related complications, e.g. non-alcoholic fatty liver disease (NAFLD) that hits >70% of diabetic patients<sup>2</sup>. Epigenetic alteration of DNA methylation in key metabolic tissues such as the liver may contribute to T2D inheritance, providing novel mechanisms linking the pathogenic environment with T2D and complications<sup>3</sup>. Recent DNA methylation studies in adipose tissue, muscle, pancreatic islets and liver from small size cohorts of subjects have reported differentially methylated genomic sites associated with T2D and/or obesity<sup>4–6</sup>. A recent large-scale epigenome-wide association study performed in blood DNA has identified many genomic sites differentially methylated according to the distribution of body mass index (BMI) which predicted future development of T2D<sup>7</sup>.

We investigated both methylome and transcriptome in livers from obese subjects presenting with T2D or with normal glucose levels. We found that a CpG site in *PDGFA*, known as a fibrosis and cancer factor, was associated with T2D risk, *PDGFA* expression, insulin resistance and steatohepatitis (NASH). We demonstrated a causative effect of high insulin serum concentrations on the methylation of this CpG site. Our human cell and animal modeling data suggested that epigenetic changes and the subsequent dramatic increase in hepatic PDGF-AA secretion links chronic hyperinsulinemia to hepatocyte insulin resistance via a vicious autocrine negative feedback loop.

## Liver epigenetic modification in T2D

The liver genome-wide methylome was assessed in 96 age- and BMI-matched obese women with T2D and 96 obese women with normal glucose levels (**Extended Data Table 1**). After adjustment for steatosis, we found that methylation at only one CpG site (cg14496282) within *PDGFA* (encoding platelet derived growth factor alpha) is genome-wide significantly associated with decreased T2D risk ( $\beta = -15.6\%$ ;  $p = 2.5 \times 10^{-8}$ ; **Fig. 1a and 1b**). We checked the methylation, at this CpG site, for possible confounding effects due to differences in cell composition<sup>8</sup> and observed consistent effects for T2D risk ( $\beta = -14.9\%$ ;  $p = 6.9 \times 10^{-7}$ ). The average DNA methylation at the cg14496282 was 41.3 % in women with T2D and 60.3 % in controls, which corresponds to a 1.46-fold decrease in the methylation level of the CpG site. We replicated this association in livers from 12 German cases with T2D and 53 German control subjects<sup>9</sup>, where T2D risk was associated with decreased methylation level at cg14496282 site ( $\beta = -14.0\%$ ;  $p = 0.01$ ) (**Extended Data Table 2**). These data were also supported by a recent study showing a decrease in *PDGFA* methylation in livers from obese men with T2D compared to non-obese controls<sup>10</sup>.

We next investigated whether the T2D-associated *PDGFA* cg14496282 hypomethylation was specific to the liver. To do so, we assessed the blood DNA methylome from 12 obese cases with T2D and 12 obese normal glucose controls presenting with extreme liver methylation levels at cg14496282. We found a significant correlation between methylation levels in blood and liver ( $r = 0.66$ ;  $p = 6.61 \times 10^{-4}$ ), and a slightly reduced methylation at the cg14496282 site ( $\beta = -1.4\%$ ;  $p = 0.01$ ) in blood of subjects with T2D when compared to controls. We also compared DNA methylation at cg14496282 in 43 liver and skeletal muscle samples from randomly selected of 192 participants, but we did not find any significant correlation ( $p > 0.05$ ).

In the 192 obese liver samples, we next investigated *cis*-located genes (within 500 kb around cg14496282) that were differentially expressed between T2D cases and controls, and which mRNA expression correlated with DNA methylation at *PDGFA* cg14496282 site. Using a false discovery rate threshold of five percent for differential expression analysis and methylation-expression correlation analysis, we identified that methylation at cg14496282 is negatively associated with the expression of *PDGFA* in T2D cases and normal glucose controls ( $p < 0.007$ ; **Table 1**).

## Decreased NASH and insulin resistance with hypomethylated *PDGFA*

In normoglycemic obese controls, we next found that *PDGFA* cg14496282 methylation is significantly associated with decreased fasting serum insulin levels and decreased insulin resistance as modeled by the homeostasis model assessment index HOMA2-IR ( $\beta = -1.45 \times 10^{-3}$ ,  $p = 2.32 \times 10^{-3}$ ; and  $\beta = -0.10$ ,  $p = 4.93 \times 10^{-3}$ , respectively; **Table 1**). In contrast, *PDGFA* liver expression was significantly associated with increased fasting serum insulin levels and increased insulin resistance ( $\beta = 6.83 \times 10^{-3}$ ,  $p = 9.49 \times 10^{-3}$ ; and  $\beta = 0.53$ ,  $p = 7.47 \times 10^{-3}$ , respectively; **Table 1**). Furthermore, in subjects with T2D and in normoglycemic controls, we found that *PDGFA* cg14496282 methylation was significantly associated with decreased NASH risk ( $p < 0.05$ ; **Table 1**), while *PDGFA* expression in the liver was associated with increased NASH risk ( $p < 0.01$ ; **Table 1**). Furthermore, in patients with T2D, *PDGFA* cg14496282 methylation was significantly associated with decreased hepatic fibrosis, decreased alanine aminotransferase levels and decreased aspartate aminotransferase levels ( $p < 0.05$ ; **Table 1**), while *PDGFA* expression in the liver was associated with increased hepatic fibrosis and increased liver enzyme levels ( $p < 0.01$ ; **Table 1**). These results were in line with previous studies which showed that *PDGFA* cg14496282 hypomethylation is associated with increased *PDGFA* liver expression in advanced *versus* mild human NAFLD<sup>11,12</sup>. Moreover, it has been previously demonstrated that the activation of PDGF receptor signaling stimulates hepatic stellate cells and thereby, promotes liver fibrosis<sup>13–15</sup>. Moreover, the overexpression of *Pdgfa* in mice liver was found to cause spontaneous liver fibrosis<sup>16</sup>.

## Insulin modifies PDGFA methylation and expression

Subsequently, we calculated a genetic risk score (GRS) as the sum of alleles increasing fasting insulin levels over 19 GWAS-identified single nucleotide polymorphisms (SNPs)<sup>17</sup>, and found that this GRS is associated with decreased DNA methylation at cg14496282 ( $\beta = -1.05\%$  per allele;  $p = 4 \times 10^{-3}$ ; **Extended Data Table 3**). This association remained significant when we analyzed T2D cases and controls separately (and then meta-analyzed) or when we adjusted for BMI, HDL cholesterol or triglycerides; these traits having a genetic overlap with fasting insulin<sup>17</sup>. These results strongly suggested that hyperinsulinemia (and the subsequent insulin resistance) contributes to decreased DNA methylation of *PDGFA* cg14496282. In contrast, the GRS including 24 SNPs associated with fasting glucose, the GRS including 65 SNPs associated with T2D and the GRS including 97 SNPs associated with BMI were not associated with cg14496282 methylation (**Extended Data Table 3**).

We then investigated the effect of hyperinsulinemia on the expression of *PDGFA* in liver cells. As the *PDGFA* cg14496282 site is not conserved between humans and mice<sup>18</sup>, we used an *in vitro* model of human hepatocytes. We first confirmed the expression of *PDGFA* in freshly isolated primary human hepatocytes and immortalized human hepatocytes (IHH) cells (**Extended Data Fig. 1**). Moreover, IHH cells and primary human hepatocytes secreted PDGF-AA homodimer at comparable levels (**Fig. 2a**). Exposure of IHH cells to 100 nM human insulin for 24, 48 and 72 hours led to a two- to five-fold increase in *PDGFA* expression (**Fig. 2b**). The stimulatory effect of insulin on *PDGFA* expression was significant as early as six hours post insulin treatment (**Fig. 2c**). Importantly, we showed in this model that the rise of *PDGFA* expression by insulin correlated with the hypomethylation of the *PDGFA* cg14496282 site (**Fig. 2d**).

Although the cg14496282 CpG site is not conserved in mice, we next investigated whether insulin stimulates the expression of *Pdgfa* in mice liver. We found that acute administration of insulin to normal mice activated Akt in the liver (**Extended Data Fig. 2**) and stimulated *Pdgfa* expression (**Fig. 2e**), which is in line with our data in humans. This result indicates that in mice other mechanisms than cg14496282 methylation levels contribute to the rise of *PDGFA* expression triggered by insulin.

Moreover, parallel to increased *PDGFA* expression, we found that the abundance of the encoded protein PDGF-AA homodimer increases in response to insulin via the canonical PI3K/AKT pathway (**Fig. 2f-g**). The PI3K inhibitor wortmannin abolished the insulin-mediated increase in *PDGFA* expression and PDGF-AA homodimer abundance (**Figs. 2g and 2h**). The increase in *PDGFA* expression by insulin was not the consequence of changes in the IHH cell number, as insulin treatment for 24 hours did not modify the cell number and viability (**Extended Data Fig. 4**). Collectively, our data strongly suggested that the increase in *PDGFA* expression by insulin involves the insulin receptor signaling. In support of this hypothesis, we found

that the pharmacological inhibition of insulin growth 1 factor receptor (IGF1R) activity does not prevent the rise of *PDGFA* expression elicited by insulin (**Extended Data Fig. 3**).

The lipid accumulation caused by insulin treatment may account for the increase in *PDGFA* expression. Our results infirmed this hypothesis as we found similar intracellular neutral lipid levels in insulin-treated IHH cells *versus* control cells, and we showed that the chronic exposure of IHH cells with palmitate that affect insulin-induced AKT activation did not modify *PDGFA* expression (**Extended Data Fig. 5**). Thus, altogether our data support a direct causative role of the insulin signaling in the expression of *PDGFA*.

## PDGF-AA induces hepatic insulin resistance

We then hypothesized that the hepatic overproduction of PDGF-AA in response to chronic hyperinsulinemia may mediate hepatocyte insulin resistance via a negative autocrine feedback loop. Furthermore, it has been proposed that this growth factor may contribute to the induction of its own expression<sup>19</sup>. Such regulation may contribute to perpetuate PDGF-AA secretion and insulin resistance. This hypothesis would be in line with the increased liver *Pdgfa* expression that we identified in several insulin resistant mice models. Indeed, we found that *Pdgfa* expression is increased by 46 % in the liver from C57BL/6J (B6) mice that are susceptible to diet-induced obesity<sup>20</sup>, as compared with control mice (i.e. that do not respond to a high-fat diet) (**Fig. 3a and 3b**). Similarly, we found that liver *Pdgfa* expression is increased in New Zealand obese (NZO) diabetes-prone females<sup>21</sup> (**Fig. 3c and 3d**), and in the BXD mice fed a high-fat diet for 21 weeks when compared to control mice (**Fig. 3e**).

To investigate the negative role of PGDF-AA on the hepatocyte insulin signaling, we have first measured the PDGF-AA secretion in IHH cells cultured with insulin. We found that PDGF-AA protein concentration in the supernatant of IHH cells progressively increased in response to insulin reaching a two-fold increase after 24 hours of incubation (**Fig. 4a**). Importantly, PDGF-AA secretion from IHH cells cultured with insulin was associated with impaired AKT phosphorylation at residue serine 473 (**Fig. 4b**). In line with AKT activation pivotal role in glycogen synthesis<sup>22</sup>, we found reduced insulin-induced glycogen production in human hepatocytes (**Fig. 4c**).

RNA sequencing of IHH cells treated or not with insulin for 24 hours revealed a profound dysregulation of expression of genes involved in both carbohydrate metabolism, inflammatory and insulin signaling pathways in response to insulin. Indeed, when we grew a network based on *PDGFA* through Ingenuity Pathway Analysis (IPA), we found a significant increase in the expression of genes of the VEGF and PDGF families, including as expected *PDGFA* ( $\log_2$  Fold Change = 0.80;  $p = 1.1 \times 10^{-11}$ ) (**Extended Data Fig. 6a and Extended Data Table 5**). Subsequently, we analyzed the diseases and/or functions highlighted by the insulin-evoked deregulated expressed genes in IHH cells. Among the significant outputs, we found a

network related to the metabolism of carbohydrates that includes *PDGFA* ( $p = 1.2 \times 10^{-6}$ ; **Extended Data Fig. 6b, Extended Data Table 6**). We also identified in cells cultured with insulin a decrease in the expression of the insulin receptor substrate 1 (*IRS1*) gene (**Fig. 4d, Extended Data Fig. 6b and Extended Data Table 6**). The decreased *IRS1* expression by insulin, confirmed by western blotting, was concomitant with the decrease of AKT activation (**Fig. 4e**). Defective *IRS1* level can therefore account for the impaired insulin signaling caused by chronic hyperinsulinemia.

PDGF-AA over-secretion may have a direct causative role in the defective insulin signaling caused by chronic incubation with insulin. We found that the culture of IHH cells with PDGF-AA inhibits insulin-induced AKT activation (**Fig. 4f**). Importantly, the incubation of IHH cells with anti-PDGF-AA blocking antibodies counteracted the negative long-term effect of insulin on AKT phosphorylation (**Fig. 4g**). We then investigated the mechanism whereby PDGF-AA inhibits insulin-induced AKT activation. Human hepatocytes express PDGF receptors (PDGFR) including PDGFR and PDGFR that both bind PDGF-AA<sup>13</sup>. Since our IHH RNA sequencing revealed the expression of these receptors, we tested the role of PDGFR signaling using the PDGFR tyrosine kinase inhibitor Ki11502<sup>23</sup>. Pre-treatment of IHH cells with Ki11502 efficiently antagonized the negative effect of chronic insulin on AKT phosphorylation thus, confirming our results obtained with the anti-PDGF-AA blocking antibodies (**Fig. 4h**). The improvement of insulin signaling by Ki11502 was further associated with increased ability of insulin to stimulate glycogen synthesis (**Fig. 4i**). To further dissect the signaling pathways by which chronic insulin and PDGF-AA impair AKT activation, we performed a global measurement of serine/threonine protein kinases (STKs) using STK PamGene arrays consisting of 140 immobilized serine/threonine-containing peptides that are targets of most known kinases<sup>24</sup>. We looked for differential STK activity between control and IHH cells cultured with insulin for 24 hours. Peptides whose phosphorylation varied significantly between the two conditions were indicative of differential specific STK activities. This unbiased kinase analyses underscored significant differences in protein kinases C (PKC $\theta$  and PKC $\epsilon$ ) activities (**Fig. 4j**). The activation of these two PKCs hampers insulin signaling in response to chronic hyperlipidemia<sup>25–27</sup>. Therefore, we treated IHH cells with phorbol 12-myristate 13-acetate (PMA), a potent activator of PKCs, and retrieved AKT inhibition (**Extended Data Fig. 7a**). PKC $\theta$  and PKC $\epsilon$  kinase activities are associated with the phosphorylation at their Serine 676 and Serine 729, respectively<sup>28,29</sup>. We found a striking phosphorylation of the two PKCs, which coincided with the decreased AKT phosphorylation in IHH cells cultured with insulin for 16 hrs or 24 hrs (**Fig. 4k**). The effect of insulin on the phosphorylation of the two kinases and IRS1 content is likely to rely on PDGF-AA, as the activation of PKC $\theta$  and PKC $\epsilon$  and the decrease of IRS1 were found in IHH cells that were exposed to the PDGF-AA for 24 h (**Fig. 4l and 4m**). The decrease of IRS1 by PDGF-AA may be independent of PKC activation as the PMA was unable to mimic the effect of the growth factor on the IRS1 content (**Extended Data Fig. 7b**).

We indeed found that the culture of IHH cells with PDGF-AA stimulated *PDGFA* expression (**Fig. 4n**) and PDGF-AA secretion (**Extended Data Fig. 8**). This effect was mediated by PDGFR as the PDGFR inhibitor ki11502 prevented the rise of *PDGFA* mRNA of cells exposed to either insulin or PDGF-AA (**Fig. 4n and 4o**). Induction of *PDGFA* by PDGF-AA may require PKC activation since PMA mimicked both insulin and PDGF-AA effects on the *PDGFA* mRNA (**Fig. 4p**) and inversely, the PKC inhibitor sotrastaurin<sup>30</sup>, which inhibits PKC $\theta$  and PKC $\epsilon$ , alleviated the rise of *PDGFA* induced by insulin for 24 hr (**Fig. 4q**) and PDGF-AA (**Fig. 4r**). The most prescribed T2D drug metformin inhibits PKC $\epsilon$ <sup>31</sup>. In line with this effect, we found that metformin also efficiently abolished the expression of insulin-induced *PDGFA* mRNA (**Fig. 4s**), protein content (**Fig. 4t**) and secretion (**Fig. 4u**).

Altogether, our data support a role for liver PDGF-AA in promoting liver insulin resistance via the decrease of IRS1 and the activation of both PKC $\theta$  and PKC $\epsilon$ . In T2D, insulin induced PDGF-AA stimulates its own expression, impairing further hepatocyte insulin signaling and possibly the hepatic fibrogenesis by activating hepatic stellate cells<sup>13–15</sup>.

## Discussion

GWAS only identified few genes involved in NAFLD<sup>32</sup> and the contribution of epigenetics to T2D liver dysfunction is still elusive. While we initially identified 381 differentially methylated sites in liver from T2D obese patients (**Extended Data Fig. 9**), after adjusting for liver steatosis, we only observed one genome-wide significant T2D differentially methylated DNA site, associated with the increase of *PDGFA* expression in cis. Elevated *PDGFA* expression was also reported in biliary atresia<sup>33</sup>, and is a diagnostic and prognostic biomarker of cholangiocarcinoma, a liver cancer of increased incidence that is not associated with obesity but with severe insulin resistance<sup>34,35</sup>. Notably, we found that liver *PDGFA* cg14496282 hypomethylation and concomitant rise in liver *PDGFA* expression were associated with systemic insulin resistance in non-diabetic obese patients but not with their glucose values.

The genetic data from our analysis of GRS related to insulin resistance, together with our *in vitro* and *in vivo* mice experiments, suggested a causative effect of insulin on methylation level, hepatic expression and secretion of this growth factor. *PDGFA* encodes a dimer disulfide-linked polypeptide (PDGF-AA) that plays a crucial role in organogenesis<sup>36</sup>. The activation of the PDGF-AA receptor signaling is involved in cirrhotic liver regeneration<sup>37</sup> and the chronic elevation of PDGF-AA in mice liver induces fibrosis<sup>38</sup>. The association of increased *PDGFA* expression with liver steatosis and fibrosis observed in our study and in others<sup>11,12,39</sup>, supports a similar fibrogenic role in human liver, in which chronic hyperinsulinemia might be instrumental<sup>40</sup>. How prolonged hyperinsulinemia hampers downstream glucose metabolism in hepatocyte is still elusive, although a desensitization mechanism operating at the insulin receptor and IRS levels is likely<sup>41</sup>. We believe

that PDGF-AA may contribute to the inhibitory effect of chronic hyperinsulinemia on hepatocyte insulin signaling via a feedback autocrine loop (**Fig. 5**). While the insulin signaling is required for stimulating the PDGFA expression, PDGF-AA stimulates its induction via the activation of PKC. This vicious cycle perpetuates high PDGF-AA level and thereby continuous insulin resistance. Our data further suggested that the negative effect of PDGF-AA on insulin signaling is mediated through the decrease of IRS1 and PKC activation including PKC $\theta$  and PKC $\epsilon$ . These two kinases are known to phosphorylate IRS-1 and the insulin receptor on serine residues, that impairs the association of the insulin receptor with IRS proteins, leading to the blockade of AKT activation and of the downstream signaling<sup>26,27</sup>.

