Université de Montréal

Épistasie en médecine évolutive

par

Isabel Gamache

Département de biochimie et médecine moléculaire
Faculté de médecine

Thèse présentée en vue de l'obtention du grade de
Philosophiæ Doctor (Ph.D.)
en Bio-informatique

Juillet 2023

Université de Montréal

Département de médecine

Cette thèse intitulée

Épistasie en médecine évolutive

présentée par

Isabel Gamache

a été évaluée par un jury composé des personnes suivantes :

Daniel Sinnett
(président-rapporteur)

Julie Hussin
(directeur de recherche)

Marie-Pierre Dubé
(codirecteur)

Michelle Scott
(membre du jury)

Claude Bherer
(examinateur externe)

Paul François
(représentant du doyen de la FESP)

Sommaire

La variabilité de la réponse aux médicaments entre les individus est en grande partie attribuable aux différences génétiques causées par des mutations génétiques. Ces mutations ont émergé au cours de l'évolution humaine et peuvent être neutres, bénéfiques ou délétères en termes de survie ou de succès reproductif. Bien que de nombreuses études identifient des variants génétiques associés à des phénotypes, comme la réponse aux médicaments, peu d'attention est accordée à l'origine de ces mutations ou à leur présence au sein des populations. La médecine évolutive entre alors en jeu en étudiant les origines évolutives des mutations associées à des phénotypes. Ce domaine se situe à l'intersection de la médecine et de la biologie évolutive, et il cherche à comprendre comment le corps humain est devenu ce qu'il est aujourd'hui. Cette thèse se concentrera sur l'évolution des gènes impliqués dans la réponse aux médicaments.

La première partie de cette thèse se penchera sur la relation entre les gènes *ADCY9* et *CETP*, qui sont liés à la réponse au médicament dalcetrapib visant à réduire les événements cardiovasculaires en ciblant la protéine CETP. Une mutation dans le gène *ADCY9* a été précédemment identifiée comme modulant la réponse à ce médicament. Nous avons identifié plusieurs pressions de sélection dans le gène *ADCY9*, mais nous avons concentré nos analyses sur son interaction épistasique, c'est-à-dire non linéaire, co-évolutive avec le gène *CETP*. Des effets de cette interaction sur plusieurs phénotypes ont été observés, et des mécanismes potentiels sous-tendant cette pression co-évolutive et son association avec le médicament ont été identifiés.

La deuxième partie de cette thèse sera la suite d'un projet portant sur l'étude des pressions de sélection sur la superfamille des cytochromes P450. Les gènes de cette superfamille sont généralement impliqués dans la détoxification de l'organisme, y compris par la métabolisation d'environ 75% des médicaments couramment prescrits. Des analyses préliminaires ont révélé

des enrichissements de pression de sélection dans deux sous-familles, à savoir les *CYP3A* et les *CYP4F*. Des phénotypes potentiellement sous pressions de sélection ont été identifiés dans la sous-famille des *CYP3A* au sein de la population africaine.

En conclusion, l'intégration de la génétique des populations avec la transcriptomique et les études d'association phénotypiques enrichit notre compréhension des liens entre les pharmacogènes au sein de diverses populations. Cette approche représente un pas de plus vers l'amélioration de la médecine de précision.

Mots clés : Génétique des populations, Transcriptomique, Études d'association phénotypique, Randomisation mendélienne, Pharmacogénomique

Summary

Variability in drug response between individuals is largely due to genetic differences caused by genetic mutations. These mutations have emerged in the course of human evolution and can be neutral, beneficial or deleterious in terms of survival or reproductive success. Although many studies identify genetic variants associated with phenotypes such as drug response, little attention is paid to the origin of these mutations or their presence in the population. This is where evolutionary medicine comes in, studying the evolutionary origins of mutations associated with phenotypes. This field lies at the intersection of medicine and evolutionary biology, and seeks to understand how the human body became what it is today. This thesis will focus on the evolution of genes involved in drug response.

The first part of this thesis will look at the relationship between the genes *ADCY9* and *CETP*, linked to the response to the drug dalcetrapib aimed at reducing cardiovascular events by targeting the CETP protein. A mutation in the *ADCY9* gene has been previously identified as modulating the response to this drug. We identified several selection pressures in the *ADCY9* gene, but focused our analyses on the co-evolutionary epistatic interactions, meaning non-linear. Effects of this interaction on several phenotypes were observed, and potential mechanisms underlying this co-evolutionary pressure and its association with the drug were identified.

The second part of this thesis will follow on from a project investigating selection pressures on the cytochrome P450 superfamily. Genes in this superfamily are generally involved in the detoxification of the body, including the metabolization of around 75% of commonly prescribed drugs. Preliminary analyses have revealed selective pressure enrichments in two subfamilies, *CYP3A* and *CYP4F*. Potential phenotypes under selective pressure were identified in the *CYP3A* subfamily in the African population.

In conclusion, the integration of population genetics with transcriptomics and phenotypic association studies enhances our understanding of the connections among pharmacogenes across diverse populations. This approach signifies another stride towards advancing precision medicine.

Keywords: Population genetics, Transcriptomics, Phenotypic association studies, Mendelian randomization, Pharmacogenomics

Table des matières

Liste des tableaux

Liste des figures

Liste des sigles et abréviations

1000G	1000 Genomes Project
3'SS	Site d'épissage en 3'
5'SS	Site d'épissage en 5'
ABCA1	ATP-binding cassette A1
ADCY9	Adénylate Cyclase 9
ADN	Acide désoxyribonucléique
ADNc	ADN complémentaire
ADME	Absorption, Distribution, Métabolisation, Excrétion
AMD	Dégénérescence maculaire liée à l'âge
AMPc	Adénosine monophosphate cyclique
ANOVA	Analyse de la variance
ARN	Acide ribonucléique
Apo	Apolipoprotéine
ARNm	ARN messager
ATP	Adénosite TriPhosphate
BPS	Site de branchement
BRAC	*BReast CAncer gene*
CAD	Maladie des artères coronariennes
CaG	CARTaGENE
CE	Ester de cholestérol
CETP	Protéine de transfert des esters de cholestérol
CTD	Domaine C-terminal
CVD	Maladie cardiovasculaire

CYP	Cytochrome P450
Domaine RS	Domaine riche en sérine et arginine
EBV	Epstein-Barr virus
EHH	Extended Haplotype Homozygosity
EPAS-1	Endothelial PAS domain-containing protein 1
ESE	activateurs d'épissage exonique
ESS	Inhibiteurs d'épissage exonique
FDR	Taux de fausse découverte
FEV1	Volume expiratoire forcé en une seconde
F_{ST}	Statistique F
FVC	Capacité vitale forcée
GLM	Modèle linéaire généralisée
GTEx	Genotype-Tissue Expression
GWAS	Étude d'association pan-génomique
HDL	Lipoprotéine à haute densité
HDL-c	Cholestérol dans les lipoprotéines à haute densité
HR	Taux de risque
ICD	Classification internationnale des maladies
iHH	integrated EHH
iHS	integrated Haplotype Score
indel	Insertion-Délétion
ISE	activateur d'épissage intronique
ISS	Inhibiteurs d'épissage intronique
IVW	Méthode pondérée par l'inverse de la variance
LCAT	Lecithin-cholesterol acyltransferase
LCL	lignée cellulaire lymphoblastoide
LD	Déséquilibre de liaison
LDL	Lipoprotéine à faible densité
LDL-c	Cholestérol dans les lipoprotéines à faible densité

LRLD	Déséquilibre de liaison sur longue distance
MR	Randomisation Mendélienne
MVMR	randomisation mendélienne multivariable
MXE	Exclusion Mutuelle d'Exons
OMS	Organisation mondiale de la santé
PBS	Statistique des branches populationnelles
PC	Composant principal
PCA	Analyse des composants principaux
PheWAS	Étude d'association pan-phénomique
Pol	Polymérase
Protéine SR	Protéine riche en sérine/arginine
PSI	*Percent Spliced-In*
RCPG	récepteurs couplés aux protéines G
QTL	loci de traits quantitatifs
RCT	Transport inverse de cholestérol
RNA-Seq	Séquençage d'ARN
SA	Sélection sexuellement antagoniste
SNP	Polymorphisme d'un seul nucléotide
snRNA	Petit ARN nucléaire
snRNP	Petite particule ribonucléoprotéique
TF	Facteur de transcription
TG	Triglycéride
tg	transgénique
UKb	Biobanque du Royaume-Uni

À mon cher oiseau Haru,
qui a courageusement tenté d'apprendre le codage et la rédaction de thèse,
même si la connaissance lui faisait défaut.

Remerciements

Je tiens à exprimer ma gratitude et mes sincères remerciements à toutes les personnes qui ont contribué à la réalisation de cette thèse. Leurs soutiens, encouragements et conseils ont été inestimables tout au long de ce parcours passionnant.

Tout d'abord, je souhaite exprimer ma profonde reconnaissance envers ma directrice de thèse, Julie Hussin, pour son expertise, sa patience et sa confiance en moi. Elle m'a accepté dans son laboratoire alors que je n'avais pas encore de connaissance en bio-informatique et elle m'a introduit à un projet qui m'a fait aimer ce domaine.

Je tiens également à remercier ma co-directrice Marie-Pierre Dubé, ainsi que nos collaborateurs, Jean-Claude Tardif et Eric Rhéaume, qui m'ont guidé sur mon projet et sans qui le projet n'aurait pas pu être ce qu'il est aujourd'hui.

Je remercie également les membres de mon laboratoire d'accueil de m'avoir supporté, conseillé et ne pas avoir perdu patience quand j'ai mis le centre de recherche sur la liste noire de nos serveurs gouvernementaux... à deux reprises... Je tiens à remercier particulièrement Jean-Christophe pour ses nombreux conseils et la résolution des problèmes informatiques de mon projet, mais également à Alex, Justin, Raphaël, Cantin, Camille, Fatima, Pamela, Savandara et Dominique pour les nombreuses discussions, liées à la recherche ou non.

Je souhaite également remercier chaleureusement les membres du comité de thèse, Pavel Hamet et Sébastien Lemieux, pour leur temps, leur expertise et leurs suggestions constructives qui ont contribué à améliorer la qualité de ce travail de recherche.

Je tiens également à remercier ma famille pour leur soutien et encouragement tout au long de mon parcours scolaire, depuis mes débuts où je demandais "Quoi? Je dois retourner à l'école demain aussi?" jusqu'à l'achèvement de ce doctorat.

Je veux également exprimer ma reconnaissance envers Magdalena Burchert pour nos dîners ensemble où nous discutions d'ADCY9 et de nos SNPs préférés.

Encore une fois, je tiens à remercier toutes les personnes qui ont joué un rôle dans la réalisation de cette thèse. Votre soutien, votre expertise et votre amitié ont été d'une valeur inestimable.

Introduction

La médecine de précision est une approche de la médecine qui vise à adapter les traitements en fonction des caractéristiques génétiques de chaque individu. Dans ce contexte, la génétique des populations joue un rôle important, car elle permet de comprendre comment les différences génétiques peuvent affecter la santé des individus dans différentes populations.