Our findings may have a major interest for the treatment of T2D and of its hepatic complications. We showed that metformin, the most-widely prescribed oral insulin sensitizer agent prevented the PDGF-AA insulin-induced vicious circle. Metformin has been specifically proposed for diabetic patients with NAFLD and hepatocarcinoma (HCC)<sup>42</sup>. Metformin reduces the risk of HCC incidence in diabetic patients in a dose-dependent manner<sup>43</sup>. Thereby metformin may improve liver insulin sensitivity at least in part through PDGF-AA liver blockade, explaining its long-term effect against HCC. Beside liver insulin sensitizers, blocking PDGF-AA activity may be a promising alternative anti-diabetic therapeutic. The anti-tumor PDGFR inhibitor imatinib demonstrated unexpected (and unexplained until now) improvement of insulin sensitivity in insulin-resistant rats<sup>44</sup> as well as a dramatic blood-glucose-lowering effect in diabetic subjects treated for leukemia<sup>45,46</sup>.

Our study also demonstrated that the human epigenome analysis, when directly performed in disease-affected tissues is an efficient tool to make progress in the pathogenesis of common diseases. Furthermore, it opens avenues in the identification of new drug targets to combat T2D, and complications linked to insulin resistance, including NAFLD and cancer.

## Experimental procedures

**Discovery study.** Liver biopsies were collected from 192 subjects from the French obesity surgery. Subjects included in the discovery study were participants of the ABOS (“Atlas Biologique de l’Obésité Sévère”) cohort (ClinicalGov NCT01129297) including 750 morbidly obese subjects whose several tissues were collected during bariatric surgery<sup>47</sup>. All subjects were unrelated, women, above 35 years of age, of European origin verified by Principal Component Analysis (PCA) using SNPs on the Metabochip array, non-smoker, non-drinker, without any history of hepatitis, and without indications of liver damage in serological analysis (normal ranges of aspartate aminotransferase, alanine aminotransferase and gamma-glutamyl transpeptidase). Overall, 96 T2D cases and 96 normoglycemic participants were selected. Normoglycemia and T2D were defined using the World Health Organization/International Diabetes Federation 2006 criteria

(Normoglycemia: fasting plasma glucose < 6.1 mmol/l or 2-h plasma glucose < 7.8 mmol/l; T2D: fasting plasma glucose ≥ 7 mmol/l or 2-h plasma glucose ≥ 11 mmol/l). Each participant of the ABOS cohort signed an informed consent. For calculation of intermediate metabolic traits (HOMA<sub>2</sub>-IR and HOMA<sub>2</sub>-B indexes), see supplemental information. All procedures were approved by local ethics committees. The main clinical characteristics were presented in **Extended Data Table 1**.

**Replication study.** The replication study was based on in silico data of liver samples analyzed by the Infinium HumanMethylation450 BeadChip, as previously reported 9. Clinical characteristics were reported in **Extended Data Table 2**. Liver samples were obtained percutaneously from subjects undergoing liver biopsy for suspected nonalcoholic fatty liver disease or intraoperatively for assessment of liver histology. Normal control samples were recruited from samples obtained for exclusion of liver malignancy during major oncological surgery. None of the normal control subjects underwent preoperative chemotherapy, and liver histology demonstrated absence of both cirrhosis and malignancy. Consenting subjects underwent a routine liver biopsy during bariatric surgery for assessment of liver affection. Biopsies were immediately frozen in liquid nitrogen, ensuring an ex vivo time of less than 40 seconds in all cases. A percutaneous follow-up biopsy was obtained in consenting bariatric patients five to nine months after surgery. Patients with evidence of viral hepatitis, hemochromatosis, or alcohol consumption greater than 20 g/day for women and 30 g/day for men were excluded. All patients provided written, informed consent. The study protocol was approved by the institutional review board (“Ethikkommission der Medizinischen Fakultät der Universität Kiel,” D425/07, A111/99) before the beginning of the study.

**Epigenome-wide DNA methylation profiling.** The epigenome-wide analysis of DNA methylation was performed using the Infinium HumanMethylation450 BeadChip (Illumina, Inc., San Diego, CA, USA) which interrogates 482,421 CpG sites and 3,091 non-CpG sites covering 21,231 RefSeq genes 48. We used 500ng of DNA from liver tissue for bisulfate conversion using the EZ DNA Methylation kit D5001 (Zymo Research, Orange, CA, USA) according to the manufacturer's instructions. Bisulfite converted DNA was amplified, fragmented and hybridized to the BeadChips following the standard Infinium protocol. All the samples were randomized across the chips and analyzed on the same machine by the same technician to reduce batch effects. After single base extension and staining, the BeadChips were imaged with the Illumina iScan. Raw fluorescence intensities of the scanned images were extracted with the GenomeStudio (V2011.1) Methylation module (1.9.0) (Illumina). The fluorescence intensity ratio was used to calculate the β-value which corresponds to the methylation score for each analyzed site according to the following equation:  $\beta\text{-value} = \text{intensity of the Methylated allele (M)} / (\text{intensity of the Unmethylated allele (U)} + \text{intensity of the Methylated allele (M)}) + 100$ . DNA methylation β-values range from zero (completely unmethylated) to one (completely methylated). All samples had high bisulfite conversion efficiency (signal intensity >4000) and

were included for further analysis based on GenomeStudio quality control steps where control probes for staining, hybridization, extension and specificity were examined. The intensity of both sample dependent and sample independent built in controls was checked for the red and green channels using GenomeStudio.

**Microarray mRNA expression analysis.** Transcriptome profiling was performed using the HumanHT-12 v4.0 Whole-Genome DASL HT Assay (Illumina). Total RNA was converted to cDNA using biotinylated oligo-dT<sub>18</sub> and random nonamer primers, followed by immobilization to a streptavidin-coated solid support. The biotinylated cDNAs were then simultaneously annealed to a set of assay-specific oligonucleotides based on content derived from the National Center for Biotechnology Information (NCBI) Reference Sequence Database (release 98). The extension and ligation of the annealed oligonucleotides generate PCR templates that are then amplified using fluorescently-labeled (P<sub>1</sub>) and biotinylated (P<sub>2</sub>) universal primers. The labeled PCR products were captured on streptavidin paramagnetic beads, to yield single-stranded fluorescent molecules which were then hybridized, via gene-specific complementarity, to the HumanHT-12 BeadChip, thereafter fluorescence intensity was measured for each bead. Hybridized chips were scanned by using iScan (Illumina) and raw measurements were extracted by GenomeStudio software version 3.0 (Illumina).

**SNP genotyping, ethnic characterization and genetic risk score.** SNP genotyping was performed with Metabochip DNA arrays (custom iSelect-Illumina genotyping arrays) using the Illumina HiScan technology and GenomeStudio software (Illumina, San Diego, CA, USA)<sup>49</sup>. We selected SNPs with a call rate  $\geq 95\%$  and with no departures from Hardy–Weinberg equilibrium ( $p > 10^{-4}$ ). A Principal Component Analysis (PCA) was performed in a combined dataset involving the 192 patients plus 272 subjects from the publicly available HapMap project database. For these 272 subjects (87 of European ancestries [HapMap CEU], 97 of Asian ancestries [HapMap CHB] and 88 of African ancestries [HapMap YRI]) genotype calls at the 106,470 SNPs present on the Metabochip were available. The first two components were sufficient to discriminate ethnic origin (**Extended Data Fig. 10**) and we observed that study participants clustered well with HapMap samples of European ancestries. We used 19 SNPs previously established for their association with fasting insulin to assess the possibility of causal direct link between DNA methylation at cg14496282. These 19 SNPs were on the Metabochip and all passed our quality control. These SNPs as well as the associated insulin raising alleles are reported in **Extended Data Table 3**. We assessed the combined effect of these SNPs on DNA methylation using either fixed-effect meta-analysis and by testing the association of DNA methylation at cg14496282 with a genetic risk score (GRS) defined for each individual as the count of the number of fasting insulin raising alleles. We also tested the association between three other GRS (T2D, BMI and fasting glucose raising alleles) and DNA methylation at cg14496282 and expression (**Extended Data Table 4**).

**Statistical Analyses.** Statistical analysis and quality control were performed with R software version 3.1.1<sup>50</sup>. Raw data (IDAT file format) from Infinium HumanMethylation450 BeadChips were imported into R using the *minfi* package (version 1.12.0 on Bioconductor)<sup>51</sup>, then we applied the preprocessing method from GenomeStudio software (Illumina) using the reverse engineered function provided in the *minfi* package. Samples were excluded when less than 75 % of the markers had detection *p*-values below  $10^{-16}$ . Markers were ruled out when less than 95 % of the samples had detection *p*-values below  $10^{-16}$ . According to this strategy, no sample was excluded and 70,314 markers (over 485,512) were excluded. To correct for Infinium HumanMethylation450 BeadChip design which includes two probe types (Type I and Type II), a Beta-Mixture Quantile normalization (BMMQ)<sup>52</sup> was performed. Moreover, we checked for outliers using Principal Component Analysis (PCA) (*flashpcaR* package, version 1.6-2 on CRAN). At this stage, 416,693 markers and 192 samples were kept for further analysis. To test the association between methylation level and diabetic status, we applied a linear regression adjusted for steatosis (in percent), presence of NASH and fibrosis. Results were corrected for multiple testing using a Bonferroni correction ( $p < 10^{-7}$ ). The association between DNA methylation and metabolic traits was analyzed using a linear regression model, including normoglycemic samples adjusted for age and BMI. Quality control was performed on the HumanHT-12 v4.0 Whole-Genome DASL HT Assay (Illumina) data, according to the following criterion: probes were kept for further analysis when the detection *p*-values provided by GenomeStudio software version 3.0 (Illumina) were below five percent for all samples. A PCA was performed to identify samples with extreme transcriptomic profiles. After the quality control just described, 18,412 probes matching 13,664 genes and 187 samples were kept and analyzed for differential expression between T2D cases and controls, using linear regression. To account for multiple testing, we used five percent as a threshold for false discovery rate (FDR). Methylation and expression data were tested for correlation.

We selected a subgroup of 24 samples among the 192 initial samples, including 12 normoglycemic and 12 T2D cases, to analyze DNA methylation in blood samples from the same donors. The 24 samples were selected based on their expression and methylation profiles using PCA to reduce the heterogeneity.

**Materials.** The PDGFR $\alpha$  inhibitor Ki11502, human PDGF-AA recombinant, Phorbol 12-Myristate 13-myristate (PMA), Metformin and Wortmannin were purchased from Sigma-Aldrich. The anti-PDGFAA blocking antibodies were from Merckmillipore. The PKC inhibitor Sautostaurin was from Selleckchem.

**Cell Culture.** Immortalized Human Hepatocytes (IHH)<sup>53</sup> were cultured in Williams E medium (Invitrogen), containing 11 mM glucose and supplemented with 10 % fetal calf serum (FCS; Eurobio), 100 U/ml penicillin, 100 µg/ml streptomycin, 20 mU/ml insulin (Sigma-Aldrich) and 50 nM dexamethasone (Sigma-Aldrich). For insulin pre-treatment, 106 cells were cultured in 6-well plates in a Dulbecco's Modified

Eagle Medium (DMEM; Invitrogen) with or without 100 nM human insulin (Novo Nordisk) supplemented with 5 mM Glucose, 2 % FCS, 100 U/ml penicillin, 100 µg/ml streptomycin for 24 hours. For monitoring insulin signaling, medium was removed and replaced by FCS- and phenol red-free DMEM medium with or without 200 nM human insulin for one hour. Human hepatocytes were isolated from liver lobectomies resected for medical reasons as described 54 in agreement with the ethics procedures and adequate authorization.

**Quantitative PCR.** Total RNA was extracted from IHH cells according to the manufacturer's protocol (RNeasy Lipid Tissue Kit, Qiagen). The RNA purity and concentration were determined by RNA Integrity Number (RNA 6000 Nano Kit, 2100 Bioanalyser, Agilent). Total RNA was transcribed into cDNA as described 55. Each cDNA sample was quantified by quantitative real-time polymerase chain reaction using the fluorescent TaqMan 5'-nuclease assays or a BioRad MyiQ Single-Color Real-Time PCR Detection System using the BioRad iQ SYBR Green Supermix, with 100 nM primers and 1 µl of template per 20 µl of PCR and an annealing temperature of 60 °C. Gene expression analysis was normalized against Beta-Glucuronidase (GUSB) expression or 60S acidic ribosomal protein Po (RPLPo). The primer sequences are available in the supplemental information.

**Western Blotting and ELISA.** Cells were scrapped in cold PBS buffer and then cells pellet was incubated for 30 minutes on ice in the following lysis buffer (20 mM Tris acetate pH 7, 0.27 mM Sucrose, 1% Triton X-100, 1 mM EDTA, 1 mM EGTA, 1 mM DTT) supplemented with antiproteases and antiphosphatases (Roche, Meylan, France). Cell lysate was centrifuged 15 minutes at 18,000g and supernatant was collected as total proteins. For Western blotting experiments, 40 µg of total protein extract was separated on 10% SDS-Polyacrylamide gel and electrically blotted to nitrocellulose membrane. The proteins were detected after an overnight incubation of the membrane at 4°C with the specific primary antibodies against AKT (Santa Cruz Biotechnology, dilution 1:1000), PKCθ (Abcam, dilution 1:1000), PKCε (Abcam, dilution 1:1000), PDGF-AA (Merck Millipore, dilution 1:1000), tubulin (Sigma, dilution 1:5000), phospho-AKT (Ser-473; Cell Signaling Biotechnology, dilution 1:1000), phospho-PKCθ (Ser-676, Abcam, dilution 1:1000), phospho-PKCε (Ser-729, Abcam, dilution 1:1000) in buffer containing 0.1% Tween 20 with either five percent BSA or five milk (for tubulin). Proteins were visualized with IRDye800 or IRDye700 (Eurobio) as secondary antibodies. Quantification was performed using the Odyssey Infrared Imaging System (Eurobio). PDGFA released in the cell supernatant was quantified by ELISA kit (R & D Systems) according to the manufacturer's protocol.

**RNA sequencing, Glycogen measurement, Global Serine/Threonine kinases activity, DNA/RNA preparation, Oil-Red staining, cell proliferation, apoptosis, intermediate metabolic traits.** See the supplemental experimental procedure in the supplemental information.

## Author contributions

PF, SC, LY and AA designed the study. LY, AA, MC, AB and PF drafted and wrote the manuscript. LY and MC performed statistical analyses. AB, OS and IR performed the bioinformatics analysis. SL, JM, JD, GL, LR, MK, MT, JB, EA, SS, PJ, RB, SGC, VP, AL and MDC performed the experiments. AA, SC, MC, RC, VR, SL, JM, LR, GL, AL, TD, PM, BS, AS, JA, CP, JH, AB, FP and PF revised the manuscript. All authors have read and approved the final version of the manuscript.

## Acknowledgements

This study was supported by nonprofit organizations and public bodies for funding of scientific research conducted in France and within the European Union: “Centre National de la Recherche Scientifique”, “Université de Lille 2”, “Institut Pasteur de Lille”, “Société Francophone du Diabète”, “Contrat de Plan Etat-Région”, “Agence Nationale de la Recherche”, ANR-10-LABX-46, ANR EQUIPEX Ligan MP: ANR-10-EQPX-07-01, European Research Council GEPIDIAB - 294785. We are grateful to Ms Estelle Leborgne for helping in the illustrations of the manuscript.

## References

1. Bonnefond, A. & Froguel, P. Rare and common genetic events in type 2 diabetes: what should biologists know? *Cell Metab.* **21**, 357–368 (2015).
2. Stefan, N. & Häring, H.-U. The Metabolically Benign and Malignant Fatty Liver. *Diabetes* **60**, 2011–2017 (2011).
3. Ling, C. & Groop, L. Epigenetics: A Molecular Link Between Environmental Factors and Type 2 Diabetes. *Diabetes* **58**, 2718–2725 (2009).
4. Kirchner, H. *et al.* Altered DNA methylation of glycolytic and lipogenic genes in liver from obese and type 2 diabetic patients. *Mol. Metab.* **5**, 171–183 (2016).
5. Muka, T. *et al.* The role of global and regional DNA methylation and histone modifications in glycemic traits and type 2 diabetes: A systematic review. *Nutr. Metab. Cardiovasc. Dis. NMCD* (2016). doi:10.1016/j.numecd.2016.04.002
6. Nilsson, E. *et al.* Altered DNA Methylation and Differential Expression of Genes Influencing Metabolism and Inflammation in Adipose Tissue From Subjects With Type 2 Diabetes. *Diabetes* **63**, 2962–2976 (2014).
7. Wahl, S. *et al.* Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature* **541**, 81–86 (2017).
8. Houseman, E. A. *et al.* Reference-free deconvolution of DNA methylation data and mediation by cell composition effects. *BMC Bioinformatics* **17**, 259 (2016).

9. Ahrens, M. *et al.* DNA methylation analysis in nonalcoholic fatty liver disease suggests distinct disease-specific and remodeling signatures after bariatric surgery. *Cell Metab.* **18**, 296–302 (2013).
10. Kirchner, H. *et al.* Altered DNA methylation of glycolytic and lipogenic genes in liver from obese and type 2 diabetic patients. *Mol. Metab.* **0**,
11. Murphy, S. K. *et al.* Relationship between methylome and transcriptome in patients with nonalcoholic fatty liver disease. *Gastroenterology* **145**, 1076–1087 (2013).
12. Zeybel, M. *et al.* Differential DNA methylation of genes involved in fibrosis progression in non-alcoholic fatty liver disease and alcoholic liver disease. *Clin. Epigenetics* **7**, 25 (2015).
13. Hayes, B. J. *et al.* Activation of Platelet-Derived Growth Factor Receptor Alpha Contributes to Liver Fibrosis. *PLoS ONE* **9**, (2014).
14. Kocabayoglu, P. *et al.*  $\beta$ -PDGF receptor expressed by hepatic stellate cells regulates fibrosis in murine liver injury, but not carcinogenesis. *J. Hepatol.* **63**, 141–147 (2015).
15. Liu, X. & Brenner, D. A. Liver: DNA methylation controls liver fibrogenesis. *Nat. Rev. Gastroenterol. Hepatol.* **13**, 126–128 (2016).
16. Thieringer, F. *et al.* Spontaneous hepatic fibrosis in transgenic mice overexpressing PDGF-A. *Gene* **423**, 23–28 (2008).
17. Scott, R. A. *et al.* Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nat Genet* **44**, 991–1005 (2012).
18. Wong, N. C. *et al.* Exploring the utility of human DNA methylation arrays for profiling mouse genomic DNA. *Genomics* **102**, 38–46 (2013).
19. Marra, F., Choudhury, G. G., Pinzani, M. & Abboud, H. E. Regulation of platelet-derived growth factor secretion and gene expression in human liver fat-storing cells. *Gastroenterology* **107**, 1110–1117 (1994).
20. Baumeier, C. *et al.* Hepatic DPP4 DNA Methylation Associates With Fatty Liver. *Diabetes* **66**, 25–35 (2017).
21. Lubura, M. *et al.* Diabetes prevalence in NZO females depends on estrogen action on liver fat content. *Am. J. Physiol. - Endocrinol. Metab.* ajpendo.00338.2015 (2015). doi:10.1152/ajpendo.00338.2015
22. Mackenzie, R. W. & Elliott, B. T. Akt/PKB activation and insulin signaling: a novel insulin signaling pathway in the treatment of type 2 diabetes. *Diabetes Metab. Syndr. Obes. Targets Ther.* **7**, 55–64 (2014).
23. Nishioka, C. *et al.* Ki11502, a novel multitargeted receptor tyrosine kinase inhibitor, induces growth arrest and apoptosis of human leukemia cells in vitro and in vivo. *Blood* **111**, 5086–5092 (2008).
24. Hilhorst, R. *et al.* in *Gene Regulation* (ed. Bina, M.) **977**, 259–271 (Humana Press, 2013).
25. Dasgupta, S. *et al.* Mechanism of lipid induced insulin resistance: Activated PKC $\epsilon$  is a key regulator. *Biochim. Biophys. Acta BBA - Mol. Basis Dis.* **1812**, 495–506 (2011).
26. Kim, J. K. *et al.* PKC- $\theta$  knockout mice are protected from fat-induced insulin resistance. *J. Clin. Invest.* **114**, 823–827 (2004).

27. Samuel, V. T. *et al.* Inhibition of protein kinase C $\epsilon$  prevents hepatic insulin resistance in nonalcoholic fatty liver disease. *J. Clin. Invest.* **117**, 739–745 (2007).
28. Cenni, V. *et al.* Regulation of novel protein kinase C $\epsilon$  by phosphorylation. *Biochem. J.* **363**, 537–545 (2002).
29. Liu, Y., Graham, C., Li, A., Fisher, R. J. & Shaw, S. Phosphorylation of the protein kinase C-theta activation loop and hydrophobic motif regulates its kinase activity, but only activation loop phosphorylation is critical to in vivo nuclear-factor- $\kappa$ B induction. *Biochem. J.* **361**, 255–265 (2002).
30. Evenou, J.-P. *et al.* The Potent Protein Kinase C-Selective Inhibitor AEB071 (Sotрастaurin) Represents a New Class of Immunosuppressive Agents Affecting Early T-Cell Activation. *J. Pharmacol. Exp. Ther.* **330**, 792–801 (2009).
31. Rodríguez *et al.* Metformin Induces Cell Cycle Arrest and Apoptosis in Drug-Resistant Leukemia Cells. *Leuk. Res. Treat.* **2015**, e516460 (2015).
32. Anstee, Q. M. & Day, C. P. The genetics of NAFLD. *Nat. Rev. Gastroenterol. Hepatol.* **10**, 645–655 (2013).
33. Cofer, Z. C. *et al.* Methylation Microarray Studies Highlight PDGFA Expression as a Factor in Biliary Atresia. *PLOS ONE* **11**, e0151521 (2016).
34. Boonjaraspinyo, S. *et al.* Platelet-derived growth factor may be a potential diagnostic and prognostic marker for cholangiocarcinoma. *Tumour Biol. J. Int. Soc. Oncodevelopmental Biol. Med.* **33**, 1785–1802 (2012).
35. Michelini, E. *et al.* Is cholangiocarcinoma another complication of insulin resistance: a report of three cases. *Metab. Syndr. Relat. Disord.* **5**, 194–202 (2007).
36. Andrae, J., Gallini, R. & Betsholtz, C. Role of platelet-derived growth factors in physiology and medicine. *Genes Dev.* **22**, 1276–1312 (2008).
37. Awuah, P. K., Nejak-Bowen, K. N. & Monga, S. P. S. Role and Regulation of PDGFR $\alpha$  Signaling in Liver Development and Regeneration. *Am. J. Pathol.* **182**, 1648–1658 (2013).
38. Thieringer, F. *et al.* Spontaneous hepatic fibrosis in transgenic mice overexpressing PDGF-A. *Gene* **423**, 23–28 (2008).
39. Kikuchi, A. & Monga, S. P. PDGFR $\alpha$  in liver pathophysiology: emerging roles in development, regeneration, fibrosis, and cancer. *Gene Expr.* **16**, 109–127 (2015).
40. Bril, F. *et al.* Relationship between disease severity, hyperinsulinemia, and impaired insulin clearance in patients with nonalcoholic steatohepatitis. *Hepatology* **59**, 2178–2187 (2014).
41. Zick, Y. Ser/Thr Phosphorylation of IRS Proteins: A Molecular Basis for Insulin Resistance. *Sci STKE* **2005**, pe4-pe4 (2005).
42. Dyson, J. K., Anstee, Q. M. & McPherson, S. Republished: Non-alcoholic fatty liver disease: a practical approach to treatment. *Postgrad. Med. J.* **91**, 92–101 (2015).
43. Bo, S., Benso, A., Durazzo, M. & Ghigo, E. Does use of metformin protect against cancer in Type 2 diabetes mellitus? *J. Endocrinol. Invest.* **35**, 231–235 (2012).

44. Hägerkvist, R., Jansson, L. & Welsh, N. Imatinib mesylate improves insulin sensitivity and glucose disposal rates in rats fed a high-fat diet. *Clin. Sci.* **114**, 65–71 (2008).
45. Breccia, M., Muscaritoli, M., Aversa, Z., Mandelli, F. & Alimena, G. Imatinib Mesylate May Improve Fasting Blood Glucose in Diabetic Ph+ Chronic Myelogenous Leukemia Patients Responsive to Treatment. *J. Clin. Oncol.* **22**, 4653–4655 (2004).
46. Veneri, D., Franchini, M. & Bonora, E. Imatinib and regression of type 2 diabetes. *N. Engl. J. Med.* **352**, 1049–1050 (2005).
47. Caiazzo, R. *et al.* Roux-en-Y gastric bypass versus adjustable gastric banding to reduce nonalcoholic fatty liver disease: a 5-year controlled longitudinal study. *Ann. Surg.* **260**, 893–898–899 (2014).
48. Bibikova, M. *et al.* High density DNA methylation array with single CpG site resolution. *Genomics* **98**, 288–295 (2011).
49. Voight, B. F. *et al.* The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet.* **8**, e1002793 (2012).
50. Team, R. C. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. (ISBN 3-900051-07-0, 2014).
51. Aryee, M. J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinforma. Oxf. Engl.* **30**, 1363–1369 (2014).
52. Teschendorff, A. E. *et al.* A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinforma. Oxf. Engl.* **29**, 189–196 (2013).
53. Samanez, C. H. *et al.* The human hepatocyte cell lines IHH and HepaRG: models to study glucose, lipid and lipoprotein metabolism. *Arch. Physiol. Biochem.* **118**, 102–111 (2012).
54. Pichard, L. *et al.* Human hepatocyte culture. *Methods Mol. Biol. Clifton NJ* **320**, 283–293 (2006).
55. Bonner, C. *et al.* Inhibition of the glucose transporter SGLT2 with dapagliflozin in pancreatic alpha cells triggers glucagon secretion. *Nat. Med.* (2015). doi:10.1038/nm.3828

## Legends of Figures

**Fig. 1.** **a)** Quantile-quantile (qq-) plot showing the residual inflation of test statistics before and after genomic-control correction. **b)** Manhattan plot centered on *PDGFA* cg14496282 methylation site showing association signal within *PDGFA* bounds.

**Fig. 2.** **a)** PDGF-AA secretion from IHH cells and primary human hepatocytes was measured by ELISA kit. **b)** and **c)** Increase of *PDGFA* mRNA by insulin. IHH cells were cultured with 100 nM human insulin (NovoNordisk) for the indicated times. The *PDGFA* mRNA level was quantified by qRT-PCR and normalized against *GUSB*. The expression levels from untreated cells were set to 100 %. Data are the mean  $\pm$  SEM (\*:  $p < 0.05$ ). **d)** Methylation levels at *PDGFA* cg14496282 in response to insulin. IHH cells were cultured in a culture medium containing 5 mM Glucose, 2 % FCS with or without 100 nM human insulin for 24 hrs. Methylation level at the cg14496282 was quantified by the Infinium HumanMethylation450 BeadChip. **e)** *PDGFA* mRNA level in mice liver in response to insulin. Insulin (I.P. 5U/Kg) or vehicle was injected for 10 minutes in overnight fasted C57Bl/6 mice males ( $n=5$ /groups). **f)** PDGF-AA abundance in IHH cells cultured with insulin. IHH cells were cultured with 100 nM human insulin for the indicated times. PDGF-AA content was quantified by Western Blotting experiments. The blot is one representative out of three independent experiments. Effect of wortmannin on **g)** the *PDGFA* mRNA by qRT-PCR and **h)** PDGF-AA protein levels by Western Blotting experiments. IHH cells were co-cultured with 100 nM human insulin in the presence of vehicle or 1 M wortmannin for the indicated times or 24 hrs.

**Fig. 3.** Hepatic *Pdgfa* expression in **a-b)** 6 weeks old male B6 mice and in **c-d)** 10 weeks old diabetes-prone female New Zealand Obese (NZO) mice. **a)** and **c)** show results of array data; **b)** and **d)** those of qRT-PCR. Differences between DIO-responder (Resp, black circle) and DIO-non-responder (nResp, white circle) mice as well as between diabetes-prone (DP, black squares) and diabetes-resistant (DR, white squares) were calculated by Student's t test. \*  $P < 0.05$ , \*\*  $P < 0.01$ , \*\*\*  $P < 0.001$ . **e)** *PDGFA* mRNA level in BXD mice fed on a HFD fed for 21 weeks. \* indicates  $p$  value  $< 0.0001$  by unpaired t test with Welch's correction.

**Fig. 4.** **a)** PDGF-AA secretion in response to insulin. IHH cells were cultured with insulin for the indicated times. The measurement of PDGF-AA from the supernatant was achieved by ELISA. **b)** Measurement of insulin-induced AKT phosphorylation in response to insulin pre-treatment. IHH cells were incubated in a culture medium containing 5 mM Glucose, 2 % FCS with or without 100 nM human insulin for the indicated times. AKT phosphorylation was stimulated by insulin for one hour. Immunoblotting for phospho-AKT (P-AKT) was done using the anti phospho-AKT (Serine 473) antibodies. The Fig. shows the result of a representative experiment out of three. **c)** Effect of insulin treatment on glycogen deposition. IHH cells were cultured with insulin for the indicated times. Thereafter, glycogen was monitored after stimulating cells in a KRP buffer without (Ctrl) or with insulin for 1 hr and 20 mM glucose. Glycogen was monitored by ELISA.

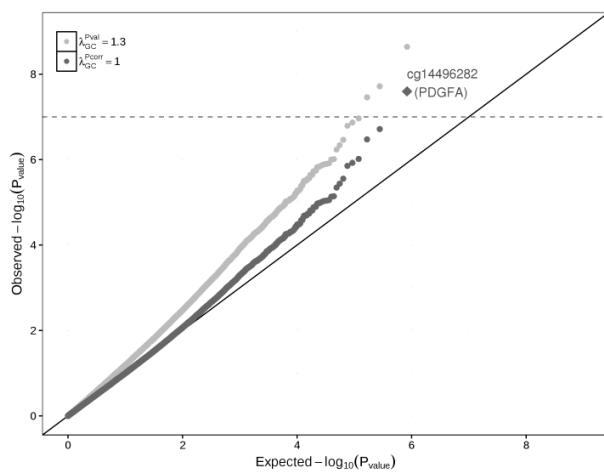
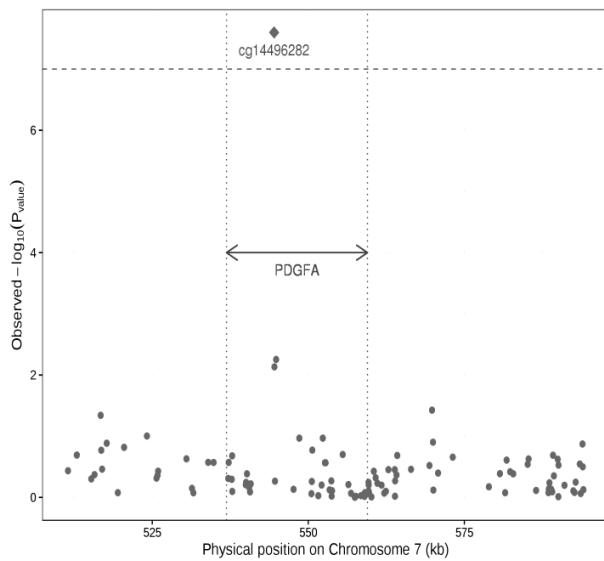
Effects of insulin on the expression of **d**) IRS1 mRNA by qRT-PCR and **e**) IRS1 protein content by Western Blotting experiments in IHH cells cultured with insulin for the indicated times. Effects of **f**) PDGF-AA, **g**) PDGFA blocking antibodies or **h**) the PDGFR inhibitor ki11502 on insulin-induced AKT activation. For **f**) activation of AKT was monitored by western blotting using total proteins from IHH cells that were cultured with the human recombinant PDGF-AA at the indicated concentrations for 24 hrs, which subsequently were incubated with 200 nM insulin for stimulating AKT phosphorylation. For **f-h**, IHH cells were co-incubated in a culture medium containing 5 mM Glucose, 2 % FCS with or without 100 nM human insulin for 24 hours plus **f**) PDGF-AA at the indicated concentration, **g**) PDGFA antibodies (+; 0.75 g or ++; 1.5 g) or **h**) ki11502 at the indicated concentration. The Figures show the result of a representative experiment out of three. **i**) Effect of the PDGFR inhibitor ki11502 on the glycogen production. Glycogen was measured by ELISA in IHH cells that were co-cultured with 5 M ki11502 and insulin for the indicated times. **j**) Volcano plot showing differences in putative serine/threonine kinase activities between control and insulin-treated IHH cells for 24 hrs. Specific and positive kinase statistic (in red) show higher activity in IHH cultured with insulin compared with control samples. Effects of **k**) insulin and **l**) PDGF-AA on the phosphorylation of PKC $\theta$  and PKC $\epsilon$ . IHH cells were cultured with insulin for the indicated times or PDGF-AA (for 24 hrs). Phosphorylation of PKC $\theta$  (Ser 676) and PKC $\epsilon$  (Ser 729) were measured by western blotting and normalized against total PKC $\theta$  and PKC $\epsilon$ . **m**) Effect of PDGF-AA on the IRS1 protein content. IHH cells were cultured for 24 hrs with PDGF-AA at the indicated concentrations. Effect of **n**) PDGF-AA and **o**) insulin on the expression of PDGFA. The PDGFA mRNA level was quantified by qRT-PCR in IHH cells co-cultured with either 100 ng/ml PDGF-AA for 24 hrs or insulin with or without the PDGFR inhibitor Ki11502 for the indicated times. Effect of **p**) PKC activator phorbol 12-myristate 13-acetate (PMA) or the PKC inhibitor sotraustorin on the expression of PDGFA mRNA either **q**) induced by insulin for the indicated times or **r**) by PDGF-AA. PDGFA mRNA was quantified in IHH cells cultured with either PMA for the indicated times or 100 ng/ml PDGF-AA in the presence or absence of 1 M sotraustorin for 24 hrs. Effect of metformin on the **s**) PDGFA mRNA level and PDGF-AA **t**) intracellular abundance and **u**) secretion induced by 100 nM insulin for the indicated times.

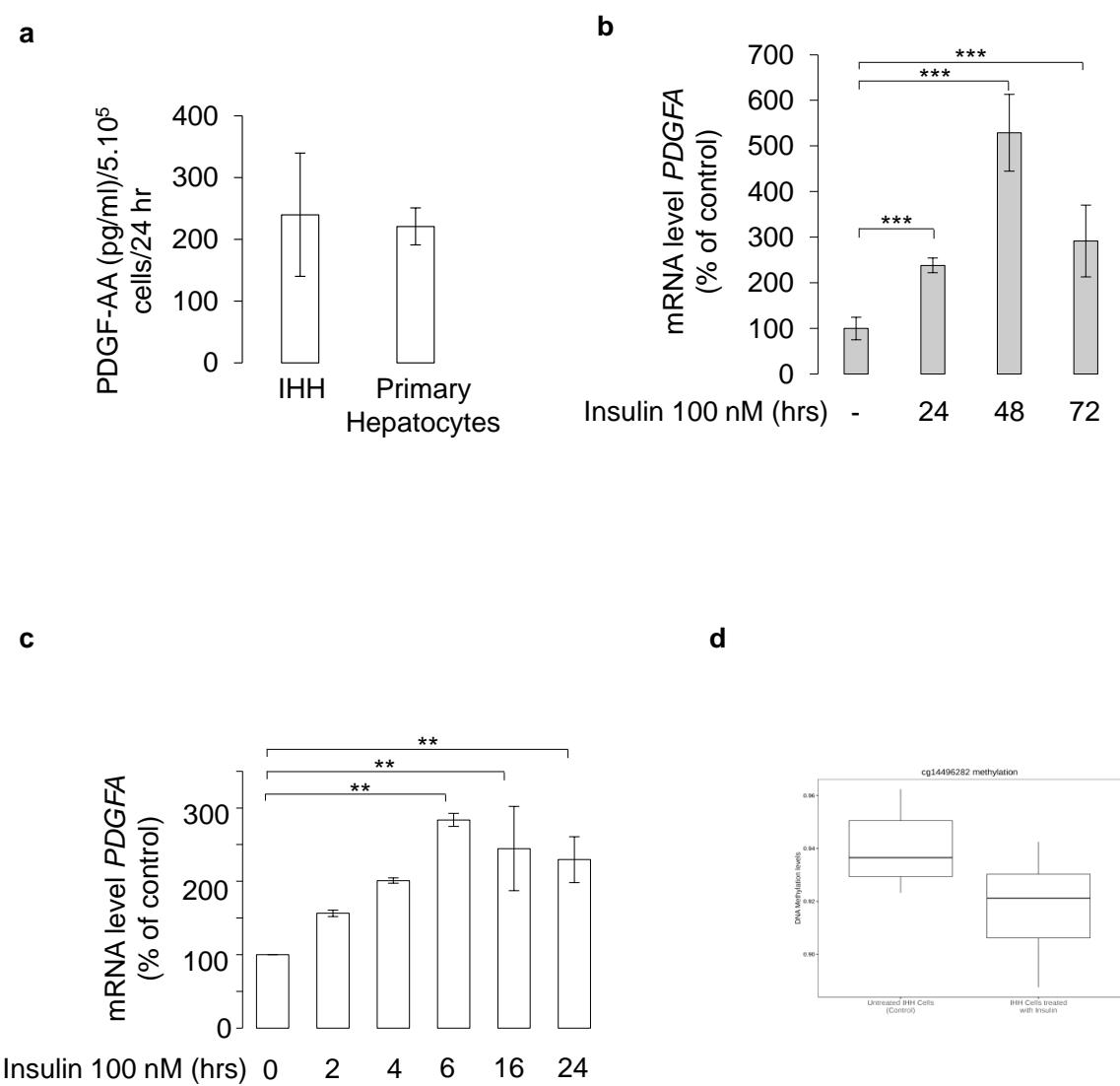
**Fig. 5:** Schematic representation of the mechanism linking chronic hyperinsulinemia to hepatic insulin resistance in T2D. Insulin promotes hypomethylation and the rise of PDGFA expression, leading to PDGF-AA secretion. In turn, PDGF-AA inhibits the insulin signaling, in a negative autocrine feedback loop, via a mechanism involving a decrease in the IRS1 abundance and PKC (PKC $\theta$  and PKC $\epsilon$ ) activation.