Les différences génétiques entre des populations s'expliquent de plusieurs manières, l'une d'elle est l'adaptation à l'environnement. Les êtres humains sont confrontés à une grande diversité d'environnements tout au long de leur vie, et ces changements environnementaux peuvent entraîner des adaptations physiologiques. Bien que l'acclimatation puisse produire des adaptations physiologiques temporaires [2], l'adaptation permanente nécessite des changements génétiques. Cette forme d'adaptation est plus lente, car elle nécessite une pression de sélection sur plusieurs générations. Plus précisément, si une variation génétique s'avère bénéfique pour l'organisme, elle est susceptible d'être transmise aux générations suivantes, favorisant ainsi l'évolution de la population - un phénomène connu sous le nom de sélection naturelle.

L'étude de l'évolution d'un génome à travers les populations et le temps est appelée génétique des populations. Cette discipline s'intéresse principalement à l'étude des mutations au sein d'une espèce, en comparant le génome de différentes populations et parfois en le comparant avec d'autres espèces. La génétique des populations est donc un outil essentiel pour la médecine de précision, car elle permet de mettre en perspective la manière dont les différences génétiques affectent la santé des individus dans différentes populations.

Dans cette thèse, nous allons explorer la relation entre des gènes liés à la médecine de précision en utilisant comme outil: la génétique des populations, la transcriptomique et l'association phénotypique, afin de comprendre les effets potentiels de ces gènes chez les humains.

Chapitre 1

Revue de littérature

1.1. Médecine de précision

La médecine de précision est une approche médicale qui vise à personnaliser les soins de santé en fonction de la génétique individuelle, de l'environnement et du mode de vie. Les mutations génétiques, qui peuvent augmenter la susceptibilité de développer certaines pathologies, constituent un élément clé de cette approche. Cependant, la relation entre une mutation ou un gène et un phénotype est souvent complexe, car elle peut être influencée par des facteurs environnementaux et par des interactions avec d'autres gènes [3]. Cette complexité peut entraîner des réponses variables aux traitements médicamenteux et des effets secondaires indésirables, augmentant les risques pour les patients [3, 4].

La médecine de précision peut agir avant l'apparition de la maladie, comme en aidant avec la prévention et le diagnostic. Elle peut aussi aider lors des traitements, tels qu'avec le domaine de la pharmacogénomique.

Dans le domaine de la prévention et du diagnostic, la médecine de précision peut aller de la vérification des antécédents familiaux au séquençage du génome de l'individu. Cette approche est particulièrement utile dans la lutte contre le cancer. Par exemple, les femmes porteuses de mutations dans les gènes *BRCA1* ou *BRCA2* (*BReast CAncer gene* en anglais) sont plus susceptibles de développer le cancer du sein ou des ovaires [5]. En détectant ces mutations, il est même possible d'intervenir avant la survenue de la maladie, tel qu'avec l'ablation préventive des organes dans le cas précédemment mentionné.

En ce qui concerne le traitement, la médecine de précision peut aller jusqu'à séquencer certains gènes spécifiques. Par exemple, une molécule médicamenteuse est métabolisée par

des protéines, ce qui peut activer son effet thérapeutique ou l'inactiver. Ainsi, toute modification de l'activité de ces protéines due à des mutations génétiques pourraient affecter la réponse au médicament, nécessitant ainsi des ajustements dans sa posologie.

1.1.1. Processus pharmacocinétique

Le processus pharmacocinétique regroupe toutes les étapes depuis l'ingestion du médicament jusqu'à son excrétion, soit les étapes de l'absorption, de la distribution, de la métabolisation et de l'excrétion (ADME) [6, 7].

L'étape d'**absorption** correspond à l'entrée du médicament dans le corps. Cette étape est influencée par les propriétés du médicament et la méthode d'administration (par exemple, ingestion, inhalation, injection ou application cutanée). À l'exception de l'injection, le médicament doit traverser des membranes avant d'entrer dans la circulation sanguine, et une partie de la dose peut être métabolisée ou éliminée au cours de cette étape.

Pendant l'étape de **distribution**, le médicament se déplace à travers le corps, souvent par le biais de la circulation sanguine. Le flot sanguin, la propriété lipophile, l'affinité avec les tissus et la taille de la molécule influencent cette étape [8].

La **métabolisation** est une étape de biotransformation de la molécule par des protéines dans un organe ou un tissu. Elle est séparée en deux principales phases afin de rendre les molécules plus solubles dans l'eau, permettant ainsi leur excrétion. Certaines molécules peuvent être actives avant cette étape, tandis que d'autres ont besoin d'être métabolisées pour le devenir [7].

L'**excrétion** est le processus par lequel la molécule métabolisée est éliminée du corps. Cependant, certaines molécules peuvent ne pas être complètement excrétées en raison de leurs propriétés particulières, ce qui peut entraîner leur accumulation dans les tissus. Par exemple, des médicaments lipophiles, c'est-à-dire solubles dans les lipides, ont tendance à s'accumuler dans les cellules adipeuses [8].

1.1.1.1. Cytochromes P450

Les deux phases de la métabolisation impliquent de nombreuses enzymes. Toute perturbation de ces enzymes peut avoir des conséquences sur la réponse du médicament.

Parmi les enzymes impliquées dans cette étape, la famille des cytochromes P450 (CYP) joue un rôle crucial dans le métabolisme de 70 à 80% des médicaments couramment utilisés

en clinique [9, 10]. Les 20 à 30% des médicaments restant incluent ceux qui ne passent pas par la phase I de la métabolisation, où les CYP agissent, ou sont métabolisés par d'autres enzymes de la phase I, comme des monooxygénases [11, 12, 13].

La superfamille des CYP est apparue tôt dans l'évolution des organismes vivants et elle est composée de 57 gènes et 58 pseudogènes, soit des gènes inactifs, regroupés en familles et sous-familles en fonction de leur similarité. Les CYP sont principalement présents dans le foie et ont évolué pour métaboliser les molécules xénobiotiques, c'est-à-dire des molécules provenant de l'extérieur du corps, telles que les médicaments. Les familles *CYP2*, *CYP3* et *CYP4* contiennent plus de gènes que les 15 autres familles combinées [14].

La famille des *CYP2* contient 36 gènes, dont le plus connu est le gène *CYP2D6*, responsable de la métabolisation d'environ 25% des médicaments [15]. Sa popularité a commencé dans les années 70 où il a été trouvé que le métabolisme de médicaments, tels que le débrisoquine, un anti-hypertenseur, et la spartéine, un antiarythmique, semblait être fortement variable et pourtant contrôlé par un seul gène autosomal [16], le gène *CYP2D6*. D'importantes variations peuvent être observées dans son activité métabolique entre les individus. C'est pourquoi, lors de la prise de certains médicaments, il est important de déterminer son niveau d'activité afin de pouvoir ajuster les doses de médicaments afin de garantir leur efficacité et minimiser les risques d'effets secondaires indésirables [4, 17].

La famille des *CYP3* ne contient que la sous-famille des *CYP3A* chez les mammifères. À elle seule, elle métabolise environ 50% des médicaments souvent utilisés, mais participe également à plusieurs autres processus biologiques, tels que la biotransformation du cholestérol, des hormones stéroïdiennes et de la vitamine D [14]. Cette famille contient quatre gènes, soit *CYP3A4, CYP3A5, CYP3A7* et *CYP3A43*, qui sont adjacents sur le chromosome 7, mais également quatre pseudogènes. Cette famille a été générée par des événements de duplication, avec le gène *CYP3A43* étant considéré comme l'ancestral [18, 19]. Parmi ces gènes, *CYP3A4* et *CYP3A5* sont les mieux caractérisés puisqu'ils sont ceux qui métabolisent le plus de médicaments. Même s'ils partagent plusieurs fonctions, ils sont régulés par différents mécanismes [14]. Le gène *CYP3A4* semble être la forme la plus exprimée chez les individus de descendance européenne, tandis que le gène *CYP3A5* est dominant dans les individus d'origine africaine [20, 21]. Cependant, la connaissance de l'effet de mutations dans cette

famille dans le domaine de la pharmacogénomique reste encore à être approfondie afin d'être utilisée dans l'ajustement des traitements, comme pour le gène *CYP2D6*.

La famille des *CYP4* contient 38 gènes dans 11 sous-familles, dont seulement six chez l'humain, soit les sous-familles A, B, F, V, X et Z. Contrairement aux *CYP* 1 à 3, cette famille a plutôt un rôle dans le métabolisme des acides gras. Même si certains rôles dans le métabolisme de médicament ont également été trouvés, ils sont moindre que ceux des trois premières familles [22, 23]. Les gènes de cette famille sont exprimés dans différents tissus. De plus, plusieurs de ses sous-familles, particulièrement pour la sous-famille des *CYP4F*, présentent un haut degré d'homologie, ce qui rend difficile le ciblage spécifique lors du développement d'anticorps ou même l'évaluation des niveaux d'expression relative entre eux [22, 23]. Cela complique donc les études fonctionnelles de ces gènes.

1.1.2. Réponses pharmacogénomiques modulées par des gènes autres que ceux du processus pharmacocinétique

Outre les gènes liés au processus pharmacocinétique, il peut arriver que les analyses de pharmacogénomique s'intéressent à des gènes qui ne sont pas directement liés à ces voies métaboliques, tels que l'exemple du gène *ADCY9* associé à la réponse du médicament dalcetrapib [24]. En effet, ce médicament a été développé pour diminuer la survenue d'événements coronariens (CVD, *CardioVascular Disease* en anglais) en inhibant la protéine CETP (voir section 1.1.2.1). Cependant, une mutation dans le gène *ADCY9* (voir section 1.1.2.2), qui n'est pas impliqué dans les étape de l'ADME, a été trouvée comme modulant la réponse au médicament. En premier, je vais présenter les concepts derrière le développement du dalcetrapib.

Les maladies cardiovasculaires sont la première cause de mortalité pathologique à travers le monde, ce qui en fait un sujet d'intérêt pour la recherche fondamentale et pharmaceutique afin de comprendre et de prévenir ces maladies. Elles comprennent un ensemble de maladies affectant le coeur et les vaisseaux sanguins. L'athérosclérose est une forme courante de maladie cardiovasculaire qui se caractérise par l'accumulation de plaque de substance, telle que de cholestérol, à l'intérieur de la paroi des artères. Cette accumulation réduit la taille de l'ouverture des vaisseaux sanguins, limitant ainsi le flux sanguin. Dans les cas les plus graves,

impliquant les artères coronariens, le flux sanguin peut s'arrêter complètement, privant ainsi le muscle cardiaque d'oxygène, entraînant un infarctus du myocarde.

Plusieurs facteurs peuvent augmenter le risque de développer l'athérosclérose, tels que le taux de cholestérol sanguin [25, 26]. Le cholestérol est transporté dans le corps par différentes lipoprotéines, notamment les lipoprotéines à faible densité (LDL, *Low Density Lipoprotein* en anglais) et très faible densité (VLDL, *Very Low Density Lipoprotein* en anglais) qui sont responsables du transport du cholestérol et des acides gras vers tous les tissus du corps, ce qui peut potentiellement mener à des accumulations de cholestérol à ces endroits. À l'inverse, le transport inverse du cholestérol (RCT, *Reverse Cholesterol Transport* en anglais) se fait via les lipoprotéines de haute densité (HDL, *High Density Lipoprotein* en anglais) qui permettent de récupérer le cholestérol des tissus, incluant ceux dans les parois des vaisseaux sanguins, et de le ramener au foie afin d'être éliminé de l'organisme [27].