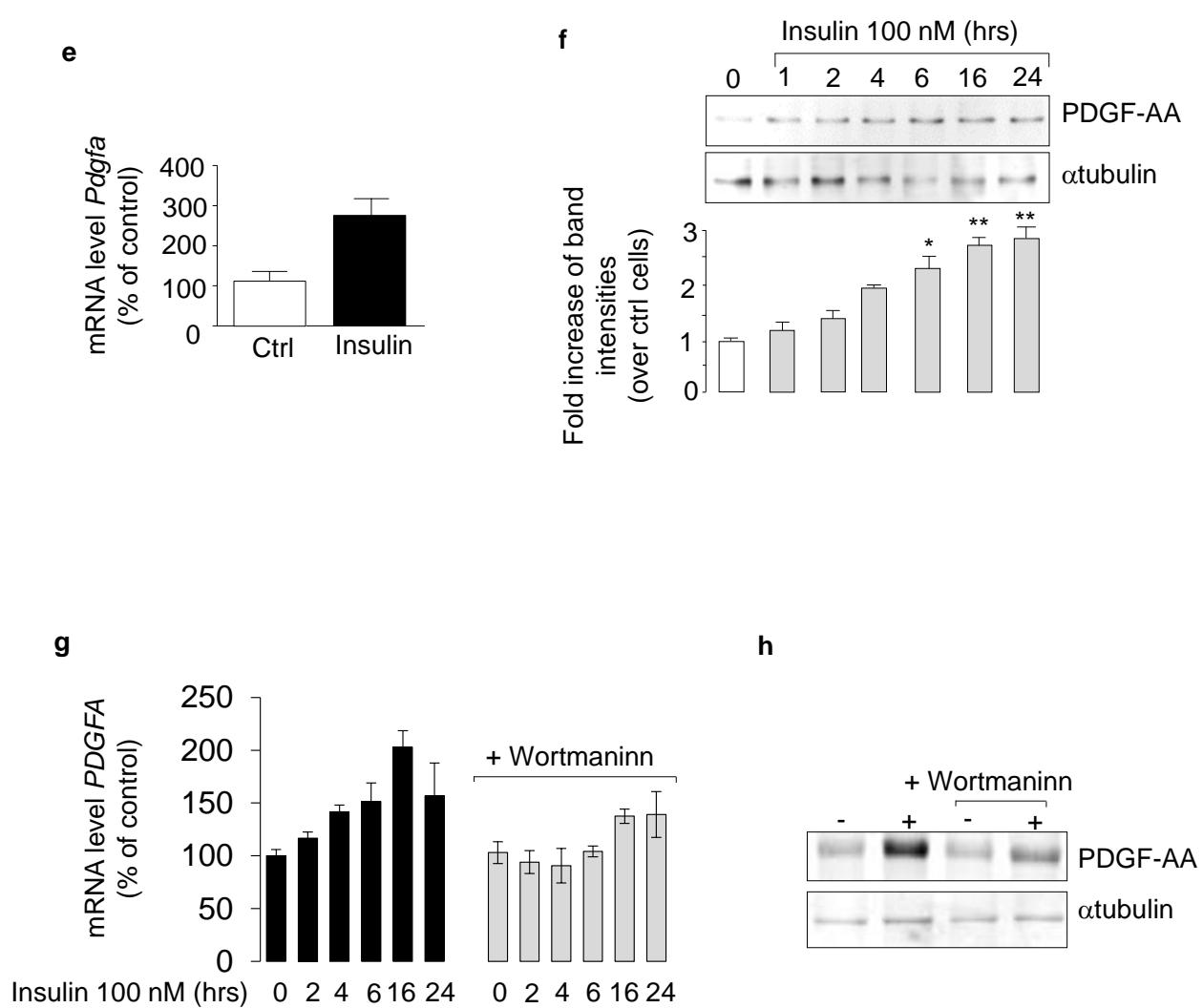
## Tables

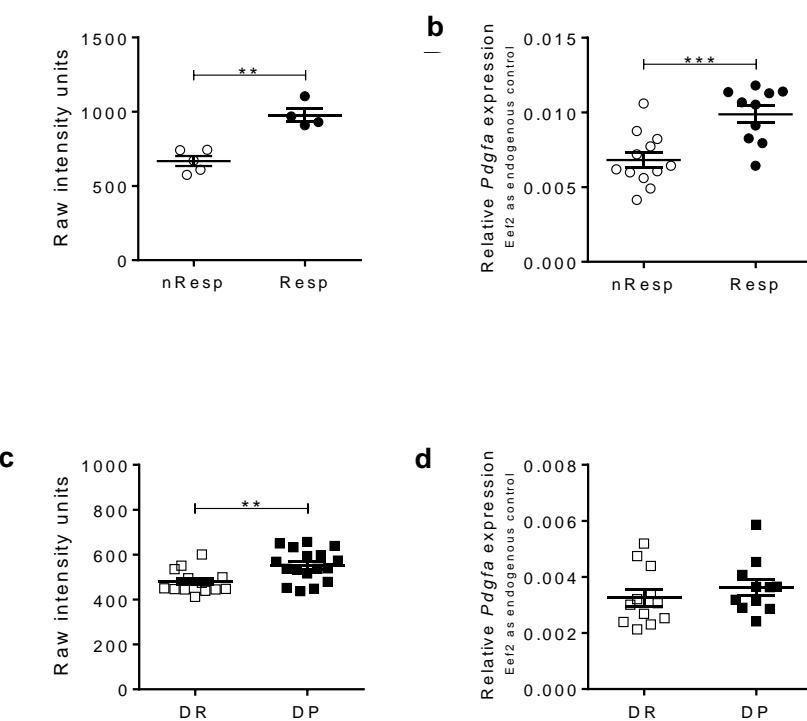
**Table 1.** Association of liver methylation levels of cg14496282 and liver *PDGFA* gene expression with multiple quantitative and binary traits. Methylation levels at cg14496282 and *PDGFA* gene expression are the endogenous variable in all linear regressions used to measure associations. SD: Standard Deviation.

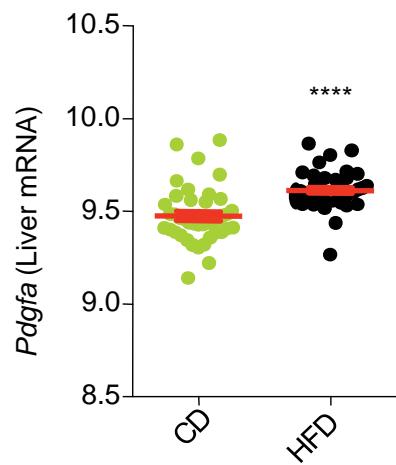
Traits(unit)	<i>PDGFA</i> cg14496282 methylation		<i>PDGFA</i> Expression	
	Effect size in % of methylation / trait unit ( <i>p</i> -value)		Effect size in SD/trait unit ( <i>p</i> -value)	
	Controls	T2D cases	Controls	T2D cases
cg14496282 methylation (%)			-1.44 (6.27×10 <sup>-3</sup> )	-2.49 (4.94×10 <sup>-3</sup> )
<i>PDGFA</i> expression (SD)	-0.05 (6.27×10 <sup>-3</sup> )	-0.03 (4.94×10 <sup>-3</sup> )		
Fasting glucose (mmol/l)	-0.01 (0.79)	-	0.29 (0.196)	-
Fasting insulin (pmol/l)	-1.45×10 <sup>-3</sup> (2.32×10 <sup>-3</sup> )	-	6.83×10 <sup>-3</sup> (9.49×10 <sup>-3</sup> )	-
HOMA2-B (unitless - log)	-0.17 (2.92×10 <sup>-3</sup> )	-	0.63 (0.038)	-
HOMA2-IR (unitless - log)	-0.10 (4.93×10 <sup>-3</sup> )	-	0.53 (7.47×10 <sup>-3</sup> )	-
QUICKI (unitless)	1.66 (0.01)	-	-9.19 (9.78×10 <sup>-3</sup> )	-
Steatosis (%)	-2.15×10 <sup>-3</sup> (0.01)	-4.34×10 <sup>-4</sup> (0.42)	0.01 (2.72×10 <sup>-3</sup> )	0.02 (2.14×10 <sup>-6</sup> )
NASH (Yes/ No)	-0.17 (0.04)	-0.072 (0.03)	2.11 (9.38×10 <sup>-7</sup> )	1.45 (3.37×10 <sup>-8</sup> )
Hepatic fibrosis (Yes/ No)	-0.07 (0.09)	-0.051 (0.04)	0.19 (0.434)	0.63 (2.66×10 <sup>-3</sup> )
Alanine aminotransferase (UI/L)	-1.44×10 <sup>-4</sup> (0.89)	-1.34×10 <sup>-3</sup> (0.03)	0.01 (0.067)	0.02 (1.46×10 <sup>-4</sup> )
Aspartate aminotransferase (UI/L)	-4.71×10 <sup>-3</sup> (0.06)	-1.76×10 <sup>-3</sup> (0.04)	0.03 (7.89×10 <sup>-3</sup> )	0.03 (2.56×10 <sup>-6</sup> )

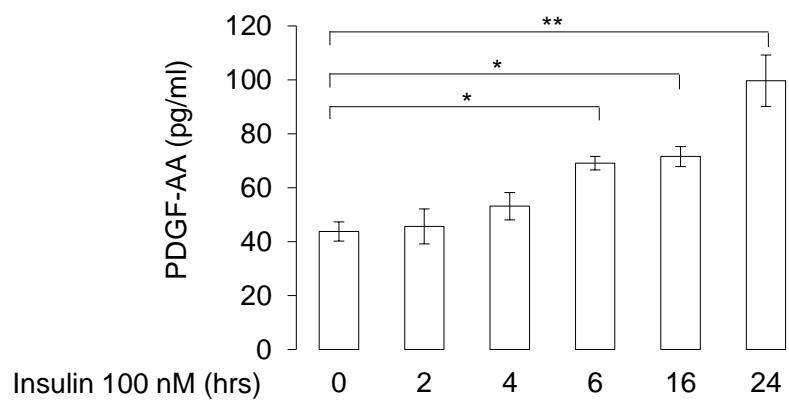
**Fig. 1****a****b**

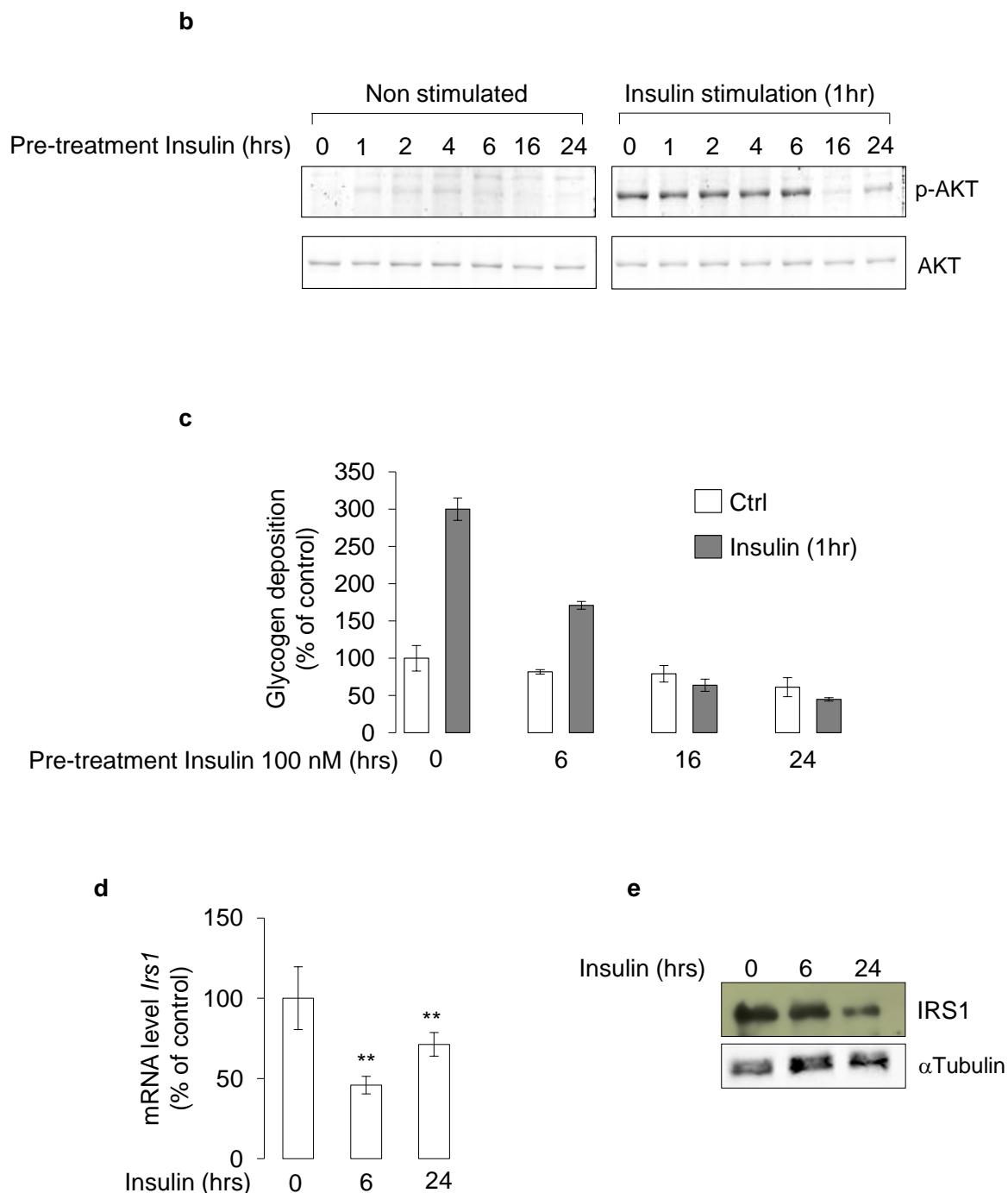
**Fig. 2**

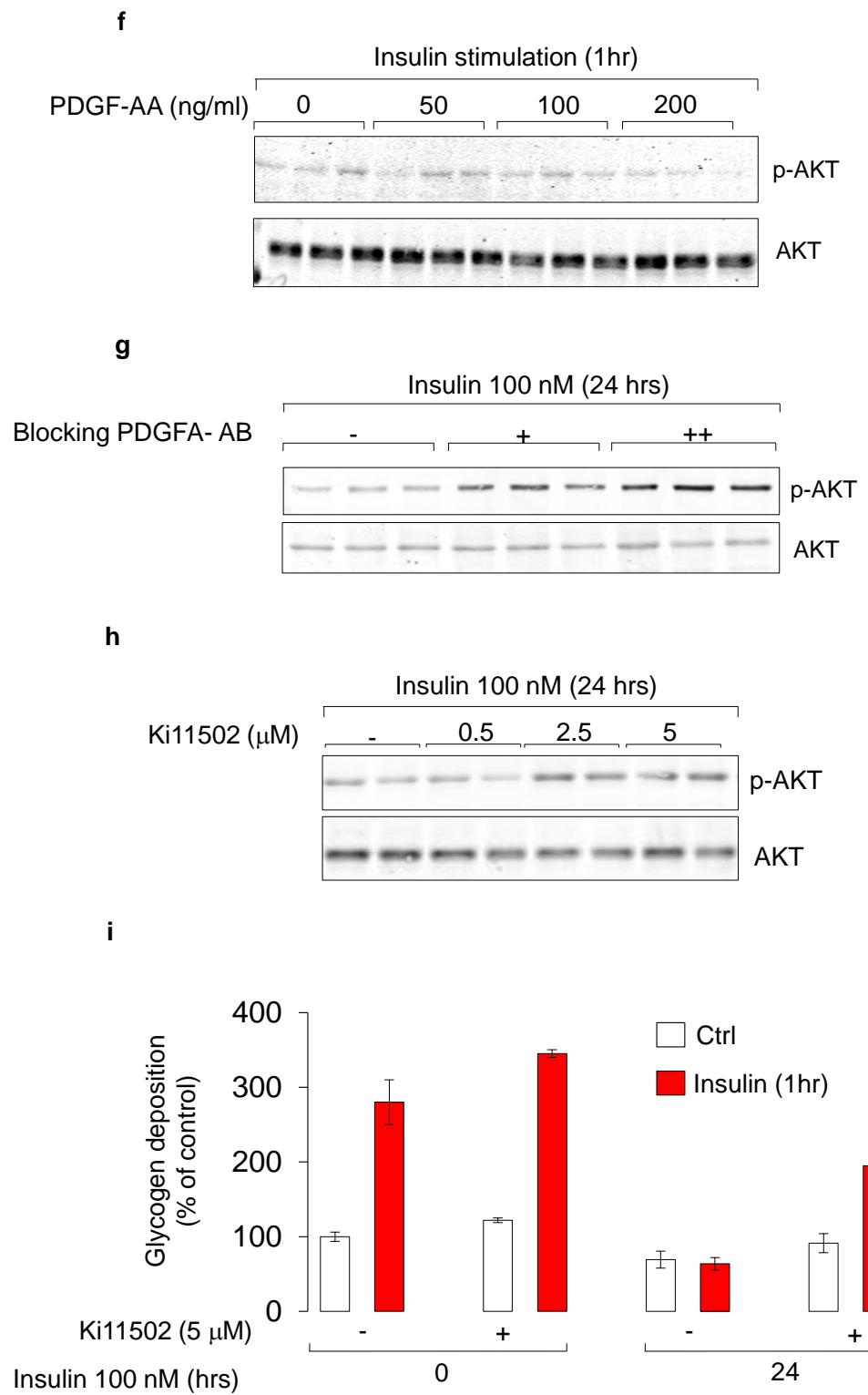
**Fig. 2**

**Fig. 3**

**Fig. 3e**

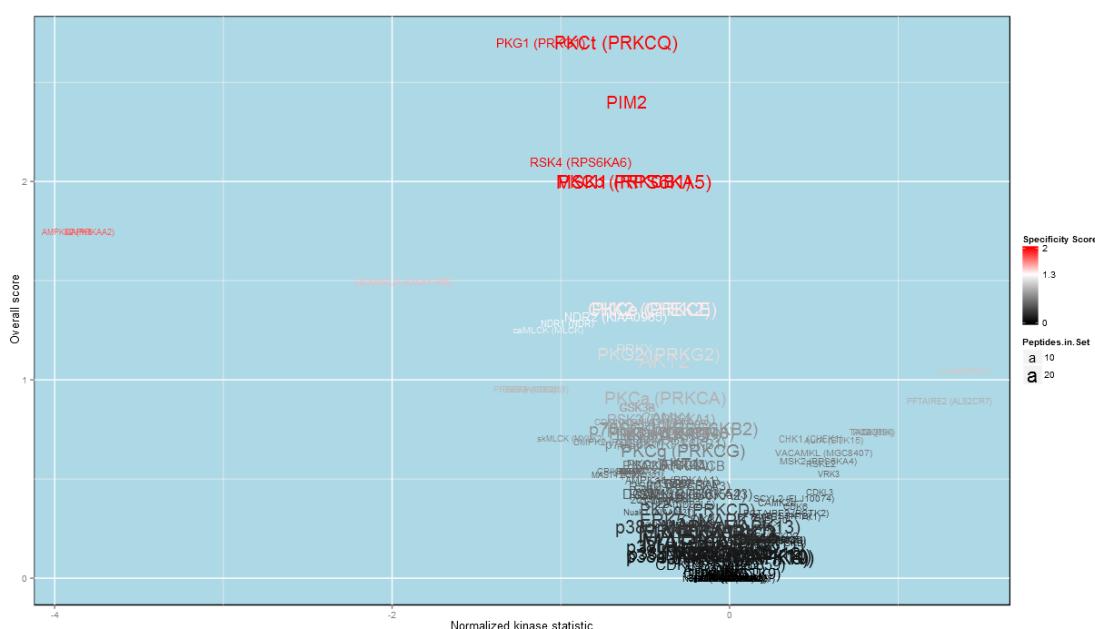
**Fig. 4a**

**Fig. 4**

**Fig. 4**

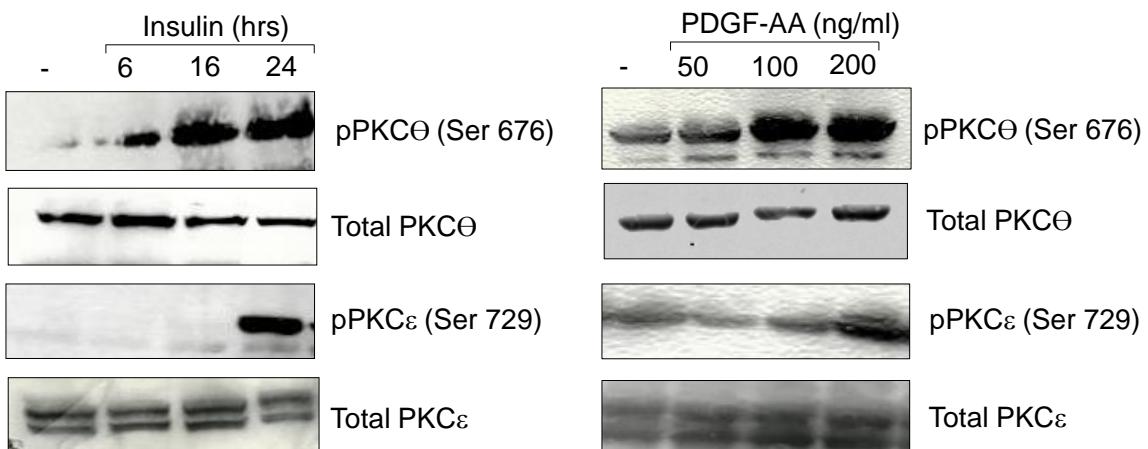
**Fig. 4**

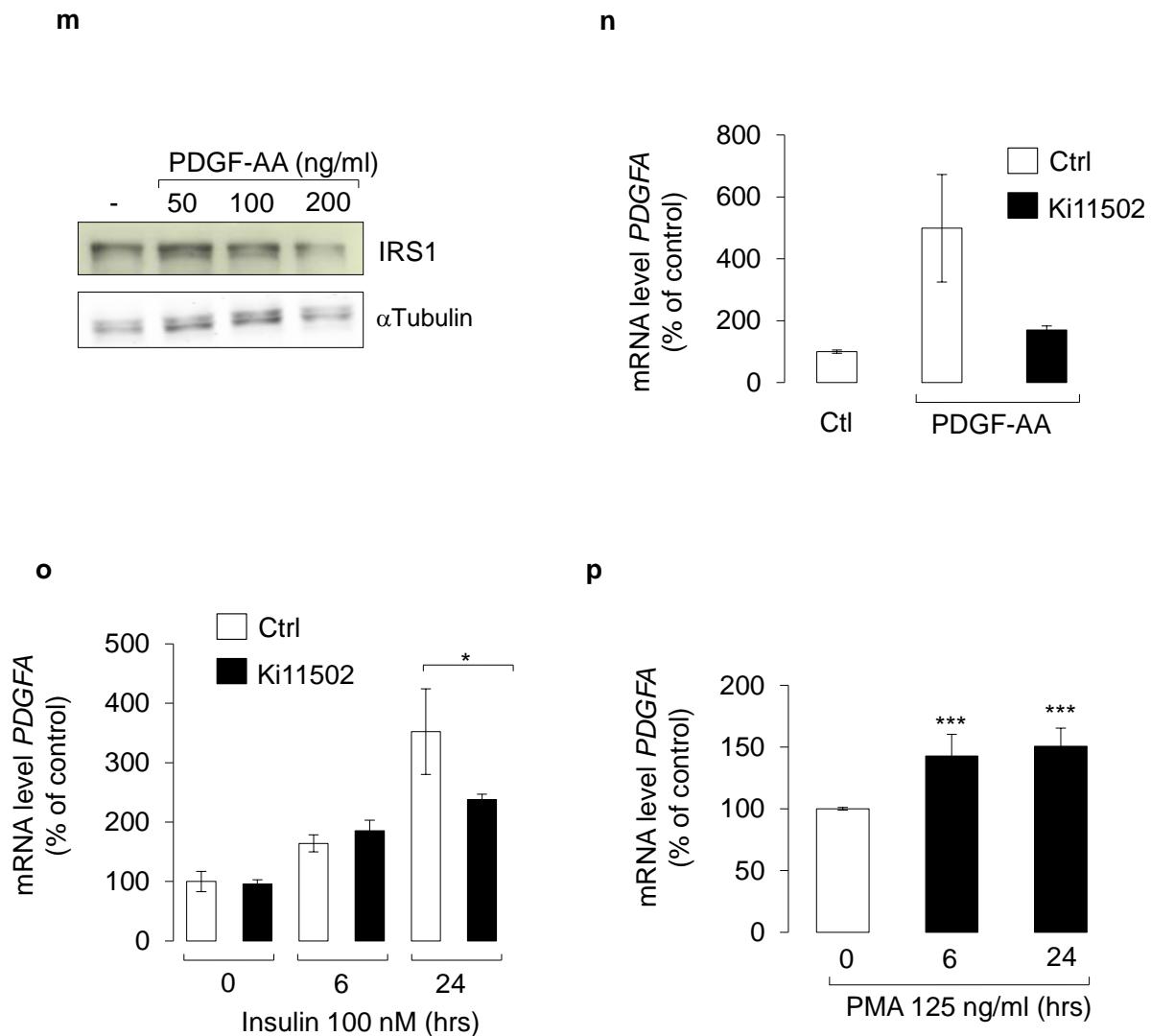
j

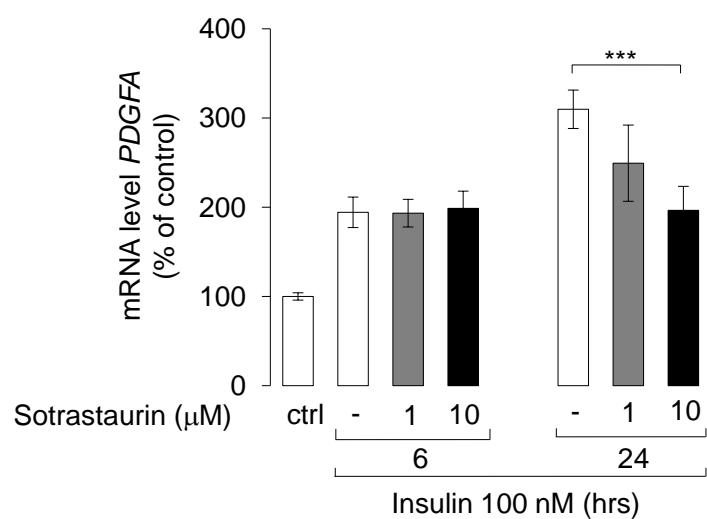
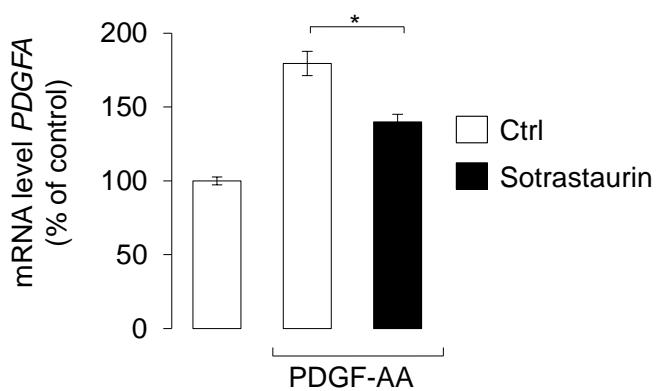


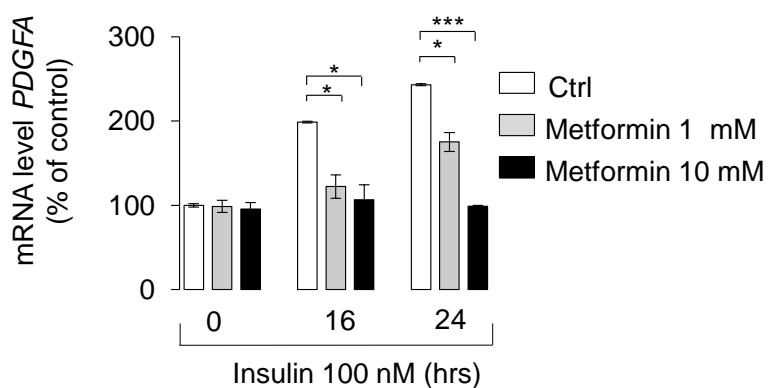
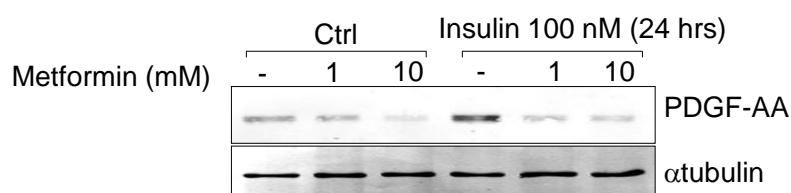
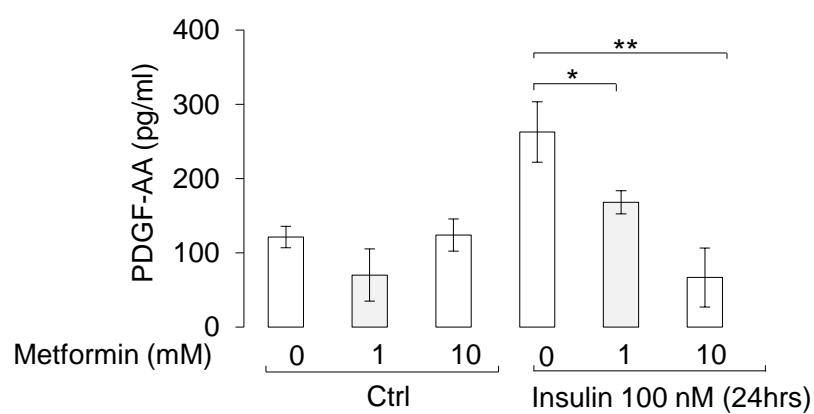
k

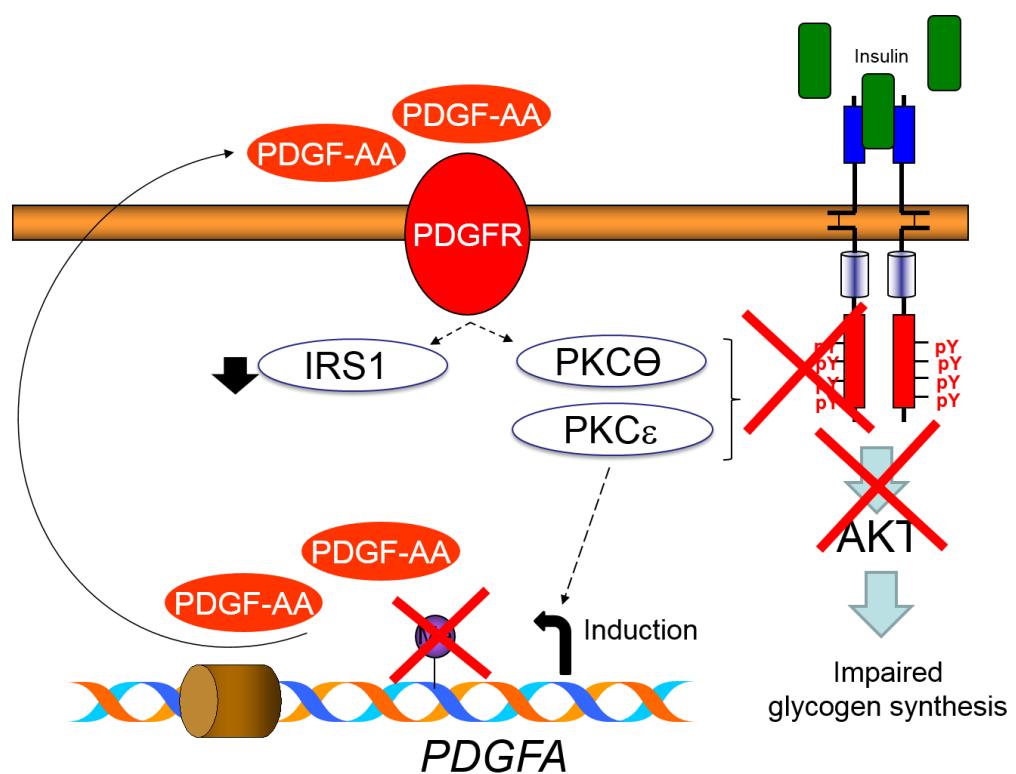
1



**Fig. 4**

**Fig. 4****q****r**

**Fig. 4****s****t****u**

**Fig. 5**



# Chapitre 4

---

## *L'Exposition à Faible Dose aux Bisphénols A, F et S des Adipocytes Primaires Humains Modifie les Profils d'ARN Codant et Non-Codant*

---

Publié dans **PLoS ONE**<sup>1</sup>.

Marie Verbanck<sup>1,\*</sup>, **Mickaël Canouil**<sup>1,\*</sup>, Audrey Leloire<sup>1</sup>, Véronique Dhennin<sup>1</sup>, Xavier Courmoul<sup>2</sup>, Loïc Yengo<sup>1</sup>, Philippe Froguel<sup>1,3,†</sup> & Odile Poulaïn-Godefroy<sup>1,†</sup>

<sup>1</sup>Univ. Lille, CNRS, CHU Lille, Institut Pasteur de Lille, UMR 8199 - EGID, F-59000 Lille, France; <sup>2</sup>INSERM UMR-S 1124, Toxicologie Pharmacologie et Signalisation cellulaire, 75006 Paris, France; Université Paris Descartes, ComUE Sorbonne Paris Cité, 75006 Paris, France; <sup>3</sup>Department of Genomics of Common Disease, School of Public Health, Imperial College London, United Kingdom.

\*Co-premier auteurs.

†Co-dernier auteurs.

---

1. <http://doi.org/10.1371/journal.pone.0179583>

---

## 1 Introduction

### 1.1 Contexte/objectifs

L'exposition des populations au bisphénol A (BPA) a été suspectée de participer à l'épidémie d'obésité et de désordres métaboliques. Le BPA, avant son interdiction en Europe et notamment en France, était utilisé dans la fabrication de plastiques et de résines époxy, et couramment utilisé dans les contenants alimentaires tels que les biberons ou les revêtements de protection des boîtes de conserve. Le BPA se retrouve également dans certains jouets, appareils médicaux et certains papiers comme les tickets de caisse. Depuis l'interdiction de l'usage du BPA, des composés analogues ont vu le jour, tels que le bisphénol S (BPS) et le bisphénol F (BPF), et sont maintenant utilisés de façon courante. À ce jour, les études toxicologiques portant sur les BPS et BPF sont peu nombreuses.

Notre hypothèse est que les substituts du BPA pourraient avoir un effet similaire à celui du BPA, notamment au niveau du tissu adipeux. À cet effet, nous avons comparé le profil d'expression des ARN codants et non-codants d'adipocytes primaires humains exposés à ces différents bisphénols au cours de leur différenciation.

### 1.2 Méthodes

Les adipocytes primaires provenant de trois patientes caucasiennes et non diabétiques ont été cultivés en présence des différents bisphénols (BPA, BPS et BPF) à deux concentrations différentes : 10 nM, correspondant à la concentration de BPA observée dans les fluides corporels de la population générale, et 10 µM, pour mesurer un potentiel effet de concentration. Une condition contrôle, correspondant au tampon utilisé pour diluer les différents bisphénols (DMSO) lors de la culture et de la différenciation des adipocytes primaires, a été utilisée pour fins de comparaison. Après différenciation des adipocytes primaires en adipocytes, l'ARN a été extrait et les profils ARN ont été évalués via une puce Agilent SurePrint G3 Human V2 pour les mRNA et lncRNA, et via une puce Agilent SurePrint G3 miRNA pour les miRNA. L'ARN d'adipocyte primaire (non différencié) a également été extrait pour évaluation du statut de différenciation des cellules. En effet, certains gènes ne sont exprimés que dans les adipocytes différenciés.

Un contrôle qualité des données générées par ces deux plateformes a ensuite été réalisé. Un premier filtre des sondes de mRNA/lncRNA est appliqué pour exclure les sondes n'étant pas exprimées. Les valeurs d'expression ont été considérées comme manquantes (car mal détectées ou non exprimées) lorsque la valeur-p de détection, telle que fournie par le logiciel d'analyse d'Agilent, était non-significative (valeur-p > 0,05). Les sondes

dont le taux de valeurs manquantes était inférieur à 5 % étaient conservées pour analyse (c.-à-d. les sondes exprimées dans au moins 95 % des échantillons). Les données ont ensuite été normalisées via une normalisation quantile implémentée dans les extensions R *limma* et *AgiMicroRna*, pour corriger un “effet plaque” résultant de l’utilisation de plusieurs puces. Afin de limiter l’impact de cet éventuel “effet plaque”, les différentes conditions expérimentales ont été réparties sur les différentes puces selon un plan factoriel. À cette étape de correction de l’effet plaque s’ajoute une normalisation de l’expression des gènes à partir de gènes de ménage (gènes ubiquitaires dont le niveau d’expression est constant dans l’ensemble des tissus), afin de rendre l’expression des gènes d’intérêt comparable, en forçant l’expression des gènes de ménage à être constante (p. ex. égal à 1 dans le cas de l’utilisation du ratio  $\frac{G_{cible}}{G_{ménage}}$ , ou de façon équivalente, égal à 0 après transformation logarithmique). Une analyse en composantes principales a été réalisée sur l’ensemble des sondes passant le contrôle qualité pour les données de mRNA/IncRNA et miRNA dans le but d’identifier une structuration des données associée au statut de différenciation des adipocytes primaires ou aux différents patients.

Pour chacun des trois patients, 4 échantillons contrôles (DMSO : contrôle négatif), 2 échantillons pour chaque combinaison de bisphénols (BPA, BPS, BPF) et de concentrations (10 nM et 10 µM) ont produit globalement  $4 + 2 \times (3 \times 2) = 16$  échantillons par patient. Les sondes mRNA, IncRNA et miRNA différentiellement exprimées ont été identifiées à l’aide d’un modèle linéaire mixte, pour un bisphénol donné, avec en effet fixe la concentration de bisphénol (10 nM et/ou 10 µM) comparée à la condition contrôle (DMSO), et avec effet aléatoire le patient. L’analyse portant sur plus de 22 000 sondes mRNA/IncRNA et 483 sondes miRNA, une correction pour tests multiples a été appliquée pour déterminer la significativité des effets selon la méthode de Benjamini et Hochberg au seuil de 5 %. Pour visualiser efficacement les similarités et dissimilarités des différentes sondes entre les conditions analysées, une représentation en “heatmap” a été utilisée.

### 1.3 Résultats

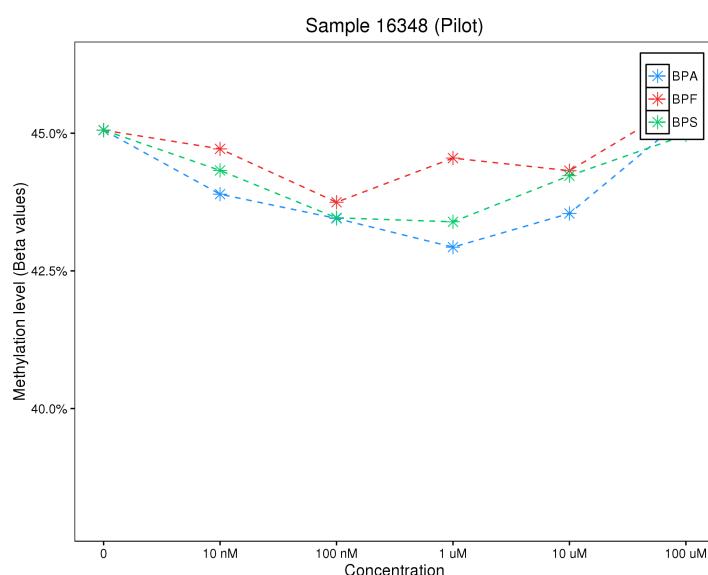
Les analyses ont permis de mettre en évidence un ensemble de 846 sondes mRNA/IncRNA dont l’expression était réduite en présence d’un bisphénol (BPA, BPF et BPS) dans une concentration “physiologique” de 10 nM et de 417 sondes dont l’expression était augmentée par rapport à la condition contrôle (DMSO). Avec une concentration “forte” de 10 µM lors de différenciation des adipocytes primaires, nous avons pu identifier 774 et 1106 sondes présentant, respectivement, une diminution et une augmentation de l’expression en présence de bisphénols. Certaines de ces dérégulations dans l’expression de ces sondes, associées à la présence de bisphénols dans le milieu de culture pendant la différenciation, sont partagées entre les BPA, BPF et BPS, mais aussi entre les deux concentrations (10 nM et 10 µM). Des résultats similaires ont également pu être observés pour l’ARN non codant (miRNA).