Un taux élevé de cholestérol transporté par les LDL (LDL-c) dans le sang augmente l'accumulation des plaques de cholestérol dans les parois artérielles, et donc, augmente le risque de développer de l'athérosclérose [26, 27]. L'infiltration et la rétention du cholestérol dans les parois artérielles vont déclencher des réponses inflammatoires. L'inflammation va recruter des monocytes, qui vont s'infiltrer dans ces parois et se différencier en macrophages. Ces cellules immunitaires vont alors ingérer le cholestérol présent dans les vaisseaux et se transformer en macrophages spumeux, c'est-à-dire des macrophages remplis de cholestérol. La présence de ces macrophages spumeux contribue à l'inflammation et facilite le développement des plaques d'athérosclérose dans les artères. Le HDL promeut l'efflux du cholestérol en dehors des cellules, incluant les macrophages spumeux, ce qui aide à la prévention de l'inflammation et donc au développement de ces plaques [26].

Les particules de HDL sont extrêmement hétérogènes en termes de structure, de composition protéique (principalement l'apolipoprotéine A-I et A-II (apoA-I et apoA-II respectivement), mais également plusieurs lipoprotéines secondaires comme apoE et apoC-III), de métabolisme et de fonction, notamment dans leurs rôles liés à l'élimination et au transport du cholestérol, ce qui permet de les différencier en plusieurs sous-types distincts de HDL, mais qu'on classifie généralement en trois classes principales, soit large, intermédiaire et petite [28]. En raison de ces différences, tous les sous-types n'ont pas le même impact sur les maladies cardiovasculaires [29].

Il existe deux principales classes étudiées, soit les larges et moins denses HDL_2, ainsi que les petits et denses HDL_3. Les petites particules de HDL, soit les HDL_3, sont davantage associées avec des propriétés anti-inflammatoires, ont une meilleure capacité d'efflux de cholestérol des macrophages via la protéine ABCA1 (*ATP-Binding Cassette A1* en anglais), possèdent potentiellement une activité athéroprotectrice, mais leur effet sur les risques cardiovasculaires est inconsistant à travers les études [28, 30]. Ces petits HDL maturent ensuite avec l'accumulation de l'ester de cholestérol, via la protéine CETP (*Cholesteryl-Ester Transfer Protein* en anglais) et sont convertis en larges HDL_2. Il a été observé que des niveaux élevés de larges particules de HDL, soit les HDL_2, présentent des fonctions antioxydantes et elles sont souvent associées à une réduction du risque cardiovasculaire [28, 31, 32, 33, 34, 35].

Des traitements visant à augmenter les niveaux de cholestérol dans les HDL (HDL-c) et à améliorer le processus RCT sont actuellement à l'étude pour prévenir ou traiter les maladies cardiovasculaires, particulièrement l'athérosclérose [26, 36, 37, 38, 39]. Plusieurs protéines sont impliquées dans ce processus, incluant les protéines ABCA1, LCAT (*Lecithin-Cholesterol AcylTransferase* en anglais) et CETP, ce qui en font des cibles thérapeutiques intéressantes pour la prévention ou le traitement de ces maladies.

1.1.2.1. CETP

CETP et son rôle dans le transport du cholestérol. Le gène *CETP* a 16 exons et se trouve sur le chromosome 16. Il est exprimé dans une grande variété de tissus, mais principalement dans les tissus riches en macrophage, tels que la rate et le foie, et riches en lipide, tels que les tissus adipeux [40, 41, 42, 43]. Sa forme protéique est majoritairement excrétée dans le plasma, principalement par le foie et les adipocytes [40], où elle exerce des fonctions de transferts d'esters de cholestérol (CE, *Cholesteryl Ester* en anglais) et triglycérides (TG) entre les HDL et les VLDL/LDL [26, 36, 44], ainsi que des transferts d'esters de cholestérol entre les différents sous-types de HDL [36]. Elle a également été trouvée dans le fluide cérébro-spinal, où il a été suggéré qu'elle puisse exercer ses fonctions sur les lipides. La protéine CETP pourrait donc avoir un rôle important dans le transport de lipides et leur redistribution à travers le système nerveux central [45]. Plusieurs facteurs ont également été identifiés comme pouvant influencer la production de CETP, tels que l'apport alimentaire en cholestérol [46, 47] et les infections par le virus Epstein-Barr (EBV, *Epstein-Barr Virus* en anglais) [48].

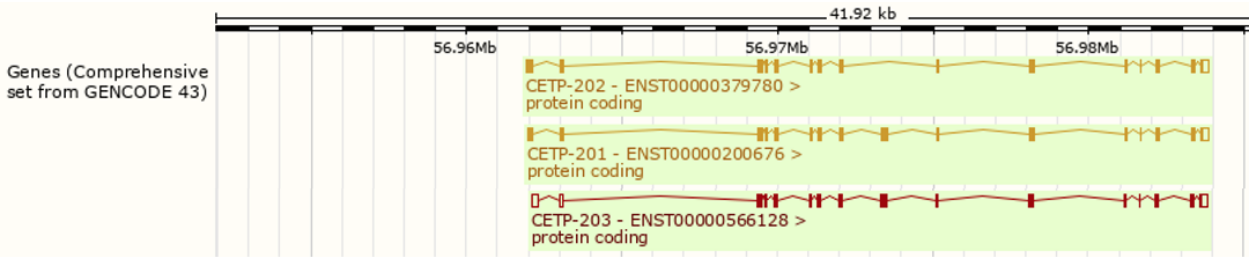

Fig. 1.1. Isoformes protéiques de *CETP*
Les trois formes protéiques du gène *CETP* selon la base de donnée d'*Ensembl* [64] avec leurs coordonnée sur le chromosome 16 sur le génome de référence GRCh38.

Des fonctions au niveau intracellulaire ont également été suggérées, proposant que cette protéine joue un rôle important dans l'homéostasie lipidique d'une cellule, telles que le transport du cholestérol, la biosynthèse des triglycérides, le stokage des lipides et la composition de la membrane [36, 44, 49, 50, 51, 52].

De par ses effets sur le fonctionnement des cellules et le contrôle du cholestérol, de nombreux phénotypes ont été trouvés comme étant associés avec l'expression de *CETP*, incluant les maladies cardiovasculaires [36, 53], la réponse à la septicémie [54, 55, 56, 57], un changement de la masse lipidique selon l'alimentation [58] et la dégénérescence maculaire liée à l'âge (AMD, *Age-related Macular Degeneration* en anglais) [59, 60, 61, 62]. Même si, généralement, la protéine CETP est connue comme favorisant le risque de développer des maladies cardiovasculaires, certaines études montrent des effets contradictoires. Par exemple, la mutation rs5882 dans l'exon 14 diminue son expression, mais cette mutation est associée à une augmentation des événements cardiovasculaires chez les hommes qui font de l'hypertriglycéridémie [63], ce qui peut suggérer que l'effet de la protéine CETP pourrait différer selon le profil métabolique des individus.

Les différents isoformes de CETP. Selon la base de données d'*Ensembl*, le gène *CETP* possède trois formes protéiques (Figure 1.1). La forme la plus abondante, appelée CETP-201, comprend les 16 exons du gène et est la seule forme présente dans le plasma, suggérant qu'elle est responsable des fonctions plasmatiques de CETP. La forme CETP-202, qui est la deuxième forme, a son neuvième exon épissé, ce qui lui enlève sa capacité de se lier aux lipoprotéines, mais pourrait avoir conservé les fonctions de transfert de cholestérol. Étant retenue dans la cellule, cela pourrait suggérer des fonctions agissant sur le cholestérol intracellulaire [43, 51, 65]. De plus, même en étant intracellulaire, elle pourrait influencer l'activité

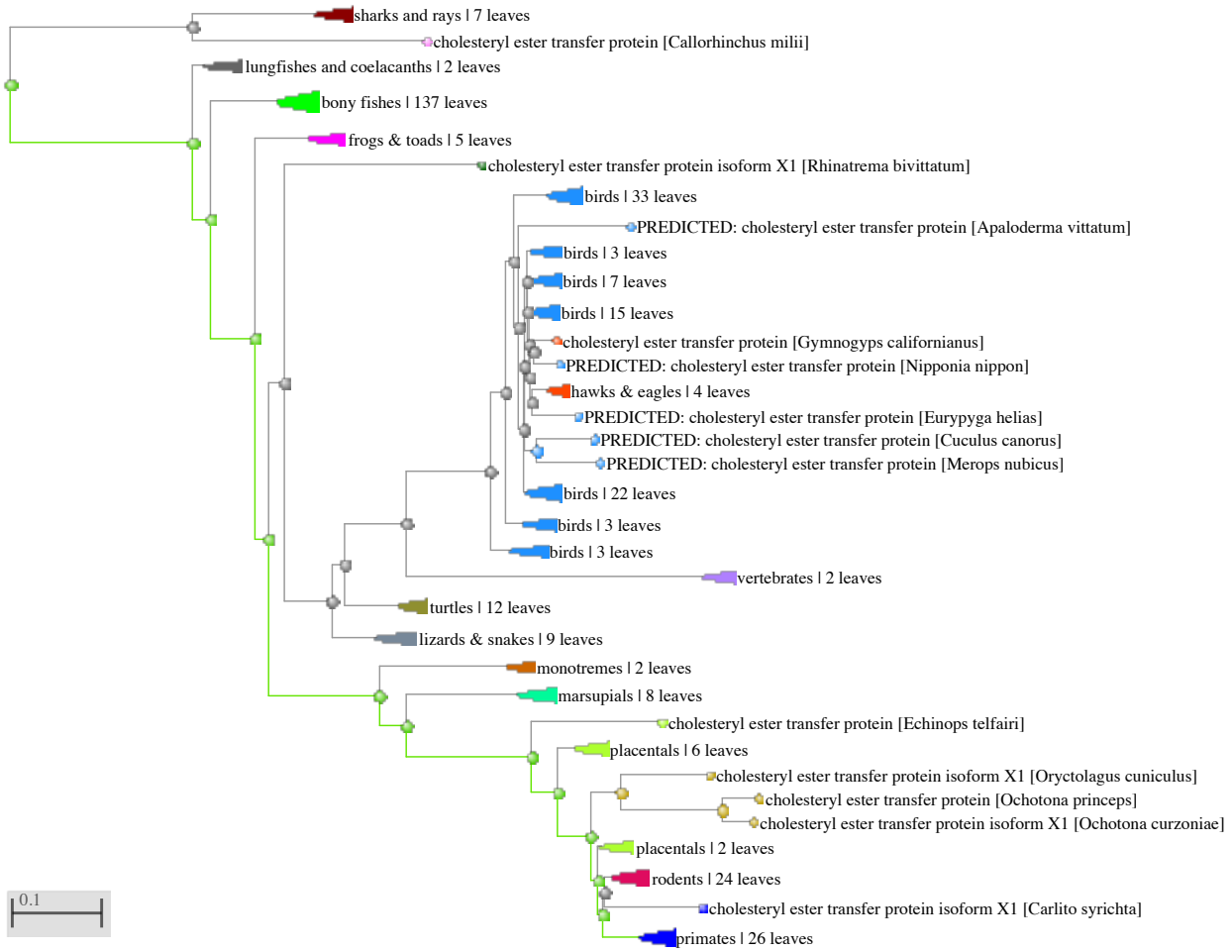

Fig. 1.2. Arbre phylogénétique du gène de *CETP* entre les espèces avec COBALT [1] de la base de donnée de NCBI
La ligne verte indique le regroupement où se trouve l'espèce *Homo Sapien*. Les paramètres utilisés sont par défauts. L'arbre a été généré en considérant 383 espèces vertébrées. L'échelle indique le nombre de changements nucléotidiques par position génétique.

de CETP plasmatique en inhibant la sécrétion de CETP-201 [42, 66]. Un changement dans le ratio d'ARN messager (ARNm) entre la forme *CETP-201* et *CETP-202* est observé dépendamment de l'alimentation et le stade de développement de tissus [67], suggérant que la régulation de leur ratio pourrait avoir des effets sur des phénotypes. La troisième forme, CETP-203, a également 16 exons, mais le premier exon diffère des deux autres formes. De ma connaissance, aucune littérature n'a étudié cette forme.

Portrait de l'évolution de CETP. Le gène *CETP* est présent dans de nombreuses espèces, incluant les mammifères, tels que le lapin et les primates, mais également dans d'autres classes, telles que chez les oiseaux et les poissons, suggérant que ce gène soit apparu tôt dû

à un besoin évolutif (Figure 1.2). Cependant, le gène est devenu totalement inactif dans plusieurs espèces, telles que les souris et les rats [68, 69, 70]. De plus, même si plusieurs espèces ont le gène, le transcrit *CETP-202* semble être apparu tard dans l'évolution des primates [69].

Plusieurs facteurs peuvent causer la perte d'un gène au cours de l'évolution [71]. Parmi ces causes, il y a la redondance de gènes ou de voies signalétiques, c'est-à-dire que les fonctions de plusieurs gènes ou voies signalétiques seront les mêmes ou seront similaires. Lorsqu'il y a une redondance, il y aura des gènes qui prendront la relève des fonctions d'un autre gène, faisant en sorte que des mutations peuvent s'accumuler dans un des gènes sans avoir d'effet important sur l'organisme jusqu'à ce que sa protéine ne soit plus fonctionnelle, générant des pseudogènes. Une autre cause de la perte de fonction de ce gène serait dû à un changement dans le besoin de l'environnement. Par exemple, la fonction du gène était nécessaire à l'organisme pour survivre aux conditions environnementales initiales. Cependant, suite à une migration, le gène n'est plus nécessaire, ou peut même devenir délétère, dans le nouvel environnement, faisant en sorte que le gène devient inactif [71].

Dans le cas du gène *CETP*, plusieurs hypothèses ont été amenées. Une des hypothèses [57] est liée au fait que les rats et les souris vivent dans un environnement qui les rend plus exposés à des pathogènes. Lors d'une infection bactérienne, les niveaux de la protéine CETP sont diminués afin d'augmenter les niveaux de HDL-c, ce qui pourrait aider à la survie des individus [57]. En effet, les molécules de HDL aident dans la réponse aux infections, telles qu'en neutralisant certaines bactéries en s'y liant, en modulant la disponibilité du cholestérol pour les cellules du système immunitaire ou en modulant la réponse inflammatoire [72]. Les espèces ayant une exposition plus élevée aux bactéries auraient donc besoin d'un niveau plus élevé de HDL-c, ce qui aurait conduit à une diminution, voire à une absence de la protéine CETP chez ces espèces. Les autres fonctions de CETP reliées aux lipides ont donc potentiellement été comblées par un autre gène, faisant en sorte que le gène *CETP* mute jusqu'à ne plus être fonctionnel, ne laissant que les reliques, des pseudogènes, observées aujourd'hui dans les génomes de ces espèces [70].

CETP comme cible pharmaceutique pour les maladies cardiovasculaires. Des mutations génétiques dans le locus de *CETP* ont été associées avec le niveau de HDL-c, mais également avec les maladies cardiovasculaires [73, 74]. De nombreuses études ont lié une

diminution de l'activité de CETP à une diminution du développement de plaques athérosclérotiques et de maladies cardiovasculaires, passant potentiellement par une augmentation du niveau de HDL-c et une diminution du niveau de LDL-c [53, 75]. Du fait de son association avec les maladies cardiovasculaires, la protéine CETP a été ciblée par plusieurs inhibiteurs développés par des compagnies pharmaceutiques (Tableau 1.1), tels que le torcetrapib [76], le dalcetrapib [77], l'anacetrapib [78], l'evacetrapib [79] et, plus récemment, l'obicetrapib [80, 81, 82]. Tous ont augmenté significativement les niveaux de HDL-c, mais l'effet sur les maladies cardiovasculaires n'a pas été suffisant [76, 77, 78, 79].

Essai clinique (Médicament)	**Effets**	**Nombre de participants**	**Référence**
ILLUMINATE (Torcetrapib)	$\uparrow$ 72% HDL-c, $\downarrow$ LDL-c, $\uparrow$ 25% CVD et mortalité	15 067	[76]
dal-OUTCOMES (Dalcetrapib)	$\uparrow$ 30% HDL-c, $\approx$ LDL-c, $\approx$ CVD	15 871	[77]
ACCELERATE (Evacetrapib)	$\uparrow$ 133% HDL-c, $\downarrow$ LDL-c, $\approx$ CVD	12 092	[79]
REVEAL (Anacetrapib)	$\uparrow$ 104% HDL-c, $\downarrow$ LDL-c, $\downarrow$ 9% CVD	30 449	[78]
TULIP (Obicetrapib)	$\uparrow$ 135-165% HDL-c, $\downarrow$ LDL-c, ? CVD	364	[80, 81, 82]

Tableau 1.1. Tableaux des essais cliniques des inhibiteurs de CETP

Les effets de la prise du médicament comparés à la prise du placebo sont indiqués par les symboles : $\uparrow$ indique une augmentation significative, $\downarrow$ indique une diminution significative, $\approx$ indique que la différence n'est pas significative et ? indique que l'effet n'est pas connu.

Le torcetrapib, durant sa phase clinique 3 de ILLUMINATE (*Investigation of Lipid Level Management to Understand its Impact in Atherosclerotic Events*), a augmenté de 25% les événements cardiovasculaires et la pression sanguine, ce qui a entraîné l'arrêt de son étude clinique. Il a été suggéré que ces effets secondaires importants sont causés par des effets hors cible [76], c'est-à-dire que la prise du médicament cause des effets qui ne sont pas attribuables à l'inhibition de CETP. Le dalcetrapib et l'evacetrapib, de par leur essai cliniques dal-OUTCOMES et ACCELERATE (*Assessment of Clinical Effects of Cholesteryl Ester Transfer Protein Inhibition with Evacetrapib in Patients at a High Risk for Vascular Outcomes*) respectivement, n'ont pas montré d'effet significatif sur la diminution des événements cardiovasculaires malgré leur augmentation significative du niveau de HDL-c, et leurs essais cliniques ont donc été arrêtés pour cause d'un manque d'efficacité. Durant l'essai clinique REVEAL (*Randomized EValuation of the Effects of Anacetrapib through Lipid modification*), évaluant les effets de l'anacetrapib, une diminution significative, mais modeste, des événements cardiovasculaires a pu être observée. Plusieurs hypothèses ont été amenées sur la raison pourquoi un effet significatif a pu être observé seulement pour l'inhibiteur anacetrapib. Une hypothèse serait la taille de sa cohorte, soit 30 449 patients, comparée à 15

871 patients pour dal-OUTCOMES et 12 092 patients pour ACCELERATE. Une autre hypothèse serait qu'anacetrapib agirait sur la protéine CETP via un mécanisme d'inhibition différent [36]. Le dernier inhibiteur, l'obicetrapib, a montré des effets prometteurs avec une forte diminution de LDL-c et augmente la concentration de HDL-c, cible secondaire de cet essai clinique. Son effet sur les événements cardiovasculaires reste encore à être étudié.

Les effets mitigés de ces inhibiteurs sur les événements cardiovasculaires pourraient partiellement être expliqués par l'amplitude de leur effet sur les sous-types des HDL. Les divers inhibiteurs agissent pour augmenter les niveaux de HDL-c, mais leur impact varie considérablement selon les sous-types de HDL. Cette différence résulte des mécanismes d'action propres à chacun de ces médicaments. À l'exception du dalcetrapib, les autres inhibiteurs agissent sur le transfert entre les sous-types de HDL, pouvant ainsi perturber l'équilibre entre ces différentes populations, ainsi que perturber l'efflux de cholestérol. Par exemple, l'evacetrapib a été associé à une augmentation significative de certains sous-types de HDL contenant des niveaux élevés de lipoprotéines spécifiques, telles que l'apoC-III, connues pour leur dysfonctionnement et leurs effets négatifs sur la santé cardiovasculaire [83, 84]. Ainsi, cette augmentation sélective de certains sous-types de HDL pourrait potentiellement annuler les bénéfices protecteurs attendus. De plus, une augmentation excessive de la concentration de HDL-c, même si elle est bénéfique, peut devenir délétère si cette concentration excède la capacité de clairance hépatique [36].

Le cas du dalcetrapib et de son étude pharmacogénomique. Parmi les inhibiteurs mentionnés ci-haut, le dalcetrapib est un modulateur partiel de CETP. Il inhibe l'échange de l'ester de cholestérol et de triglycéride entre les VLDL/LDL et HDL, mais n'impacte pas la fonction de l'échange entre les sous-types de HDL (Figure 1.3). Une analyse pharmacogénomique a été effectuée dans un sous-groupe de l'essai clinique dal-OUTCOMES afin d'identifier si la génétique pouvait influencer les différences de réponse à ce médicament. Durant cette étude, la mutation rs1967309 dans le gène *ADCY9* a été trouvée comme modulant la réponse à ce médicament. Les individus qui ont le génotype AA (fréquence allélique de l'allèle A de 41% dans la cohorte) ont montré une diminution de 39% des événements cardiovasculaires avec le dalcetrapib alors que les individus qui ont le génotype GG ont montré une augmentation de 27% de ces mêmes événements [24]. La région génétique de la mutation rs1967309 a également été associée avec d'autres phénotypes, tels qu'un changement dans