L'utilisation de l'outil IPA (*Ingenuity Pathway Analysis*) a permis d'identifier des voies/fonctions biologiques/métaboliques présentant un enrichissement des sondes identifiées dans ces voies. Les dérégulations associées à la présence de bisphénols pourraient être impliquées dans les voies liées au "cancer" et à des "anomalies et blessures de l'organisme". IPA a permis également d'analyser les régulateurs en amont des dix gènes dérégulés communs pour les trois bisphénols et les deux concentrations. Parmi ces régulateurs, la voie des estrogènes a été mise en évidence.

#### 1.4 Conclusion

L'identification de profils transcriptomiques dérégulés similaires entre les bisphénols A, S et F, ainsi que l'identification de gènes dont les éléments de régulation sont d'origine hormonale, suggèrent un potentiel caractère de perturbateur endocrinien pour les BPF et BPS, au même titre que le BPA. Aussi, les résultats de notre étude suggèrent, qu'en raison des fortes similarités entre les substituts du BPA, que sont les BPF et BPS, que ceux-ci devraient être soumis aux mêmes restrictions.

#### 1.5 Note



**FIGURE 32.** Profil de methylation globale selon différentes concentrations de bisphénols A, F et S.

Dans une étude pilote portant sur les cellules d'un patient, la méthylation globale d'environ 381 000 sites CpG (puce Illumina HumanMethylation 450K) a été analysée en fonction de la concentration en bisphénol (10 nM, 100 nM, 1 µM, 10 µM et 100 µM), introduite dans le milieu de culture lors de la différenciation des adipocytes primaires. Ces premiers résultats suggéraient un effet dose-réponse de la méthylation, pouvant être modélisé par un polynôme de degré deux (Figure 32), dont les effets les plus importants ont été observés

au niveau du gène PTPRN2 (Protein Tyrosine Phosphatase, Receptor type N2) avec 91 sites CpG présentant un effet d'ordre deux significatif du logarithme de la concentration de bisphénol. Ces sites présentaient une hypométhylation pour des concentrations non physiologiques (c.-à-d. supérieures à 10 nM), et un retour proche du niveau de méthylation pour une forte concentration de 100 µM, suggérant ainsi que la méthylation des adipocytes pouvait être sensible à l'ajout de bisphénols dans le milieu, notamment pour des concentrations de 10 nM et de 10 µM, tel qu'observé sur la méthylation globale.

Conjointement à l'étude du transcriptome des adipocytes, le méthylome de ces mêmes cellules a été examiné. Cependant, la forte variabilité inter et intra-patients n'a pas permis, en plus du faible nombre de patients ( $n = 3$  en duplicitats) d'identifier des sites CpG différemment méthylés partagés entre les différents bisphénols. De plus, contrairement à l'analyse transcriptomique où le statut de différenciation des adipocytes a pu être évalué (via l'utilisation de gènes spécifiquement exprimés dans les adipocytes matures), le méthylome des adipocytes comparés aux adipocytes primaires n'a pas permis d'identifier de marques de méthylation spécifiques reflétant cet état de différenciation. Les mRNA et les miRNA identifiés dans l'analyse transcriptomique ont fait l'objet d'une étude de corrélation avec les sites CpG annotés (Illumina) sur les gènes correspondants. Aucune corrélation significative entre les sondes miRNA/mRNA et la méthylation n'a pu être mise en évidence, vraisemblablement en raison d'un manque de puissance. Le faible nombre d'échantillons combiné à la variabilité de la méthylation observée dans ces échantillons a probablement altéré notre capacité à détecter des sites CpG impactés de façon similaire par les bisphénols, malgré des études ayant démontré un effet sur la méthylation du BPA, comme la déméthylation globale de l'ADN au cours de la différenciation d'adipocytes (lignée cellulaire 3T3-L1) [Bastos Sales et al., 2013].

---

## 2 Article

Article disponible en ligne sur **PLoS ONE** (<http://doi.org/10.1371/journal.pone.0179583>)



---

## **Conclusion**

---

Nous proposons dans cette thèse une contribution au domaine de la statistique génétique appliquée à la pathologie du diabète de type 2. Cette contribution se scinde en deux aspects. Le premier aspect est un développement méthodologique visant à améliorer les connaissances actuelles et exploiter au mieux les données disponibles. Le second aspect se concentre sur le support méthodologique et l'application des méthodes adaptées aux questionnements inhérents aux différents projets de recherche sur les données “omiques”.

---

### **1 Développement méthodologique**

Deux principaux développements ont été réalisés au cours de cette thèse (Chapitres 1 et 2) :

- Le premier est une nouvelle approche, dans le domaine de la génétique, permettant de modéliser conjointement l'évolution d'un trait longitudinal et la survenue d'un événement, tout en évaluant l'effet des SNPs simultanément sur ces deux traits. Contrairement aux approches dites “classiques” des GWAS qui ont étudié l'effet des SNPs, d'une part, sur la glycémie à jeun dans des groupes d'individus normoglycémiques, et d'autre part, l'association des SNPs dans des études cas/témoins, l'approche par modèle joint que nous proposons dans cette thèse permet d'identifier des SNPs associés à l'évolution de la glycémie sans qu'ils ne soient nécessairement associés au risque de développement d'un diabète. De plus, nous apportons une solution au problème computationnel de l'application de cette approche à l'échelle du génome (p. ex. puce-à-ADN imputée, séquençage, etc.), au moyen d'une méthode approchée dite en “deux étapes”. Nous avons montré que l'approche en “deux étapes” est aussi robuste et précise que l'approche par modèle joint tout en offrant un gain en terme de temps de calcul.
- Le second développement réalisé porte sur l'amélioration et l'automatisation de la récupération et l'analyse des données provenant de la technologie NanoString et de culture cellulaire. Dans ce second développement, j'ai réalisé deux applications web Shiny, nommées *NanoStringTissueCartography* et *EndoC\_Beta* :
  - L'application *NanoStringTissueCartography* permet, en premier lieu, à partir des données brutes gé-

nérées via la technologie NanoString, d'effectuer et de visualiser les différentes étapes du contrôle qualité, ainsi que les niveaux d'expressions de plusieurs gènes de ménage, et en second lieu, d'analyser et de visualiser les données importées dans l'application, tout en offrant un cadre interactif. Par exemple, il est possible de sélectionner un ou plusieurs gènes selon une fonction, mais également de sélectionner les tissus d'intérêts, pour lesquels l'analyse d'expression et l'étude enrichissement doivent être réalisées.

- L'application *EndoC\_Beta* permet d'inclure des fichiers de résultats au format Excel (ici, des mesures d'absorbance réalisées sur un modèle de cellule  $\beta$ ) dès le remplissage de ces fichiers par les personnes en charge des expérimentations. Lagrégation de ces résultats dans un format standardisé permet de mesurer et de prendre en compte plusieurs facteurs techniques tels que l'expérimentateur, le jour d'expérimentation et la qualité de l'étalonnage des appareils, et de ce fait permet de réduire les biais que ces facteurs peuvent entraîner dans l'analyse. En outre, cette application permet différents niveaux de contrôle qualité. Par exemple, j'ai incorporé une étape de contrôle de la gamme étalon utilisée pour estimer la sécrétion d'insuline. Cette étape vise à vérifier que les mesures d'absorbances observées, pour chaque concentration de référence, restent homogènes d'une expérience à l'autre, puisque le matériel (spectromètre et produits de référence de la gamme) est théoriquement le même. De plus, une erreur d'estimation de la pente de la gamme étalon (relation linéaire entre la concentration et l'absorbance) impactera l'ensemble des mesures utilisant celle-ci. Cette application, en plus de prendre en compte les facteurs techniques dans l'analyse, fournit un critère objectif quant à la mesure de sécrétion d'insuline par les cellules  $\beta$ . En effet, le regroupement de l'ensemble des expérimentations réalisées sur les cellules  $\beta$  a permis d'établir un seuil à partir du ratio de la quantité d'insuline sécrétée dans la condition cible et dans la condition témoin. Ce seuil fournit une indication relative au développement des cellules  $\beta$  et donc indirectement sur la qualité de la culture.

---

## 2 Support méthodologique

La nature et la complexité des données "omiques" nécessitent de pouvoir identifier et appliquer une grande variété de méthodes connues et/ou nouvelles pour s'adapter au mieux à la problématique des projets de recherche. Il existe de nombreux articles traitant des méthodes d'analyses des données de génomique, de transcriptomique et de méthylomique. Cependant, le domaine de la méthylomique reste quant à lui en retard par rapport aux deux autres, principalement en raison de son caractère récent et en développement.

Les Chapitres 3 et 4, sont le résultat des études méthodologiques sur les trois “omiques” discutées dans cette thèse, d'une part, au niveau des méthodes de contrôle qualité plus ou moins spécifiques, et d'autre part, au niveau des méthodes d'analyse statistique. Dans ces deux chapitres, les études décrites ont fait l'objet d'une étude pilote, en particulier pour étudier la méthylation de l'ADN. Ces études pilotes ont permis l'identification et le développement d'un processus d'analyse allant de la lecture des données à la génération des résultats d'analyse, sur lequel se sont appuyées les études définitives et les demandes de financement subséquentes. En effet, ces résultats ont permis d'établir le nombre d'échantillon nécessaire et/ou d'évaluer la puissance de l'étude, et d'élaborer les plans d'expériences les mieux adaptés afin de réduire les biais techniques ou les facteurs de confusion.

---

### 3 Multi-omiques & Perspectives

Les différents projets abordés dans les Chapitres 1, 2, 3 et 4 s'inscrivent dans une démarche multi-omiques, notamment par l'intégration de la méthylomique, de la transcriptomique, de la génomique et de la phénotypique (phénotype). Néanmoins, cette intégration est partielle dans le sens où l'analyse est réalisée sous la forme de filtres successifs : par exemple dans le Chapitre 3, l'analyse s'est concentrée sur la méthylomique, puis les autres “omiques” ont été incluses pour apporter une nouvelle couche “mécanistique” et ainsi remonter à la pathophysiologie. Cette approche d'intégration successive des “omiques” offrent un cadre de mise en œuvre simple et robuste, et permet de mettre en évidence un aspect mécanistique plus proche des hypothèses biologiques et cliniques, que ne le permet une analyse séparée des “omiques”.

Avec le développement de la médecine de précision, qui vise à caractériser, à plusieurs niveaux, un individu et prédire son avenir médical, des méthodes d'intégrations de données se sont développées conjointement aux méthodes relatives au “Big Data” (c.-à-d. “machine learning”, tels que les “neural network”) permettant de traiter des gros volumes de données [Huang et al., 2017; Lin and Lane, 2017]. La cohorte D.E.S.I.R. fournit un cadre idéal d'application et de développement des approches pour données longitudinales et des approches d'intégration de données. En effet, dans cette cohorte interrogée lors de quatre vagues de mesure conduites à intervalle régulier sur une période 9 ans, des données de différentes natures sont disponibles telles que plus de 200 variables phénotypiques, des données de génomique, de métabolomique (deux temps de mesure) et prochainement des données de méthylomique (deux temps de mesure). À cela s'ajoute la mise en place de consortia permettant de regrouper plusieurs laboratoires et cohortes (p. ex. RHAPSODY portant sur l'évaluation

du risque et de la progression du diabète de type 2 et du pré-diabète), établissant un cadre de plus en plus adapté aux différentes méthodes décrites et développées dans cette thèse.

---

## **Liste des communications scientifiques**

---

### **1 Communications en lien avec la thèse**

#### **1.1 Conférences**

- Rocheleau, G., Canouil, M., Yengo, L., Froguel, P. (2015). Application of Joint Models in Genetic Association Studies. *International Genetic Epidemiology Society - IGES*, Baltimore, United-States.
- Canouil, M., Rocheleau, G., Yengo, L., Froguel, P. (2016). Longitudinal Genetic Modelling : Revisiting Associations of SNPs Associated with Blood Fasting Glucose in Normoglycemic Individuals. *Statistical Methods for Post Genomic Data - SMPGD*, Lille, France.
- Canouil, M., Froguel, P., Rocheleau, G. (2016). Single Nucleotide Polymorphisms Associated with Fasting Blood Glucose Trajectory and Type 2 Diabetes Incidence : A Joint Modelling Approach. *International Genetic Epidemiology Society - IGES*, Toronto, Canada.
- Canouil, M., Froguel, P., Rocheleau, G. (2016). Single Nucleotide Polymorphisms Associated with Fasting Blood Glucose Trajectory and Type 2 Diabetes Incidence : A Joint Modelling Approach. *4th Symposium European Genomic Institute for Diabetes (E.g.i.d)*, Lille, France.
- Canouil, M., Froguel, P., Rocheleau, G. (2017). Variants Génétiques Associés à la Trajectoire de la Glycémie à Jeun et à l'Incidence du Diabète de Type 2 : Une Approche par Modèle Joint. *Annual Congress of Société Francophone du Diabète (SFD)*, Lille, France.

## 1.2 Articles publiés dans des revues internationales à comité de lecture

- Ndiaye, F. K., Ortalli, A., **Canouil, M.** et al. (2017). Expression and functional assessment of candidate type 2 diabetes susceptibility genes identify four new genes contributing to human insulin secretion. *Molecular Metabolism*.
  - Verbanck, M., **Canouil, M.** et al. (2017). Low-dose exposure to bisphenols A, F and S of human primary adipocyte impacts coding and non-coding RNA profiles. *PLOS ONE*, 12(6), e0179583.
- 

## 2 Autres communications dans le domaine de la génétique

- Yengo, L., Jacques, J., Biernacki, C. and **Canouil, M.** (2016). Variable Clustering in High-Dimensional Linear Regression : The R Package clere. *The R Journal*, 8(1), 92–106.
- Baumeier, C., Saussenthaler, S., Kammel, A., Jähnert, M., Schlüter, L., Hesse, D., **Canouil, M.** et al. (2017). Hepatic DPP4 DNA Methylation Associates With Fatty Liver. *Diabetes*, 66(1), 25–35.
- Carrat, G. R., Hu, M., Nguyen-Tu, M.-S., Chabosseau, P., Gaulton, K. J., van de Bunt, M., Siddiq, A., Falchi, M., Thurner, M., **Canouil, M.** et al. (2017). Decreased STARD10 Expression Is Associated with Defective Insulin Secretion in Humans and Mice. *The American Journal of Human Genetics*, 100(2), 238–256.
- Bonnefond, A., Yengo, L., Dechaume, A., **Canouil, M.** et al. (2017). Relationship between salivary/pancreatic amylase and body mass index : a systems biology approach. *BMC Medicine*, 15(1).

---

## Bibliographie

---

- Abecasis, G. R., Cherny, S. S., and Cardon, L. R. (2001). The impact of genotyping error on family-based analysis of quantitative traits. *European Journal of Human Genetics*, 9(2) :130–134.
- Almasy, L., Dyer, T. D., Peralta, J. M., Jun, G., Wood, A. R., Fuchsberger, C., Almeida, M. A., Kent, J. W., Fowler, S., et al. (2014). Data for Genetic Analysis Workshop 18 : Human whole genome sequence, blood pressure, and simulated phenotypes in extended pedigrees. *BMC Proceedings*, 8(Suppl 1) :S2.
- Ambros, V. (2004). The functions of animal microRNAs. *Nature*, 431(7006) :350–355.
- Antequera, F. (2003). Structure, function and evolution of CpG island promoters. *Cellular and molecular life sciences : CMLS*, 60(8) :1647–1658.
- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., et al. (2000). Gene Ontology : Tool for the unification of biology. *Nature Genetics*, 25(1) :25–29.
- Ball, M. P., Li, J. B., Gao, Y., Lee, J.-H., LeProust, E. M., Park, I.-H., Xie, B., Daley, G. Q., and Church, G. M. (2009). Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nature Biotechnology*, 27(4) :361–368.
- Bartel, D. P. (2004). MicroRNAs : Genomics, Biogenesis, Mechanism, and Function. *Cell*, 116(2) :281–297.
- Bastos Sales, L., Kamstra, J., Cenijn, P., van Rijt, L., Hamers, T., and Legler, J. (2013). Effects of endocrine disrupting chemicals on in vitro global DNA methylation and adipocyte differentiation. *Toxicology in Vitro*, 27(6) :1634–1643.
- Beer, N. L., Tribble, N. D., McCulloch, L. J., Roos, C., Johnson, P. R. V., Orho-Melander, M., and Gloyn, A. L. (2009). The P446L variant in GCKR associated with fasting plasma glucose and triglyceride levels exerts its effect through increased glucokinase activity in liver. *Human Molecular Genetics*, 18(21) :4081–4088.
- Bell, C. G., Finer, S., Lindgren, C. M., Wilson, G. A., Rakyan, V. K., Teschendorff, A. E., Akan, P., Stupka, E., Down, T. A., et al. (2010). Integrated Genetic and Epigenetic Analysis Identifies Haplotype-Specific Methylation in the FTO Type 2 Diabetes and Obesity Susceptibility Locus. *PLOS ONE*, 5(11) :e14040.

- Beyene, J. and Hamid, J. S. (2014). Longitudinal Data Analysis in Genome-Wide Association Studies. *Genetic Epidemiology*, 38(S1) :S68–S73.
- Bi, Y., Wang, W., Xu, M., Wang, T., Lu, J., Xu, Y., Dai, M., Chen, Y., Zhang, D., et al. (2015). Diabetes genetic risk score modifies effect of bisphenol A exposure on deterioration in glucose metabolism. *The Journal of Clinical Endocrinology & Metabolism*, pages jc.2015–3039.
- Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J. M., Delano, D., Zhang, L., Schroth, G. P., et al. (2011). High density DNA methylation array with single CpG site resolution. *Genomics*, 98(4) :288–295.
- Bijanzadeh, M. (2017). The recurrence risk of genetic complex diseases. *Journal of Research in Medical Sciences*, 22(1) :32.
- Bird, A. P. (1980). DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Research*, 8(7) :1499–1504.
- Bouatia-Naji, N., Bonnefond, A., Cavalcanti-Proen  a, C., Spars  , T., Holmkvist, J., Marchand, M., Delplanque, J., Lobbens, S., Rocheleau, G., et al. (2009). A variant near MTNR1B is associated with increased fasting plasma glucose levels and type 2 diabetes risk. *Nature Genetics*, 41(1) :89–94.
- Bouaziz, M., Ambroise, C., and Guedj, M. (2011). Accounting for Population Stratification in Practice : A Comparison of the Main Strategies Dedicated to Genome-Wide Association Studies. *PLoS ONE*, 6(12) :e28845.
- Burton, P. R., Clayton, D. G., Cardon, L. R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D. P., McCarthy, M. I., Ouwehand, W. H., et al. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145) :661–678.
- Canivell, S., Ruano, E. G., Sis  -Almirall, A., Kostov, B., Gonz  lez-de Paz, L., Fernandez-Rebollo, E., Hanzu, F. A., P  rrizas, M., Novials, A., et al. (2014). Differential methylation of TCF7L2 promoter in peripheral blood DNA in newly diagnosed, drug-na  ve patients with type 2 diabetes. *PloS One*, 9(6) :e99310.
- Cardenas, A., Allard, C., Doyon, M., Houseman, E. A., Bakulski, K. M., Perron, P., Bouchard, L., and Hivert, M.-F. (2016). Validation of a DNA methylation reference panel for the estimation of nucleated cells types in cord blood. *Epigenetics*, 11(11) :773–779.
- Cardon, L. R. and Palmer, L. J. (2003). Population stratification and spurious allelic association. *The Lancet*, 361(9357) :598–604.
- Case, J., Willoughby, D., Haley-Zitlin, V., and Maybee, P. (2006). Preventing type 2 diabetes after gestational diabetes. *The Diabetes Educator*, 32(6) :877–886.