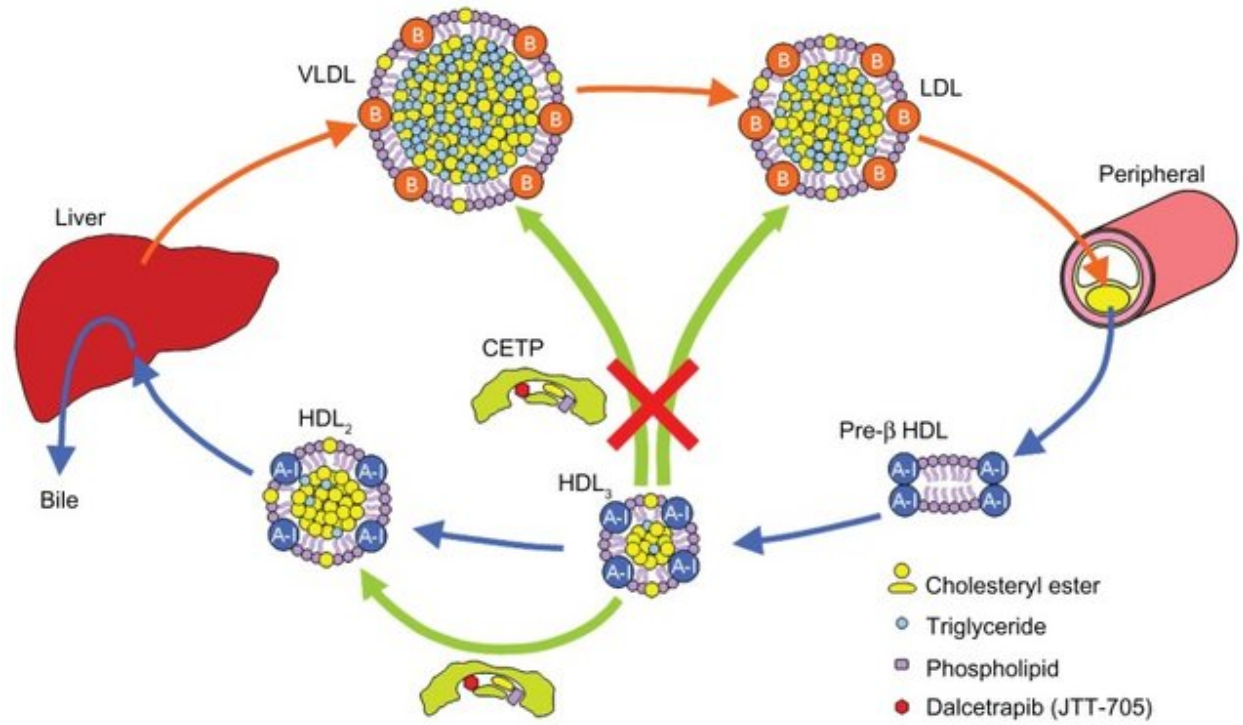

Fig. 1.3. Les effets du dalcetrapib sur la protéine CETP plasmatique
Schématisation de l'effet du modulateur dalcetrapib sur la fonction de la protéine CETP dans le transport inverse du cholestérol. Les flèches oranges indiquent le transport du cholestérol vers les tissus périphériques. Les flèches bleues indiquent le transport inverse du cholestérol des tissus vers le foie. Les flèches vertes représentent la fonction de transfert du cholestérol par la protéine CETP. Le X rouge représente la fonction inhibée de la protéine CETP par le médicament dalcetrapib, représenté par un petit cercle rouge. Source : Shinkai et al. [36].

l'épaisseur de l'intima-média de la carotide, dans la prise de poids et sur les niveaux de la protéine C-réactive (CRP, *C-Reactive Protein* en anglais) durant l'étude dal-PLAQUE-2, qui s'assure de la sécurité et de l'efficacité du dalcetrapib [24]. À la suite de ces observations, un essai clinique dal-GenE a été effectué afin d'évaluer l'effet du dalcetrapib dans le sous-groupe des individus qui ont le génotype AA pour la mutation rs1967309 [85]. Malgré une légère diminution des événements cardiovasculaires (taux de risque 0.88 [0.75-1.03]) chez les individus traités, la différence ne s'est pas avérée significative. Cependant, l'étude clinique a débuté avant le début de la pandémie de COVID-19 dû au virus du SARS-CoV-2 et s'est terminée pendant la pandémie. Quand la pandémie a débuté, les visites à l'hôpital ont diminué et plusieurs diagnostics ont potentiellement été manqués, ce qui inclut les maladies cardiovasculaires [86, 87]. Le nombre d'événements non mortels a été diminué, probablement dû à la diminution des visites aux hôpitaux, tandis que la mortalité, principalement

celle qui n'est pas liée aux événements cardiovasculaires, a augmenté. En ne gardant que les résultats obtenus avant la pandémie, une diminution nominale du nombre d'événements a pu être observée dans le groupe traité (taux de risque=0.82 [0.68-0.98]). Cependant, cette diminution des visites ne devrait pas avoir affecté un groupe plus que l'autre. Cela pourrait signifier qu'il pourrait y avoir d'autres facteurs qui ont influencé cette perte de significativité, tels que la perte de participants, causant une diminution de la puissance statistique. Une autre explication possible serait un effet d'interaction entre SARS-CoV-2 et la génétique, où la diminution de l'activité de CETP combiné à un génotype AA (dans rs1967309) n'agit pas de la même manière lors d'une infection par le virus de SARS-CoV-2, puisque l'activité de CETP peut être modulée lors d'une infection [48].

1.1.2.2. ADCY9

***ADCY9* et sa relation avec les inhibiteurs de CETP.** La relation du gène *ADCY9* avec la réponse au médicament dalcetrapib n'est pas encore bien comprise. De plus, même si la relation a été identifiée avec le dalcetrapib, la mutation rs1967309 n'est pas significativement associée avec les inhibiteurs evacetrapib et anacetrapib (valeurs p de 0.17 et 0.96 respectivement) [78, 88]. Cette différence pourrait être causée par le mécanisme d'action de ces inhibiteurs. CETP a des fonctions d'échange entre les HDL et les VLDL/LDL, mais également entre les sous-types des HDL. Le dalcetrapib inhibe seulement la première fonction contrairement aux autres inhibiteurs qui inhibent complètement CETP. De plus, les inhibiteurs de CETP entrent dans les cellules, où la protéine CETP est également présente sous différentes formes. Il est possible que l'action des inhibiteurs affecte différemment les isoformes, notamment dans le cas du dalcetrapib qui ne devrait théoriquement pas influencer l'isoforme CETP-203. En effet, ce médicament cible un résidu présent dans le premier exon, qui est absent dans cet isoforme spécifique [36].

Dans une récente étude faite chez des souris, les auteurs ont identifié une interaction entre les gènes *Adcy9* et $CETP^{transgenique(tg)}$ [89]. Les souris ont été génétiquement modifiées afin d'exprimer la protéine CETP. Ils ont comparé différents groupes de souris afin de voir l'effet de l'absence ou de la présence d'*Adcy9* et/ou $CETP^{tg}$. Ils ont trouvé que les souris n'ayant aucun des deux gènes sont protégées de l'athérosclérose et il y a une amélioration des fonctions endothéliales, ainsi que plusieurs effets sur le système nerveux autonome. Cela

montre donc, au moins chez les souris transgéniques, la présence d'une interaction entre les gènes *Adcy9* et *CETP*.

Rôles des ADCY. Les protéines de la famille des ADénylates CYclases (ADCY) jouent un rôle important dans toutes les cellules de l'organisme. Il existe dix gènes connus de cette famille. À l'exception d'ADCY10, les ADCY sont des protéines à douze passages transmembranaires [90]. Parmi les dix, la protéine ADCY9 est l'une des moins bien caractérisée.

La principale fonction de cette famille est la conversion des molécules d'adénosine triphosphate (ATP) en adénosine monophosphate cyclique (AMPc, *cyclic Adenosine Mono-Phosphate* en anglais) et en pyrophosphate. Les molécules d'AMPc vont servir de signaux régulateurs via des protéines spécifiques, telles que des facteurs de transcription, d'autres enzymes et des transporteurs d'ion. Les adénylates cyclases sont les principaux effecteurs de la signalisation trans-membranaire du récepteur couplé à la protéine G (RCPG, *G Protein-Coupled Receptors* en anglais) [91]. Ces récepteurs sont les cibles de nombreuses médications, afin de traiter, entre autres, la haute pression sanguine et l'asthme. Les RCPG incluent la classe des récepteurs adrénergiques, sous-divisés en deux groupes principaux α et β, et ont comme ligands la classe de l'épinéphrine. Les récepteurs α_1-adrénergiques sont jumelés au protéine G_q et les α_2-adrénergiques sont jumelés aux protéines G inhibitrices (G_i). Les récepteurs β sont divisés en trois tous liés aux protéines G stimulatrices (G_s). Les protéines G_i et G_s sont liées aux adénylates cyclases et les activent.

Chez les mammifères, lorsque l'épinéphrine se lie aux récepteurs β-adrénergiques, il y aura activation des protéines G_s qui activeront les ADCY afin de mener à la production d'AMPc. Le type d'effet de la liaison dépend du type cellulaire où celle-ci se fait, montrant que les effets des ADCY sont spécifiques aux tissus. Par exemple, si la liaison se fait sur les types hépatiques et adipeux, il va y avoir une libération de glucose et d'acide gras. Si elle se fait sur les muscles cardiaques, il va y avoir une augmentation du taux de contraction, augmentant ainsi l'apport sanguin aux tissus [92]. Toujours chez les mammifères, si l'épinéphrine se lie aux récepteurs α-adrénergiques, il y aura l'activation des G_i ou G_q. Si c'est l'activation des G_i, il y aura la diminution de la production d'AMPc. Si celle-ci se fait sur des artères, elle va se contracter et donc couper la circulation vers les organes périphériques [92]. Même si les effets sont différents, l'activation de cette famille a un but commun, soit fournir l'énergie

pour le mouvement rapide des principaux muscles pour répondre efficacement au stress de la situation [92].

Les phénotypes associés à ADCY9. Plusieurs associations avec des phénotypes ont été identifiées pour le gène *ADCY9*. Un premier exemple est une association avec l'asthme [93, 94], qui est une maladie respiratoire qui implique l'obstruction des voies respiratoires. Il y a trois mécanismes qui causent cette maladie, soit l'inflammation de la paroi interne des bronches, la contraction des fibres musculaires qui entourent les bronches et la production de surplus de mucus qui bloque les bronches. Pour contrer ces symptômes, plusieurs médicaments ont été conçus. Une des approches cible les récepteurs β_2-adrénergiques, puisqu'ils sont impliqués dans la régulation des voies respiratoires au niveau des muscles lisses des poumons. Lors de l'activation d'ADCY9 par ces récepteurs, il y a la production d'AMPc, qui active les protéines kinases A (PKA), qui vont phosphoryler les protéines associées aux muscles des voies respiratoires, entraînant la bronchodilatation. Des variants génétiques dans le gène *ADCY9* ont été trouvés comme modulant la réponse thérapeutique de cette approche [93], telle que la modulation du volume expiratoire forcé en une seconde (FEV1, *Forced Expiratory Volume in the first second* en anglais) [94]. Le gène *ADCY9* a également été associé avec les niveaux du chimiotactisme des neutrophiles, puisque son absence entraîne une diminution de la migration de ces cellules [95, 96], qui sont également associées avec la sévérité de l'asthme [97]. Cela montrerait qu'*ADCY9* aurait des fonctions importantes pour les capacités pulmonaires.

Un autre phénotype suggéré comme ayant une association avec *ADCY9* est la réponse immunitaire à une infection de la malaria. La malaria, ou le *Paludisme*, est une maladie potentiellement mortelle dans plusieurs régions tropicales, comme en Afrique. Selon le rapport de 2022 de l'Organisation Mondiale de la Santé (OMS) [98], il y a plus de 100 pays et territoires qui ont un risque important de transmission de la malaria, principalement les pays chauds. Elle est causée par un parasite protozoaire de la famille des *Plasmodium*, dont cinq espèces causent la maladie chez l'humain. Dans des études vérifiant l'association de la sévérité de la maladie avec la génétique humaine, des variants introniques dans le génome de l'hôte ont été potentiellement associés avec des symptômes de la malaria. Une des mutations est la mutation rs2230739 dans le gène *ADCY9*. Cette mutation a été associée au risque d'acidose durant la malaria [99], ainsi que la régulation de la transcription du gène

$TGF-\beta$, qui est impliqué dans la régulation des réponses immunitaires [100]. Une autre mutation, soit rs10775349 dans *ADCY9*, a été associée à la protection contre l'hyperpyrexie, soit une augmentation extrême de la température corporelle, généré pendant l'infection à la malaria [100]. Ces deux associations montre qu'*ADCY9* joue un rôle lors de la réponse du système immunitaire.

1.1.3. Analyses phénotypiques

Depuis le siècle dernier, la génétique est passée du simple concept de gènes à une cartographie génétique complète des organismes, en passant par la découverte des effets de mutations sur l'organisme. La connaissance des effets qu'une mutation d'un gène a sur des phénotypes a aidé à comprendre les fonctions de ces gènes et donc d'associer les mutations à des maladies.

Les avancées dans la génomique ont permis des progrès significatifs dans la médecine de précision, qui repose sur une meilleure compréhension de la variabilité génétique interindividuelle. Plusieurs types d'analyses peuvent être effectués afin d'étudier la relation entre la génétique et un phénotype.

On peut définir ici deux types de maladies: les maladies mendéliennes et les maladies complexes. Dans le cas de maladies mendéliennes, une seule mutation est nécessaire afin de permettre le développement de la maladie. Cependant, dans le cas des maladies complexes, plusieurs mutations peuvent permettre d'augmenter les chances de la développer, mais des facteurs environnementaux peuvent aussi grandement l'influencer. Dans certain cas, l'influence environnementale a même plus d'effet que l'influence génétique [101].

Ronald Fisher, un des pères de la génétique des populations, a développé le domaine de la génétique quantitative qui dit que la variance phénotypique dans une population est due à un grand nombre de facteurs génétiques, donc que c'est leur contribution additive qui forme le phénotype observable [102]. Il a fait usage de la dominance et de l'épistasie (voir section 1.4), mais il les a présenté comme étant des paramètres nuisibles pour expliquer les anomalies au-lieu de concepts génétiques d'intérêt. Il a émis une série d'hypothèses au sujet de la structure des populations, du nombre de gènes associé au phénotype, et du fait que les gènes travaillent ensemble pour générer la relation entre les génotypes et le phénotype [103]. Une des hypothèses est qu'il y a beaucoup de gènes associés à un phénotype, mais que leur

contribution est additive et faible. Dans le cas théorique d'une population dont la taille tend vers l'infini, un phénotype continu a une distribution normale et peut être estimé à l'aide d'un modèle additif (Équation 1.1.1).

$$E(Phenotype) = \alpha + \sum_{i=1}^{n} \beta_i G_i + \epsilon \tag{1.1.1}$$

Dans cette équation, il y a n loci, pour lesquels l'allèle G_i (qui peut être 0,1,2 selon son nombre d'allèle alternatif au locus i présent chez l'individu) a une contribution pondérée par β_i avec une valeur phénotypique de base α, et avec une erreur marginale pour les sources d'erreur externe ϵ. La contribution d'une mutation sur un trait peut être estimée grâce à des tests d'association phénotypique.

1.1.3.1. Test d'association phénotypique

Afin d'identifier les mutations génétiques associées à un phénotype, tel qu'une maladie, une première approche serait une étude d'association pangénomique (GWAS, *Genome-Wide Association Study* en anglais) [104]. Les bases de données biologiques, ou biobanques, sont fréquemment utilisées pour ce type d'analyse. En effet, ces biobanques contiennent des données recueillies sur un grand nombre d'individus, comprenant à la fois des informations génétiques et phénotypiques, ce qui permet d'identifier des associations. Cette association peut soit se faire en utilisant la formule 1.1.1 pour les traits continus ou en comparant les fréquences entre le groupe des cas et des contrôles pour les phénotypes binaires. L'avantage majeure du GWAS est qu'il permet de faire une étude sans hypothèse initiale en recherchant sur tout le génome et de trouver les régions associées au phénotype d'intérêt [105].

Valeur significative p. L'association entre la région et le phénotype est dite significative si sa valeur d'association p avec le phénotype se trouve inférieur à un seuil significatif α, souvent établi à 0.05. Cette valeur indique que la probabilité que l'association soit due au hasard est de 5%. Cependant, lors d'un GWAS, un grand nombre de variants sont évalués, ce qui augmente la probabilité d'avoir des faux positifs, soit des erreurs de type I, avec le seuil $\alpha = 0.05$. Une des méthodes les plus utilisées pour éviter ce problème est d'utiliser la méthode de correction de *Bonferroni* qui est de diviser α par le nombre de tests indépendants à effectuer. Si on a k tests, alors notre seuil significatif sera de $p = \alpha/k$ [106]. Cependant, ce test est très conservateur, ce qui va augmenter les faux-négatifs, donc les erreurs de type

II. Plusieurs alternatives ont été proposées, tels que l'approche du taux de fausse découverte (FDR, *False Discovery Rate* en anglais) ou les tests de permutation. Durant les GWAS, le seuil est traditionnellement mis à $p = 10^{-8}$, en tenant compte de la non-indépendance entre les sites à cause du déséquilibre de liaison, ce qui diminue le nombre de test indépendant [104].

Covariable. Des facteurs autres que la génétique peuvent influencer un phénotype, tels que le sexe, l'âge et les habitudes de vie des individus, qui sont représentés par la valeur ϵ dans l'équation 1.1.1. C'est pourquoi il est important de les considérer comme covariables dans les analyses. La structure populationnelle peut également avoir des impacts importants sur les résultats obtenus lors de ces analyses. En effet, la fréquence allélique et les phénotypes peuvent différer entre les populations (voir section 1.2.1), ce qui peut générer des fausses associations entre une mutation et un phénotype, alors que ce n'est dû qu'à la différence entre les populations. Il est possible d'inférer les origines ethniques en étudiant la similarité des génomes de chaque individu par rapport à un échantillon de référence à l'aide des analyses de composantes principales (PCA, *Principal Component Analysis* en anglais) [107]. Cette méthode va diminuer le nombre de variables, ici les positions génétiques, en estimant leur variance observée entre les données génétiques. Même si la cohorte n'est échantillonnée que dans une seule population, il est quand même important de considérer les composantes principales car il y aura quand même de la structure fine entre les individus.

Étude d'Association Panphénotypes. Lorsqu'un variant a été identifié dans des études d'association telles que les GWAS ou dans le contexte de la détection de signature de sélection, il serait intéressant de déterminer les phénotypes auxquels il est associé. Pour ce faire, il existe une technique complémentaire au GWAS : une étude d'association panphénotypes (PheWAS, *Phenome-Wide Association Study* en anglais). Cette technique permet d'évaluer l'association d'une mutation avec chaque phénotype auquel elle est liée [108]. Cela permet d'identifier les propriétés pléiotropiques des variants d'intérêt (voir section 1.1.3.3), c'est-à-dire l'association d'un variant avec plusieurs phénotypes, et de documenter éventuellement le rôle métabolique des gènes qui contiennent ces variants [109].

1.1.3.2. Randomisation mendélienne

Suite à l'identification d'association entre une mutation et un phénotype, plusieurs questions se posent : Par quel moyen cette mutation influence le phénotype? Est-ce par l'expression du gène? Par la modulation d'un autre phénotype?

L'analyse de randomisation mendélienne (MR, *Mendelian Randomization* en anglais) permet de répondre à ces questions en établissant une relation causale entre des variables d'intérêt [110]. Cette approche est prometteuse pour prédire l'efficacité des interventions thérapeutiques et éclairer les voies métaboliques associées aux maladies. Par exemple, les analyses MR ont permis de trouver des relations causales des maladies cardiovasculaires avec le profile lipidique sanguin [111, 112], la consommation d'alcool [113], l'adiposité [114] et plusieurs autres phénotypes.

Afin d'étudier les effets d'une mutation génétique sur un phénotype, l'utilisation des biobanques est essentielle. Cependant, une biobanque n'a pas nécessairement toutes les données d'intérêt. Par exemple, la base de données du Royaume-Uni [115] (*UK biobank* en anglais, section 1.5.3.3) contient une grande quantité de phénotypes, mais ne contient pas de données d'expression. Dans le cas des bases de données d'expression, telles que GTEx [116] (section 1.5.3.4), le nombre d'individu est limité, souvent moins de 1000, et le nombre de phénotypes est souvent restreint. Même si on veut regarder une association entre un phénotype et l'expression, dû au faible nombre d'individus, la puissance statistique ne sera pas nécessairement présente. Cependant, ces bases de données ont une variable en commun, soit la génétique.

Afin de combler cette limitation, l'approche de randomisation mendélienne sur deux cohortes a été développée [117]. Cette approche utilise les résultats d'association avec les variables dans les deux bases de données. En comparant les effets des mêmes mutations génétiques sur les deux variables, il est possible de vérifier s'il existe une corrélation dans leur comportement.

Par exemple, pour évaluer l'impact d'une augmentation de notre exposition, telle que l'expression d'un gène, sur une issue, telle que la taille des individus, on peut estimer les effets (β de l'équation 1.1.1) des mutations génétiques sur ces deux variables à partir des données de leur biobanque respective. Ensuite, on peut comparer la corrélation entre les estimations de ces deux variables. Si les mutations qui présentent les niveaux d'expression

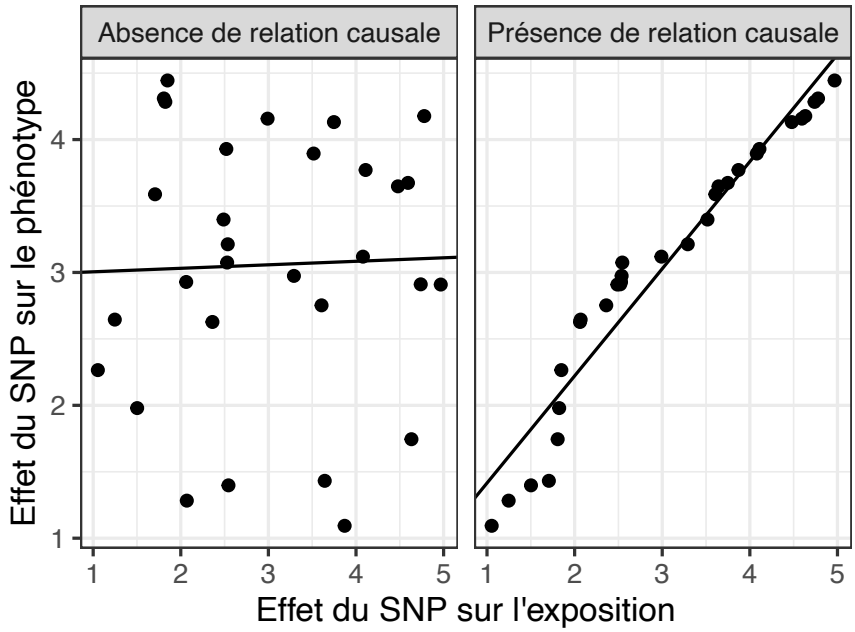

Fig. 1.4. Exemple de relation causale en randomisation mendélienne
Les graphiques représentent l'effet d'une donnée, obtenue de manière aléatoire dans cet exemple, sur l'exposition (axe des X) et le phénotype étudié (axe des Y). La ligne noire représente la corrélation entre ces deux variables. En absence de relation causale, il n'y a pas de corrélation entre les effets des variables instrumentales, comme des variants génétiques, sur l'exposition et leurs effets sur le phénotype. En présence de relation causale, il est possible d'observer une corrélation entre les effets de ces variables instrumentales sur les deux variables.

les plus élevés sont les mêmes que celles associées aux tailles les plus grandes, cela indique une corrélation entre l'expression du gène et la taille des individus (Figure 1.4).