- Caussinus, H. (1986). Models and uses of principal component analysis. *Multidimensional data analysis*, 86 :149–170.
- Chambers, J. C., Loh, M., Lehne, B., Drong, A., Kriebel, J., Motta, V., Wahl, S., Elliott, H. R., Rota, F., et al. (2015). Epigenome-wide association of DNA methylation markers in peripheral blood from Indian Asians and Europeans with incident type 2 diabetes : A nested case-control study. *The Lancet. Diabetes & Endocrinology*, 3(7) :526–534.
- Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK : Rising to the challenge of larger and richer datasets. *GigaScience*, 4 :7.
- Chang, S.-W., McDonough, C. W., Gong, Y., Johnson, T. A., Tsunoda, T., Gamazon, E. R., Perera, M. A., Takahashi, A., Tanaka, T., et al. (2016). Genome-wide association study identifies pharmacogenomic loci linked with specific antihypertensive drug treatment and new-onset diabetes. *The Pharmacogenomics Journal*.
- Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2017). *shiny : Web Application Framework for R*. R package version 1.0.3.
- Chen, Y.-a., Lemire, M., Choufani, S., Butcher, D. T., Grafodatskaya, D., Zanke, B. W., Gallinger, S., Hudson, T. J., and Weksberg, R. (2013). Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics*, 8(2) :203–209.
- Cheung, V. G. and Spielman, R. S. (2009). Genetics of human gene expression : Mapping DNA variants that influence gene expression. *Nature Reviews Genetics*, 10(9) :595–604.
- Clark, A. G. and Li, J. (2007). Conjuring SNPs to detect associations. *Nature Genetics*, 39(7) :815–816.
- Clayton, D. G., Walker, N. M., Smyth, D. J., Pask, R., Cooper, J. D., Maier, L. M., Smink, L. J., Lam, A. C., Ovington, N. R., et al. (2005). Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nature Genetics*, 37(11) :1243–1246.
- Cooper, D. N. and Krawczak, M. (1989). Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. *Human Genetics*, 83(2) :181–188.
- Costanza, M. C., Beer-Borst, S., James, R. W., Gaspoz, J.-M., and Morabia, A. (2012). Consistency between cross-sectional and longitudinal SNP : Blood lipid associations. *European Journal of Epidemiology*, 27(2) :131–138.
- Dabelea, D., Hanson, R. L., Lindsay, R. S., Pettitt, D. J., Imperatore, G., Gabir, M. M., Roumain, J., Bennett, P. H., and Knowler, W. C. (2000). Intrauterine exposure to diabetes conveys risks for type 2 diabetes and obesity : A study of discordant sibships. *Diabetes*, 49(12) :2208–2211.

- Dayeh, T., Tuomi, T., Almgren, P., Perfilyev, A., Jansson, P.-A., de Mello, V. D., Pihlajamäki, J., Vaag, A., Groop, L., et al. (2016). DNA methylation of loci within ABCG1 and PHOSPHO1 in blood DNA is associated with future type 2 diabetes risk. *Epigenetics*, 11(7) :482–488.
- Dayeh, T., Volkov, P., Salö, S., Hall, E., Nilsson, E., Olsson, A. H., Kirkpatrick, C. L., Wollheim, C. B., Eliasson, L., et al. (2014). Genome-wide DNA methylation analysis of human pancreatic islets from type 2 diabetic and non-diabetic donors identifies candidate genes that influence insulin secretion. *PLoS genetics*, 10(3) :e1004160.
- Dayeh, T. A., Olsson, A. H., Volkov, P., Almgren, P., Rönn, T., and Ling, C. (2013). Identification of CpG-SNPs associated with type 2 diabetes and differential DNA methylation in human pancreatic islets. *Diabetologia*, 56(5) :1036–1046.
- Deaton, A. M. and Bird, A. (2011). CpG islands and the regulation of transcription. *Genes & Development*, 25(10) :1010–1022.
- Dedeurwaerder, S., Defrance, M., Calonne, E., Denis, H., Sotiriou, C., and Fuks, F. (2011). Evaluation of the Infinium Methylation 450K technology. *Epigenomics*, 3(6) :771–784.
- Devlin, B. and Roeder, K. (1999). Genomic Control for Association Studies. *Biometrics*, 55(4) :997–1004.
- Diamond, J. (2003). The double puzzle of diabetes. *Nature*, 423(6940) :599–602.
- Du, P., Zhang, X., Huang, C.-C., Jafari, N., Kibbe, W. A., Hou, L., and Lin, S. M. (2010). Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, 11(1) :587.
- Dudbridge, F. and Gusnanto, A. (2008). Estimation of significance thresholds for genomewide association scans. *Genetic Epidemiology*, 32(3) :227–234.
- Dunn, O. J. (2012). Multiple Comparisons among Means. *Journal of the American Statistical Association*.
- Dupuis, J., Langenberg, C., Prokopenko, I., Saxena, R., Soranzo, N., Jackson, A. U., Wheeler, E., Glazer, N. L., Bouatia-Naji, N., et al. (2010). New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nature Genetics*, 42(2) :105–116.
- Elliott, H. R., Shihab, H. A., Lockett, G. A., Holloway, J. W., McRae, A. F., Smith, G. D., Ring, S. M., Gaunt, T. R., and Relton, C. L. (2017). The Role of DNA Methylation in Type 2 Diabetes Aetiology – Using Genotype as a Causal Anchor. *Diabetes*, page db160874.
- Evangelou, E. and Ioannidis, J. P. A. (2013). Meta-analysis methods for genome-wide association studies and beyond. *Nature Reviews Genetics*, 14(6) :379–389.

- Fardo, D. W., Ionita-Laza, I., and Lange, C. (2009). On Quality Control Measures in Genome-Wide Association Studies : A Test to Assess the Genotyping Quality of Individual Probands in Family-Based Association Studies and an Application to the HapMap Data. *PLoS Genetics*, 5(7) :e1000572.
- Flannick, J. and Florez, J. C. (2016). Type 2 diabetes : Genetic data sharing to advance complex disease research. *Nature Reviews Genetics*, advance online publication.
- Gardiner-Garden, M. and Frommer, M. (1987). CpG islands in vertebrate genomes. *Journal of Molecular Biology*, 196(2) :261–282.
- Georgiopoulos, G. and Evangelou, E. (2016). Power considerations for  $\lambda$  inflation factor in meta-analyses of genome-wide association studies. *Genetics Research*, 98.
- Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., Ch'ang, L.-Y., Huang, W., Liu, B., et al. (2003). The International HapMap Project. *Nature*, 426(6968) :789–796.
- Gibson, G. and Weir, B. (2005). The quantitative genetics of transcription. *Trends in Genetics*, 21(11) :616–623.
- Gordon, D. and Ott, J. (2001). Assessment and management of single nucleotide polymorphism genotype errors in genetic association analysis. In *Pac Symp Biocomput*, volume 6, pages 18–29.
- Graham, J. W. (2009). Missing Data Analysis : Making It Work in the Real World. *Annual Review of Psychology*, 60(1) :549–576.
- Grant, S. F. A., Thorleifsson, G., Reynisdottir, I., Benediktsson, R., Manolescu, A., Sainz, J., Helgason, A., Stefansson, H., Emilsson, V., et al. (2006). Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nature Genetics*, 38(3) :320–323.
- Groves, C. J., Zeggini, E., Minton, J., Frayling, T. M., Weedon, M. N., Rayner, N. W., Hitman, G. A., Walker, M., Wiltshire, S., et al. (2006). Association analysis of 6,736 U.K. subjects provides replication and confirms TCF7L2 as a type 2 diabetes susceptibility gene with a substantial effect on individual risk. *Diabetes*, 55(9) :2640–2644.
- Hales, C. N. and Barker, D. J. (1992). Type 2 (non-insulin-dependent) diabetes mellitus : The thrifty phenotype hypothesis. *Diabetologia*, 35(7) :595–601.
- Hall, E., Volkov, P., Dayeh, T., Bacos, K., Rönn, T., Nitert, M. D., and Ling, C. (2014). Effects of palmitate on genome-wide mRNA expression and DNA methylation patterns in human pancreatic islets. *BMC medicine*, 12 :103.
- Hansen, K. D., Langmead, B., and Irizarry, R. A. (2012). BSsmooth : From whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biology*, 13(10) :R83.

- Hansen, K. D., Timp, W., Bravo, H. C., Sabuncyan, S., Langmead, B., McDonald, O. G., Wen, B., Wu, H., Liu, Y., et al. (2011). Increased methylation variation in epigenetic domains across cancer types. *Nature Genetics*, 43(8) :768–775.
- Hellman, A. and Chess, A. (2007). Gene body-specific methylation on the active X chromosome. *Science (New York, N.Y.)*, 315(5815) :1141–1143.
- Hochberg, Y. and Benjamini, Y. (1990). More powerful procedures for multiple significance testing. *Statistics in Medicine*, 9(7) :811–818.
- Hossain, A. and Beyene, J. (2014). Analysis of baseline, average, and longitudinally measured blood pressure data using linear mixed models. *BMC Proceedings*, 8(Suppl 1) :S80.
- Houseman, E., Accomando, W. P., Koestler, D. C., Christensen, B. C., Marsit, C. J., Nelson, H. H., Wiencke, J. K., and Kelsey, K. T. (2012). DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*, 13(1) :86.
- Houseman, E. A., Kelsey, K. T., Wiencke, J. K., and Marsit, C. J. (2015). Cell-composition effects in the analysis of DNA methylation array data : A mathematical perspective. *BMC Bioinformatics*, 16(1).
- Houseman, E. A., Kile, M. L., Christiani, D. C., Ince, T. A., Kelsey, K. T., and Marsit, C. J. (2016). Reference-free deconvolution of DNA methylation data and mediation by cell composition effects. *BMC Bioinformatics*, 17(1).
- Houseman, E. A., Molitor, J., and Marsit, C. J. (2014). Reference-free cell mixture adjustments in analysis of DNA methylation data. *Bioinformatics*, 30(10) :1431–1439.
- Houwing-Duistermaat, J. J., Helmer, Q., Balliu, B., van den Akker, E., Tsonaka, R., and Uh, H.-W. (2014). Gene analysis for longitudinal family data using random-effects models. *BMC Proceedings*, 8(Suppl 1) :S88.
- Howie, B., Marchini, J., and Stephens, M. (2011). Genotype Imputation with Thousands of Genomes. *G3&#38; Genes|Genomes|Genetics*, 1(6) :457–470.
- Hu, Y., Hui, Q., and Sun, Y. V. (2014). Association analysis of whole genome sequencing data accounting for longitudinal and family designs. *BMC Proceedings*, 8(Suppl 1) :S89.
- Huang, S., Chaudhary, K., and Garmire, L. X. (2017). More Is Better : Recent Progress in Multi-Omics Data Integration Methods. *Frontiers in Genetics*, 8.
- Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., et al. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, 12(2) :115–121.

- Irizarry, R. A. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2) :249–264.
- Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1) :118–127.
- Jones, P. A. (1999). The DNA methylation paradox. *Trends in genetics : TIG*, 15(1) :34–37.
- Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG : New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(D1) :D353–D361.
- Kaprio, J., Tuomilehto, J., Koskenvuo, M., Romanov, K., Reunanen, A., Eriksson, J., Stengård, J., and Kesäniemi, Y. A. (1992). Concordance for Type 1 (insulin-dependent) and Type 2 (non-insulin-dependent) diabetes mellitus in a population-based cohort of twins in Finland. *Diabetologia*, 35(11) :1060–1067.
- Kerner, B., North, K. E., and Fallin, M. D. (2009). Use of longitudinal data in genetic studies in the genome-wide association studies era : Summary of Group 14. *Genetic Epidemiology*, 33(S1) :S93–S98.
- Klein, R. J. (2005). Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science*, 308(5720) :385–389.
- Kleiner, D. E., Brunt, E. M., Van Natta, M., Behling, C., Contos, M. J., Cummings, O. W., Ferrell, L. D., Liu, Y.-C., Torbenson, M. S., et al. (2005). Design and validation of a histological scoring system for nonalcoholic fatty liver disease. *Hepatology (Baltimore, Md.)*, 41(6) :1313–1321.
- Knowler, W. C., Saad, M. F., Pettitt, D. J., Nelson, R. G., and Bennett, P. H. (1993). Determinants of diabetes mellitus in the Pima Indians. *Diabetes Care*, 16(1) :216–227.
- Kobberling, J. and Tillil, H. (1982). Empirical risk figures for first degree relatives of non-insulin dependent diabetics. *The genetics of diabetes mellitus*, 1(982) :201–209.
- Kurdyukov, S. and Bullock, M. (2016). DNA Methylation Analysis : Choosing the Right Method. *Biology*, 5(1) :3.
- Laird, N. M., Fitzmaurice, G. M., and Schwartz, A. G. (2000). 15 The analysis of case-control data : Epidemiologic studies of familial aggregation. In *Handbook of Statistics*, volume 18, pages 465–482. Elsevier.
- Laird, N. M. and Ware, J. H. (1982). Random-Effects Models for Longitudinal Data. *Biometrics*, 38(4) :963–974.
- Larsen, F., Gundersen, G., Lopez, R., and Prydz, H. (1992). CpG islands as gene markers in the human genome. *Genomics*, 13(4) :1095–1107.

- Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N., and Davey Smith, G. (2008). Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*, 27(8) :1133–1163.
- Lee, Y., Park, S., Moon, S., Lee, J., Elston, R. C., Lee, W., and Won, S. (2014). On the Analysis of a Repeated Measure Design in Genome-Wide Association Analysis. *International Journal of Environmental Research and Public Health*, 11(12) :12283–12303.
- Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., and Storey, J. D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, 28(6) :882–883.
- Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., and Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10) :733–739.
- Levy, J. C., Matthews, D. R., and Hermans, M. P. (1998). Correct homeostasis model assessment (HOMA) evaluation uses the computer program. *Diabetes Care*, 21(12) :2191–2192.
- Li, G., Chang, H., Xia, W., Mao, Z., Li, Y., and Xu, S. (2014). F0 maternal BPA exposure induced glucose intolerance of F2 generation through DNA methylation change in Gck. *Toxicology Letters*, 228(3) :192–199.
- Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1) :13–22.
- Lin, E. and Lane, H.-Y. (2017). Machine learning and systems genomics approaches for multi-omics data. *Biomarker Research*, 5(1).
- Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., Nery, J. R., Lee, L., Ye, Z., et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271) :315–322.
- Liu, J., Huang, J., and Ma, S. (2014). Penalized multivariate linear mixed model for longitudinal genome-wide association studies. *BMC Proceedings*, 8(Suppl 1) :S73.
- Lopez-Romero, P. (2016). *AgiMicroRna : Processing and Differential Expression Analysis of Agilent microRNA chips*. R package version 2.24.0.
- Lu, N., Tang, W., He, H., Yu, Q., Crits-Christoph, P., Zhang, H., and Tu, X. (2009). On the Impact of Parametric Assumptions and Robust Alternatives for Longitudinal Data Analysis. *Biometrical Journal*, 51(4) :627–643.

- Lupski, J. R., Belmont, J. W., Boerwinkle, E., and Gibbs, R. A. (2011). Clan genomics and the complex architecture of human disease. *Cell*, 147(1) :32–43.
- Lyssenko, V., Jonsson, A., Almgren, P., Pulizzi, N., Isomaa, B., Tuomi, T., Berglund, G., Altshuler, D., Nilsson, P., et al. (2008). Clinical risk factors, DNA variants, and the development of type 2 diabetes. *The New England Journal of Medicine*, 359(21) :2220–2232.
- Lyssenko, V. and Laakso, M. (2013). Genetic Screening for the Risk of Type 2 Diabetes : Worthless or valuable? *Diabetes Care*, 36(Supplement\_2) :S120–S126.
- MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., et al. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research*, 45(D1) :D896–D901.
- Mackay, T. F. C., Stone, E. A., and Ayroles, J. F. (2009). The genetics of quantitative traits : Challenges and prospects. *Nature Reviews Genetics*, 10(8) :565–577.
- Maksimovic, J., Gordon, L., and Oshlack, A. (2012). SWAN : Subset-quantile Within Array Normalization for Illumina Infinium HumanMethylation450 BeadChips. *Genome Biology*, 13(6) :R44.
- Mandereau-Bruno, L., Denis, P., Fagot-Campagna, A., and Fosse-Edorh, S. (2014). Prévalence du diabète traité pharmacologiquement et disparités territoriales en France en 2012. *Bull Epidémiol Hebd*, pages 30–31.
- Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., et al. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265) :747–753.
- Marabita, F., Almgren, M., Lindholm, M. E., Ruhrmann, S., Fagerström-Billai, F., Jagodic, M., Sundberg, C. J., Ekström, T. J., Teschendorff, A. E., et al. (2013). An evaluation of analysis pipelines for DNA methylation profiling using the Illumina HumanMethylation450 BeadChip platform. *Epigenetics*, 8(3) :333–346.
- Marquard, V., Beckmann, L., Heid, I. M., Lamina, C., and Chang-Claude, J. (2009). Impact of genotyping errors on the type I error rate and the power of haplotype-based association methods. *BMC Genetics*, 10(1) :3.
- Marullo, L., El-Sayed Moustafa, J. S., and Prokopenko, I. (2014). Insights into the Genetic Susceptibility to Type 2 Diabetes from Genome-Wide Association Studies of Glycaemic Traits. *Current Diabetes Reports*, 14(11).
- Matthews, D. R., Hosker, J. P., Rudenski, A. S., Naylor, B. A., Treacher, D. F., and Turner, R. C. (1985). Homeostasis model assessment : Insulin resistance and beta-cell function from fasting plasma glucose and insulin concentrations in man. *Diabetologia*, 28(7) :412–419.

- McCarthy, M. I. and Zeggini, E. (2009). Genome-wide association studies in type 2 diabetes. *Current diabetes reports*, 9(2) :164–171.
- Mei, H., Chen, W., Jiang, F., He, J., Srinivasan, S., Smith, E. N., Schork, N., Murray, S., and Berenson, G. S. (2012). Longitudinal Replication Studies of GWAS Risk SNPs Influencing Body Mass Index over the Course of Childhood and Adulthood. *PLoS ONE*, 7(2) :e31470.
- Meigs, J. B., Cupples, L. A., and Wilson, P. W. (2000). Parental transmission of type 2 diabetes : The Framingham Offspring Study. *Diabetes*, 49(12) :2201–2207.
- Meyre, D., Delplanque, J., Chèvre, J.-C., Lecoeur, C., Lobbens, S., Gallina, S., Durand, E., Vatin, V., Degraeve, F., et al. (2009). Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in European populations. *Nature Genetics*, 41(2) :157–159.
- Moarii, M., Boeva, V., Vert, J.-P., and Reyal, F. (2015). Changes in correlation between promoter methylation and gene expression in cancer. *BMC Genomics*, 16.
- Moran, S., Arribas, C., and Esteller, M. (2016). Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics*, 8(3) :389–399.
- Morris, A. P., Voight, B. F., Teslovich, T. M., Ferreira, T., Segrè, A. V., Steinthorsdottir, V., Strawbridge, R. J., Khan, H., Grallert, H., et al. (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genetics*, 44(9) :981–990.
- Morris, A. P. and Zeggini, E. (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genetic Epidemiology*, 34(2) :188–193.
- Musolf, A., Nato, A. Q., Londono, D., Zhou, L., Matise, T. C., and Gordon, D. (2014). Mapping genes with longitudinal phenotypes via Bayesian posterior probabilities. *BMC Proceedings*, 8(Suppl 1) :S81.
- Mykkänen, L., Kuusisto, J., Pyörälä, K., and Laakso, M. (1993). Cardiovascular disease risk factors as predictors of type 2 (non-insulin-dependent) diabetes mellitus in elderly subjects. *Diabetologia*, 36(6) :553–559.
- Nica, A. C., Montgomery, S. B., Dimas, A. S., Stranger, B. E., Beazley, C., Barroso, I., and Dermitzakis, E. T. (2010). Candidate Causal Regulatory Effects by Integration of Expression QTLs with Complex Trait Genetic Associations. *PLoS Genetics*, 6(4) :e1000895.
- Noble, D., Mathur, R., Dent, T., Meads, C., and Greenhalgh, T. (2011). Risk models and scores for type 2 diabetes : Systematic review. *BMJ*, 343 :d7163.

- Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., et al. (2008). Genes mirror geography within Europe. *Nature*, 456(7218) :98–101.
- Nygaard, V., Rødland, E. A., and Hovig, E. (2015). Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics*, page kxv027.
- Patterson, N., Price, A. L., and Reich, D. (2006). Population Structure and Eigenanalysis. *PLoS Genetics*, 2(12) :e190.
- Pavkov, M. E., Hanson, R. L., Knowler, W. C., Sievers, M. L., Bennett, P. H., and Nelson, R. G. (2010). Effect of Intrauterine Diabetes Exposure on the Incidence of End-Stage Renal Disease in Young Adults With Type 2 Diabetes. *Diabetes Care*, 33(11) :2396–2398.
- Pearson, T. A. (2008). How to Interpret a Genome-wide Association Study. *JAMA*, 299(11) :1335.
- Peters, T. J., Buckley, M. J., Statham, A. L., Pidsley, R., Samaras, K., V Lord, R., Clark, S. J., and Molloy, P. L. (2015). De novo identification of differentially methylated regions in the human genome. *Epigenetics & Chromatin*, 8 :6.
- Pettitt, D. J., Aleck, K. A., Baird, H. R., Carragher, M. J., Bennett, P. H., and Knowler, W. C. (1988). Congenital susceptibility to NIDDM. Role of intrauterine environment. *Diabetes*, 37(5) :622–628.
- Pettitt, D. J., Baird, H. R., Aleck, K. A., Bennett, P. H., and Knowler, W. C. (1983). Excessive obesity in offspring of Pima Indian women with diabetes during pregnancy. *The New England Journal of Medicine*, 308(5) :242–245.
- Philipson, P., Sousa, I., Diggle, P. J., Williamson, P., Kolamunnage-Dona, R., Henderson, R., and Hickey, G. L. (2017). *joinerR : Joint Modelling of Repeated Measurements and Time-to-Event Data*. R package version 1.2.0.
- Hipkiss, B., Lee, S., Majewski, I. J., Alexander, W. S., and Smyth, G. K. (2016). Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. *The Annals of Applied Statistics*, 10(2) :946–963.
- Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8) :904–909.
- Price, M. E., Cotton, A. M., Lam, L. L., Farré, P., Emberly, E., Brown, C. J., Robinson, W. P., and Kobor, M. S. (2013). Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium Human-Methylation450 BeadChip array. *Epigenetics & Chromatin*, 6(1) :4.
- Prokopenko, I., Langenberg, C., Florez, J. C., Saxena, R., Soranzo, N., Thorleifsson, G., Loos, R. J. F., Manning, A. K., Jackson, A. U., et al. (2009). Variants in MTNR1B influence fasting glucose levels. *Nature Genetics*, 41(1) :77–81.

- Purcell, S. and Chang, C. (2015). PLINK v1.90b3.36.
- Quackenbush, J. (2002). Microarray data normalization and transformation. *Nature Genetics*, 32(Supp) :496–501.
- Raciti, G. A., Nigro, C., Longo, M., Parrillo, L., Miele, C., Formisano, P., and Béguinot, F. (2014). Personalized medicine and type 2 diabetes : Lesson from epigenetics. *Epigenomics*, 6(2) :229–238.
- Rajesh, P. and Balasubramanian, K. (2015). Gestational exposure to di(2-ethylhexyl) phthalate (DEHP) impairs pancreatic  $\beta$ -cell function in F1 rat offspring. *Toxicology Letters*, 232(1) :46–57.
- Ravelli, A. C., van der Meulen, J. H., Michels, R. P., Osmond, C., Barker, D. J., Hales, C. N., and Bleker, O. P. (1998). Glucose tolerance in adults after prenatal exposure to famine. *Lancet (London, England)*, 351(9097) :173–177.
- Ricci, P., Blotière, P.-O., Weill, A., Simon, D., Tuppin, P., Ricordeau, P., and Allemand, H. (2010). Diabète traité : Quelles évolutions entre 2000 et 2009 en France. *Bull Epidemiol Hebd*, 42(43) :425–431.
- Risch, N. and Merikangas, K. (1996). The Future of Genetic Studies of Complex Human Diseases. *Science*, 273(5281) :1516–1517.
- Ritchie, M. E., Silver, J., Oshlack, A., Holmes, M., Diyagama, D., Holloway, A., and Smyth, G. K. (2007). A comparison of background correction methods for two-colour microarrays. *Bioinformatics (Oxford, England)*, 23(20) :2700–2707.
- Rizopoulos, D. (2016a). *JM: Joint Modeling of Longitudinal and Survival Data*. R package version 1.4-5.
- Rizopoulos, D. (2016b). *JMbayes: Joint Modeling of Longitudinal and Time-to-Event Data under a Bayesian Approach*. R package version 0.8-0.
- Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of Regression Coefficients When Some Regressors are not Always Observed. *Journal of the American Statistical Association*, 89(427) :846–866.
- Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data. *Journal of the American Statistical Association*, 90(429) :106–121.
- Rockman, M. V. and Kruglyak, L. (2006). Genetics of global gene expression. *Nature Reviews Genetics*, 7(11) :862–872.
- Roglic, G. and World Health Organization, editors (2016). *Global Report on Diabetes*. World Health Organization, Geneva, Switzerland. OCLC : ocn948336981.

- Roslin, N. M., Hamid, J. S., Paterson, A. D., and Beyene, J. (2009). Genome-wide association analysis of cardiovascular-related quantitative traits in the Framingham Heart Study. *BMC Proceedings*, 3(Suppl 7) :S117.
- Sawicki, M. P., Samara, G., Hurwitz, M., and Passaro, E. (1993). Human Genome Project. *The American Journal of Surgery*, 165(2) :258–264.
- Saxena, R., Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of BioMedical Research, Voight, B. F., Lyssenko, V., Burtt, N. P., de Bakker, P. I. W., Chen, H., Roix, J. J., Kathiresan, S., et al. (2007). Genome-Wide Association Analysis Identifies Loci for Type 2 Diabetes and Triglyceride Levels. *Science*, 316(5829) :1331–1336.
- Saxonov, S., Berg, P., and Brutlag, D. L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences of the United States of America*, 103(5) :1412–1417.
- Schadt, E. E., Lamb, J., Yang, X., Zhu, J., Edwards, S., GuhaThakurta, D., Sieberts, S. K., Monks, S., Reitman, M., et al. (2005). An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics*, 37(7) :710–717.
- Schork, N. J., Murray, S. S., Frazer, K. A., and Topol, E. J. (2009). Common vs. rare allele hypotheses for complex diseases. *Current Opinion in Genetics & Development*, 19(3) :212–219.
- Schultz, M. D., He, Y., Whitaker, J. W., Hariharan, M., Mukamel, E. A., Leung, D., Rajagopal, N., Nery, J. R., Urich, M. A., et al. (2015). Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature*, 523(7559) :212–216.
- Scott, L. J., Mohlke, K. L., Bonycastle, L. L., Willer, C. J., Li, Y., Duren, W. L., Erdos, M. R., Stringham, H. M., Chines, P. S., et al. (2007). A Genome-Wide Association Study of Type 2 Diabetes in Finns Detects Multiple Susceptibility Variants. *Science*, 316(5829) :1341–1345.
- Sham, P., Cherny, S., Purcell, S., and Hewitt, J. (2000). Power of Linkage versus Association Analysis of Quantitative Traits, by Use of Variance-Components Models, for Sibship Data. *The American Journal of Human Genetics*, 66(5) :1616–1630.
- Sikorska, K., Montazeri, N. M., Uitterlinden, A., Rivadeneira, F., Eilers, P. H., and Lesaffre, E. (2015). GWAS with longitudinal phenotypes : Performance of approximate procedures. *European Journal of Human Genetics*.
- Sikorska, K., Rivadeneira, F., Groenen, P. J., Hofman, A., Uitterlinden, A. G., Eilers, P. H., and Lesaffre, E. (2013). Fast linear mixed model computations for genome-wide association studies with longitudinal data. *Statistics in Medicine*, 32(1) :165–180.

- Silver, J. D., Ritchie, M. E., and Smyth, G. K. (2009). Microarray background correction : Maximum likelihood estimation for the normal-exponential convolution. *Biostatistics*, 10(2) :352–363.
- Sitlani, C. M., Rice, K. M., Lumley, T., McKnight, B., Cupples, L. A., Avery, C. L., Noordam, R., Stricker, B. H. C., Whitsel, E. A., et al. (2015). Generalized estimating equations for genome-wide association studies using longitudinal phenotype data. *Statistics in Medicine*, 34(1) :118–130.
- Siva, N. (2008). 1000 Genomes project. *Nature Biotechnology*, 26(3) :256–256.
- Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P., Vincent, D., Belisle, A., et al. (2007). A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 445(7130) :881–885.
- Smith, E. N., Chen, W., Kähönen, M., Kettunen, J., Lehtimäki, T., Peltonen, L., Raitakari, O. T., Salem, R. M., Schork, N. J., et al. (2010). Longitudinal Genome-Wide Association of Cardiovascular Disease Risk Factors in the Bogalusa Heart Study. *PLoS Genet*, 6(9) :e1001094.
- Smyth, G., Hu, Y., Ritchie, M., Silver, J., Wettenhall, J., McCarthy, D., Wu, D., Shi, W., Phipson, B., et al. (2017). *limma : Linear Models for Microarray Data*. R package version 3.30.13.
- Smyth, G. K. (2004). Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1) :1–25.
- Stitzel, M. L., Sethupathy, P., Pearson, D. S., Chines, P. S., Song, L., Erdos, M. R., Welch, R., Parker, S. C. J., Boyle, A. P., et al. (2010). Global epigenomic analysis of primary human pancreatic islets provides insights into type 2 diabetes susceptibility loci. *Cell Metabolism*, 12(5) :443–455.
- Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., et al. (2005). Gene set enrichment analysis : A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43) :15545–15550.
- Sudell, M., Kolamunnage-Dona, R., and Tudur-Smith, C. (2016). Joint models for longitudinal and time-to-event data : A review of reporting quality with a view to meta-analysis. *BMC Medical Research Methodology*, 16(1).
- Sun, Z., Chai, H. S., Wu, Y., White, W. M., Donkena, K. V., Klein, C. J., Garovic, V. D., Therneau, T. M., and Kocher, J.-P. A. (2011). Batch effect correction for genome-wide methylation data with Illumina Infinium platform. *BMC Medical Genomics*, 4(1).
- Syed, H., Jorgensen, A. L., and Morris, A. P. (2016a). Evaluation of methodology for the analysis of ‘time-to-event’ data in pharmacogenomic genome-wide association studies. *Pharmacogenomics*, 17(8) :907–915.

- Syed, H., Jorgensen, A. L., and Morris, A. P. (2016b). SurvivalGWAS\_Power : A user friendly tool for power calculations in pharmacogenetic studies with “time to event” outcomes. *BMC Bioinformatics*, 17(1).
- Tam, C. H. T., Ho, J. S. K., Wang, Y., Lee, H. M., Lam, V. K. L., Germer, S., Martin, M., So, W. Y., Ma, R. C. W., et al. (2010). Common polymorphisms in MTNR1B, G6PC2 and GCK are associated with increased fasting plasma glucose and impaired beta-cell function in Chinese subjects. *PloS One*, 5(7) :e11428.
- Teschendorff, A. E., Breeze, C. E., Zheng, S. C., and Beck, S. (2017). A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. *BMC Bioinformatics*, 18(1).
- Teschendorff, A. E., Marabita, F., Lechner, M., Bartlett, T., Tegner, J., Gomez-Cabrero, D., and Beck, S. (2013). A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics*, 29(2) :189–196.
- Thanabalasingham, G. and Owen, K. R. (2011). Diagnosis and management of maturity onset diabetes of the young (MODY). *BMJ*, 343(oct19 3) :d6044–d6044.
- The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature*, 526(7571) :68–74.
- Toporoff, G., Aran, D., Kark, J. D., Rosenberg, M., Dubnikov, T., Nissan, B., Wainstein, J., Friedlander, Y., Levy-Lahad, E., et al. (2012). Genome-wide survey reveals predisposing diabetes type 2-related DNA methylation variations in human peripheral blood. *Human Molecular Genetics*, 21(2) :371–383.
- Touleimat, N. and Tost, J. (2012). Complete pipeline for Infinium<sup>®</sup> Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. *Epigenomics*, 4(3) :325–341.
- Triche, T. J., Weisenberger, D. J., Van Den Berg, D., Laird, P. W., and Siegmund, K. D. (2013). Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Research*, 41(7) :e90–e90.
- Tu, I.-P. and Whittemore, A. S. (1999). Power of Association and Linkage Tests When the Disease Alleles Are Unobserved. *The American Journal of Human Genetics*, 64(2) :641–649.
- van den Oord, E. J. (2008). Controlling false discoveries in genetic studies. *American Journal of Medical Genetics Part B : Neuropsychiatric Genetics*, 147B(5) :637–644.
- Verbanck, M., Canouil, M., Leloire, A., Dhennin, V., Coumoul, X., Yengo, L., Froguel, P., and Poulain-Codefroy, O. (2017). Low-dose exposure to bisphenols A, F and S of human primary adipocyte impacts coding and non-coding RNA profiles. *PLOS ONE*, 12(6) :e0179583.

- Verbeke, G., Spiessens, B., and Lesaffre, E. (2001). Conditional Linear Mixed Models. *The American Statistician*, 55(1) :25–34.
- Voight, B. F., Kang, H. M., Ding, J., Palmer, C. D., Sidore, C., Chines, P. S., Burtt, N. P., Fuchsberger, C., Li, Y., et al. (2012). The Metabochip, a Custom Genotyping Array for Genetic Studies of Metabolic, Cardiovascular, and Anthropometric Traits. *PLoS Genetics*, 8(8) :e1002793.
- Voight, B. F., Scott, L. J., Steinhorsdottir, V., Morris, A. P., Dina, C., Welch, R. P., Zeggini, E., Huth, C., Aulchenko, Y. S., et al. (2010). Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nature Genetics*, 42(7) :579–589.
- Volkmar, M., Dedeurwaerder, S., Cunha, D. A., Ndlovu, M. N., Defrance, M., Deplus, R., Calonne, E., Volkmar, U., Igoillo-Esteve, M., et al. (2012). DNA methylation profiling identifies epigenetic dysregulation in pancreatic islets from type 2 diabetic patients. *The EMBO Journal*, 31(6) :1405–1426.
- Wagner, J. R., Busche, S., Ge, B., Kwan, T., Pastinen, T., and Blanchette, M. (2014). The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. *Genome Biology*, 15 :R37.
- Wang, S., Gao, W., Ngwa, J., Allard, C., Liu, C.-T., and Cupples, L. A. (2014a). Comparing baseline and longitudinal measures in association studies. *BMC Proceedings*, 8(Suppl 1) :S84.
- Wang, S., Zhang, J., and Lu, W. (2014b). Sample size calculation for the proportional hazards model with a time-dependent covariate. *Computational Statistics & Data Analysis*, 74 :217–227. WOS :000333781500017.
- Weijnen, C. F., Rich, S. S., Meigs, J. B., Krolewski, A. S., and Warram, J. H. (2002). Risk of diabetes in siblings of index cases with Type 2 diabetes : Implications for genetic studies. *Diabetic Medicine*, 19(1) :41–50.
- Wittke-Thompson, J. K., Pluzhnikov, A., and Cox, N. J. (2005). Rational Inferences about Departures from Hardy-Weinberg Equilibrium. *The American Journal of Human Genetics*, 76(6) :967–986.
- World Health Organization (1999). Definition, diagnosis and classification of diabetes mellitus and its complications : Report of a WHO consultation. Part 1, Diagnosis and classification of diabetes mellitus.
- World Health Organization and International Diabetes Federation (2006). *Definition and Diagnosis of Diabetes Mellitus and Intermediate Hyperglycaemia : Report of a WHO/IDF Consultation*.
- Wu, Y. Y. and Briollais, L. (2014). Mixed-effects models for joint modeling of sequence data in longitudinal studies. *BMC Proceedings*, 8(Suppl 1) :S92.

- Xu, Z., Niu, L., Li, L., and Taylor, J. A. (2015). ENmix : A novel background correction method for Illumina HumanMethylation450 BeadChip. *Nucleic Acids Research*, page gkv907.
- Xu, Z., Shen, X., Pan, W., and for the Alzheimer's Disease Neuroimaging Initiative (2014). Longitudinal Analysis Is More Powerful than Cross-Sectional Analysis in Detecting Genetic Association with Neuroimaging Phenotypes. *PLoS ONE*, 9(8) :e102312.
- Yaghootkar, H. and Frayling, T. M. (2013). Recent progress in the use of genetics to understand links between type 2 diabetes and related metabolic traits. *Genome biology*, 14(3) :203.
- Yan, Q., Chen, R., Sutcliffe, J. S., Cook, E. H., Weeks, D. E., Li, B., and Chen, W. (2016). The impact of genotype calling errors on family-based studies. *Scientific Reports*, 6(1).
- Yousefi, P., Huen, K., Schall, R. A., Decker, A., Elboudwarej, E., Quach, H., Barcellos, L., and Holland, N. (2013). Considerations for normalization of DNA methylation data by Illumina 450K BeadChip assay in population studies. *Epigenetics*, 8(11) :1141–1152.
- Zeggini, E. and Ioannidis, J. P. (2009). Meta-analysis in genome-wide association studies. *Pharmacogenomics*, 10(2) :191–201.
- Zhang, C., Qi, L., Hunter, D. J., Meigs, J. B., Manson, J. E., van Dam, R. M., and Hu, F. B. (2006). Variant of transcription factor 7-like 2 (TCF7L2) gene and the risk of type 2 diabetes in large cohorts of U.S. women and men. *Diabetes*, 55(9) :2645–2648.
- Zhang, X., Mu, W., and Zhang, W. (2012). On the Analysis of the Illumina 450k Array Data : Probes Ambiguously Mapped to the Human Genome. *Frontiers in Genetics*, 3.
- Zhao, Q., Xiao, J., He, J., Zhang, X., Hong, J., Kong, X., Mills, K. T., Weng, J., Jia, W., et al. (2014). Cross-Sectional and Longitudinal Replication Analyses of Genome-Wide Association Loci of Type 2 Diabetes in Han Chinese. *PLoS ONE*, 9(3) :e91790.
- Zhi, D., Aslibekyan, S., Irvin, M. R., Claas, S. A., Borecki, I. B., Ordovas, J. M., Absher, D. M., and Arnett, D. K. (2013). SNPs located at CpG sites modulate genome-epigenome interaction. *Epigenetics*, 8(8) :802–806.
- Zhu, J., Wiener, M. C., Zhang, C., Fridman, A., Minch, E., Lum, P. Y., Sachs, J. R., and Schadt, E. E. (2007). Increasing the Power to Detect Causal Associations by Combining Genotypic and Expression Data in Segregating Populations. *PLoS Computational Biology*, 3(4) :e69.
- Ziegler, A., Kastner, C., and Blettner, M. (1998). The Generalised Estimating Equations : An Annotated Bibliography. *Biometrical Journal*, 40(2) :115–139.

Zierath, J. R. and Barrès, R. E. (2011). Research Highlights : Nutritional status affects the epigenomic profile of peripheral blood cells. *Epigenomics*, 3(3) :259–260.