Cependant, plusieurs suppositions doivent être remplies [110] afin de pouvoir effectuer ces analyses que l'on peut résumer en trois principales suppositions (Figure 1.5).

Première supposition. Les variables instrumentales, soit les mutations génétiques, doivent être associées avec l'exposition. Cela peut être validé en filtrant sur les valeurs p et/ou la statistique F-dérivée. Cette dernière statistique regarde la force d'association entre le variant et la variable étudiée. En général, un seuil de 10 est considéré comme fortement associé, ce qui va limiter les biais de faibles instruments. Si cette supposition n'est pas valide, cela créera un biais de faibles instruments qui faussera les estimés de corrélation.

Deuxième supposition. Il n'existe pas de confondant non-mesuré lors de l'association génétique entre le variant génétique et l'issue d'intérêt. Une variable confondante est une troisième variable non-mesurée qui influence à la fois l'exposition et l'issue. Par exemple,

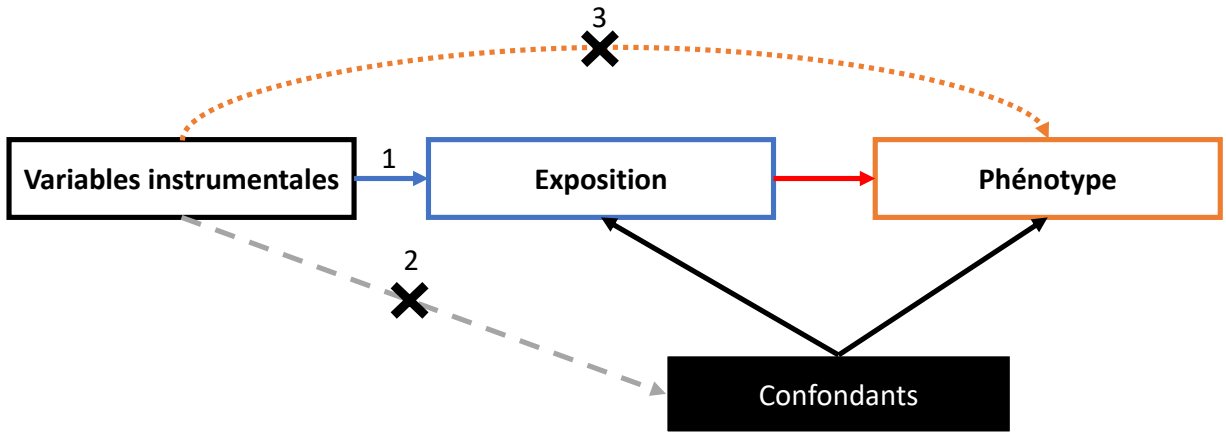

Fig. 1.5. Représentation des trois suppositions de la randomisation mendélienne (1) Les variants génétiques (Variables instrumentales) doivent être associés avec l'exposition. (2) Les variants génétiques ne doivent pas être associations avec des confondants. (3) Les variants génétiques doivent influencer le phénotype par le biais de l'exposition seulement et pas via un autre chemin. La flèche rouge de l'exposition vers le phénotype représente l'association causale mesurée par l'analyse MR dans le cas que les assomptions sont valides.

on veut étudier si la consommation de crème glacée influence le nombre de coup de soleil. Une analyse permettra d'observer une augmentation significative du nombre de coup de soleil quand la consommation de crème glacée augmente, et donc, la conclusion serait qu'il existe une relation. Cependant, une variable confondante qui n'a pas été tenue compte dans cette analyse est la température. Une augmentation de la température augmentera la consommation de crème glacée car les personnes voudront se refroidir, mais également le nombre de coup de soleil car le soleil sera plus présent. Si on considère une température stable, la relation entre ces deux variables ne sera probablement pas significative. Durant les analyses génétiques, il existe plusieurs variables confondantes à tenir en compte, telles que l'ethnicité. Afin d'éviter de créer de fausse association, il est donc important que les effets de ces variables confondantes soient minimisées, par exemple, en les ajoutant comme covariable pour les analyses d'association initiale. L'effet de l'ethnicité peut être corrigé avec la méthode des composantes principales (PCA), mais ces valeurs sont spécifiques à chaque base de données. Il faudra donc que les statistiques sommaires calculées dans les deux bases de données soient faites sur des populations similaires.

Troisième supposition. L'association entre le variant et l'issue d'intérêt doit être entièrement médiée par l'exposition étudiée, sans présence de pléiotropie (voir section 1.1.3.3).

Cela signifie que l'effet du variant sur l'issue d'intérêt ne peut pas être influencé par d'autres mécanismes ou voies biologiques indépendantes de l'exposition étudiée.

La relation entre une exposition et l'issue d'intérêt peut être directe ou modulée par une variable médiatrice, c'est-à-dire une variable qui est également influencée par l'exposition et qui affecte également l'issue. Il y a donc trois paramètres typiques qui peuvent être estimés dans les analyses MR [118]. Tout d'abord, il y a l'effet total, qui représente l'effet de l'exposition sur l'issue d'intérêt via tous les chemins possibles. Lorsque l'on tient compte de la médiation, il y a l'effet direct, qui correspond à l'effet de l'exposition directement sur l'issue, et l'effet indirect, qui se produit via des chemins de médiation.

1.1.3.3. Pléiotropie

L'effet d'une mutation ne va pas toujours se limiter à un seul phénotype. Ce phénomène s'appelle de la pléiotropie. Cela peut se faire par plusieurs mécanismes d'action. Si un gène est impliqué dans plusieurs voies signalétiques, alors les mutations influençant le niveaux d'expression ou les fonctions de ce gène vont également influencer ces voies signalétiques et donc influencer différents phénotypes [119]. Il existe deux types de pléiotropie, soit la pléiotropie verticale et la pléiotropie horizontale.

Le premier type est quand un locus a des effets sur deux phénotypes, mais ceux-ci se trouvent sur la même voie signalétique. Par exemple, on trouve une association causale entre l'expression du gène X et du phénotype Y. Cependant, l'association entre X et Y n'est pas directe, mais elle se ferait par l'intermédiaire de la variable W. Ce type de pléiotropie ne va pas affecter les résultats des analyses de randomisation mendélienne, puisque c'est l'essence de cette approche.

Le deuxième type est quand un locus va affecter deux voies signalétiques distinctives. Par exemple, on trouve une association causale entre l'expression du gène X et du phénotype Y. Cependant, ce locus, ou un deuxième locus en déséquilibre de liaison, pourrait également affecter le gène Z et ce serait ce gène qui, en réalité, serait associé avec le phénotype Y. C'est ce type de pléiotropie qui va entraîner une violation de la troisième supposition et peut fausser les associations causales [120].

1.1.3.4. Statistiques

Il existe plusieurs méthodes statistiques pour évaluer la causalité entre deux variables avec des analyses de randomisation mendélienne.

La méthode principale est une méthode pondérée par l'inverse de la variance (IVW, *Inverse-Variance Weighted* en anglais)[121] qui est la plus efficace tant que les variants génétiques sont valides et indépendants ou que la pléiotropie est balancée à travers les variants. Cette méthode donne l'estimé d'un ratio obtenu en divisant l'effet d'une mutation par l'effet de cette mutation sur l'exposition. Les effets individuels sont ensuite pondérés par l'inverse de leur variance, puis les estimés de toutes les variables instrumentales sont combinées. Cela fait en sorte que les instruments qui sont plus précis vont avoir plus de poids dans l'estimé final [121]. C'est la première méthode pour assumer qu'il existe une relation causale, tant que les suppositions sont respectées. Il existe plusieurs méthodes pour vérifier cela.

La méthode de MR-Egger [122] permet de vérifier la troisième supposition en tenant compte de la pléiotropie horizontale. Son approche ressemble à celle d'IVW, mais elle ne force pas son ordonnée à l'origine à 0, ce qui permet d'avoir un estimé de l'effet moyen de la pléiotropie horizontale via la valeur de l'ordonnée. Cependant, afin de tenir compte de la pléiotropie, sa puissance statistique est diminuée par rapport à que celle d'IVW. Une autre limitation est qu'elle est sensible aux valeurs aberrantes qui peuvent influencer son ordonnée.

Une autre méthode est celle de la médiane pondérée [123] qui donne un estimé causal consistant à celle d'IVW, tant qu'au moins 50% des instruments génétiques utilisées sont valides. Elle est robuste aux valeurs aberrantes, mais sensible à l'ajout et la suppression de variants.

Initialement, les statistiques de randomisation mendélienne permettent d'estimer l'effet total d'une exposition sur l'issue. Cependant, il peut y avoir de la pléiotropie ou l'effet de l'exposition peut passer par des variables médiatrices. Afin d'établir le chemin entre l'exposition et l'issue, il est possible d'estimer ou de contrôler pour les variables pléiotropiques ou médiatrices. Les analyses de randomisation mendélienne multivariable (MVMR, *Multi-Variable Mendelian Randomisation* en anglais)[124] sont utilisées pour tenir compte de ces divers chemins. Au lieu d'estimer uniquement l'effet de l'exposition primaire sur l'issue, les autres variables potentiellement alternatives sont incluses dans le modèle pour réduire les effets pléiotropiques. Les instruments pour l'exposition primaire et l'exposition médiatrice

seront donc inclus dans l'analyse. Cette approche permet d'obtenir l'effet direct de l'exposition sur l'issue en contrôlant les effets des autres expositions. Cependant, il est supposé qu'il n'y a pas pas d'interaction entre l'exposition et le médiateur [118].

Après avoir identifié des associations entre une mutation et un phénotype, nous pouvons caractériser cette mutation avec des analyses de génétique des populations et d'analyses de transcriptomique.

1.2. Génétique des populations

1.2.1. Définitions et notions de base

La génétique des populations se concentre sur l'étude des variations génétiques au sein des populations humaines. En comprenant comment la fréquence de certaines mutations varie dans des environnements spécifiques, nous pouvons mieux appréhender l'impact de ces variations génétiques sur la santé et la maladie. Cette approche nous permet également d'explorer les principes de la médecine évolutionniste, ou médecine darwinienne, qui met l'accent sur l'influence de la sélection naturelle et des pressions environnementales sur l'évolution de la santé humaine. En combinant la génétique des populations et la médecine évolutionniste, nous pouvons identifier les fonctions spécifiques des mutations et comprendre comment elles contribuent à l'adaptation des populations à leur environnement. Ainsi, la génétique des populations et la médecine évolutionniste s'ajoutent à la médecine de précision en nous permettant d'utiliser les aspects évolutifs de la génétique médicale pour mieux comprendre les bases génétiques des différences de santé observées entre les populations.

1.2.1.1. Principe d'Hardy-Weinberg

Afin d'évaluer les forces qui maintiennent en équilibre les différentes populations, plusieurs concepts ont été établis. Un de ces principes est celui de l'équilibre d'Hardy-Weinberg, calculé avec la formule 1.2.1. Ce principe stipule qu'un équilibre s'établit de génération en génération entre les fréquences alléliques (p et q) et génotypiques, et que la population sera en équilibre d'Hardy-Weinberg si elle respecte la formule 1.2.1.

$$p^2 + 2pq + q^2 = 1 \tag{1.2.1}$$

Afin d'appliquer le principe d'Hardy-Weinberg, la population doit être de taille infinie, diploïde, de reproduction sexuée, sans consanguinité et qu'il n'y ait ni migration, ni mutation, ni sélection naturelle [125]. Le non respect de ces conditions peut faire dévier la population de l'équilibre.

1.2.1.2. Mutations

Les mutations génétiques réfèrent à un changement dans la séquence génétique. Elles sont responsables de la diversité dans les organismes et peuvent se produire à différents stades de la vie et dans différents types cellulaires, potentiellement en réponse à des facteurs environnementaux, ou être héritées des parents. On peut distinguer deux types de mutations. Le changement d'une seule paire de base est appelé un polymorphisme d'un seul nucléotide (SNP, *Single Nucleotide Polymorphism* en anglais). Par ailleurs, les disparitions ou les ajouts de nucléotides sont appelés des insertion-délétions (indel). Les mutations sont dites neutres, bénéfiques ou délétères selon la répercussion sur des traits phénotypiques (voir section 1.2.1.5).

1.2.1.3. Migration et dérive génétique

La migration entre les populations et la dérive génétique sont des facteurs qui peuvent influencer les fréquences alléliques entre les populations et perturber l'équilibre d'Hardy-Weinberg.

L'échange d'individus entre les populations, soit la migration, peut entraîner un enrichissement des homozygotes sans changement de la fréquence des hétérozygotes lors de la première génération [126].

La dérive génétique, quant à elle, est causée par des événements aléatoires et peut amener un allèle jusqu'à un état de fixation ou de disparition dans une population, surtout dans les petites populations [127, 128].

Ces phénomènes peuvent affecter n'importe quelle position génétique dans le génome et sont des mécanismes de base de l'évolution.

1.2.1.4. Déséquilibre de liaison

Le déséquilibre de liaison (LD, *Linkage Disequilibrium* en anglais) est l'association entre deux mutations, autrement dit, la probabilité que deux mutations soient transmises ensemble à la prochaine génération. Durant la méiose, des événements de recombinaison entre les

chromosomes parentaux peuvent éliminer cette association. Certaines régions génomiques sont plus propices à des événements de recombinaison et sont appelées des points chauds de recombinaison (ou *hotspots*) [129]. En général, plus deux mutations sont éloignées, plus fortes seront les chances d'événement de recombinaison entre les deux positions et elles auront donc un LD moins fort. Trois statistiques peuvent être utilisées pour mesurer le LD, soit D, D' et r^2.

- La valeur de D explique de combien la fréquence observée des haplotypes diffère de la valeur attendue
- La valeur D' est une valeur D ajustée pour la fréquence des allèles
- La valeur r^2 montre la force de la corrélation entre les deux SNPs

Supposons que nous avons deux loci avec deux allèles pour chaque locus, soit A,a et B,b. La fréquence de l'haplotype composé de A (P_A) et B (P_B) observée est de P_{AB}. Dans ce cas, la valeur de D est obtenue par l'équation $D = P_{AB} - (P_A \times P_B) = P_{Ab} - (P_A \times P_b) = P_{aB} - (P_a \times P_B) = P_{ab} - (P_a \times P_b)$ [130]. Si les loci sont indépendants, donc sans LD, leur valeur sera proche de 0, puisque la fréquence de la combinaison est égale à la multiplication des fréquences des deux allèles séparés. Cependant, cela n'indique pas la force de l'association, car l'amplitude est influencée par la fréquence des allèles. La valeur D' a été développée afin de corriger cet effet de fréquence. Quand la valeur D est positive, l'équation de D' est $D' = \frac{D}{D_{max}}$, où D_{max} est la valeur minimale entre $P_A \times P_b$ et $P_a \times P_B$. Dans le cas contraire, l'équation devient $D' = \frac{D}{D_{min}}$, où D_{min} est la valeur maximale entre $-P_A \times P_B$ et $-P_a \times P_b$. Les valeurs de D' vont donc varier entre -1 et 1. Plus la valeur absolue de D' s'approche de 1, plus cela signifie qu'il n'y a pas eu beaucoup de recombinaison entre les deux SNP. Cependant, l'interprétation des données qui ne sont pas de -1, 0 ou 1 n'est pas évidente, car leur amplitude n'indique pas si l'association est statistiquement significative.

La dernière mesure est le r^2 (Formule 1.2.2).

$$r^2 = \frac{D^2}{P_A \times P_a \times P_b \times P_B} \tag{1.2.2}$$

La plus grande valeur possible est de 1. L'obtention de cette valeur signifie que les deux mutations sont toujours transmises ensemble, donc ont un déséquilibre de liaison "parfait". Cela peut seulement arriver quand les fréquences des allèles mineures sont les mêmes [130].

1.2.1.5. Sélection naturelle

Le génome est un manuel d'instructions pour la production des protéines. Cependant, l'apparition des mutations peut changer ces instructions, ce qui aura un impact sur l'organisme. Tel que mentionné plus tôt, une mutation peut avoir un effet bénéfique, délétère ou neutre sur les traits d'un individu dans son environnement. C'est Charles Darwin qui a proposé, avec sa théorie de l'*Évolution des espèces*, que si une mutation est bénéfique, alors elle sera gardée dans la population [131]. La valeur adaptative (*fitness* en anglais) est un concept central de la génétique des populations, puisqu'elle permet de décrire la capacité d'un individu avec un certain génotype d'avoir une descendance. La valeur adaptative en équilibre est de 1, c'est-à-dire qu'il ne va pas y avoir un génotype qui produira plus de descendances qu'un autre.

Un écart à la valeur adaptative de 1 est appelé un coefficient de sélection (s). On attribue la valeur adaptative de 1 à un des génotypes, auquel on compare les deux autres génotypes qui auront le coefficient de sélection s. Un coefficient de sélection positif signifie que le génotype donne un avantage sur le succès de reproduction et/ou de survie comparé au génotype de référence, tandis qu'un coefficient négatif signifie que le génotype donne un désavantage.

Trois types de sélection sont communément distingués : la sélection positive, négative et balancée. Dans certaines circonstances, ces types de sélection peuvent avoir d'autres nominations, telles que la sélection sexuellement antagonistique (voir section 1.2.1.6) et la co-évolution génétique (voir section 1.4.3). Chacun d'entre eux laisse une marque distinctive dans le génome, ce qui permet leur identification grâce à des tests statistiques tels que ceux présentés dans les sous-sections suivantes (Sections 1.2.2 et 1.2.3).

Sélection positive. La sélection positive survient lorsque la nouvelle mutation donne un avantage sélectif, ce qui mènera à un coefficient de sélection positif. Sa fréquence allélique augmentera, possiblement, jusqu'à fixation (fréquence allélique de 100%) [132]. Les porteurs auront un avantage dans leur environnement, ils auront donc plus de chance de survivre et de se reproduire. Par exemple, une mutation qui permet d'augmenter l'efficacité des échanges gazeux lors de la respiration sera avantageuse dans les environnements qui ont une pression partielle d'oxygène réduite, comme celles situées en haute altitude, mais procurera peu d'avantage dans les autres environnements. La fréquence de cet allèle augmentera donc dans les populations vivant dans des environnements situés en haute altitude.

Lorsqu'une mutation est sujette à la sélection positive, il est courant qu'elle soit en LD avec d'autres variants génétiques situés à proximité. Par conséquent, la sélection exercée sur cette mutation peut également avoir un impact sur les variants en LD avec elle, ce qui crée une signature génétique identifiable à l'aide de statistiques telles que le D de Tajima et l'iHS [128, 133]. La vitesse d'augmentation de la fréquence de la mutation sous sélection dépend de la force de sélection. Plus la force est élevée, plus la fréquence de la mutation augmentera rapidement. Cependant, cela peut également dépendre du degré de dominance associé à l'allèle bénéfique. Un faible degré de dominance indique que la force de sélection n'agit pas fortement sur les individus hétérozygotes, ce qui ralentit la fixation de la mutation [128].

Sélection négative. La sélection négative est le contraire de la sélection positive. La mutation entraîne un désavantage de l'individu dans l'environnement et tend à disparaître [128]. Un exemple pour cela serait une mutation associée à une maladie mortelle qui survient en bas âge. Dans ce cas, les porteurs vivront moins longtemps et auront moins de temps pour se reproduire. Ils auront moins de descendants, ce qui mènera à une diminution de la fréquence de l'allèle délétère. Cela entraînera également une diminution dans la diversité génétique dans la région en LD, également identifiable avec la statistique du D de Tajima.

Sélection balancée. Une sélection balancée se produit lorsqu'il y a un avantage sélectif pour deux allèles différents pour une même mutation ou pour l'état hétérozygote. Dans ce cas, le polymorphisme est maintenu et la diversité génétique autour de la mutation est augmentée, ce qui empêche la fixation d'un seul allèle. Les allèles soumis à une sélection balancée atteignent un équilibre dans leur fréquence en fonction de la force de sélection exercée sur eux [128].

1.2.1.6. Sélection sexuellement antagoniste

La sélection sexuellement antagoniste (SA, *Sexually Antagonistic Selection* en anglais) est un type particulier de pression de sélection qui survient lorsque les traits sont partagés entre les sexes, mais que la sélection agit différemment sur ces traits selon le sexe. Ce phénomène se produit lorsque les allèles sont bénéfiques pour un sexe et délétères pour l'autre, et est observé dans les populations où la reproduction est sexuelle et les variants génétiques se maintiennent grâce aux traits sexuellement dimorphiques [134].

L'avenir d'une mutation SA dépend de sa valeur de sélection dans chaque sexe et de sa localisation. Si la mutation est délétère pour les femelles mais bénéfique pour les mâles et se trouve sur un chromosome autosomal, l'allèle avec la valeur adaptative la plus élevée augmentera en fréquence. En revanche, si la mutation est située sur le chromosome X, l'effet bénéfique de la mutation sera automatiquement exprimé chez les mâles, tandis que l'effet délétère sera présent chez les femelles homozygotes, mais variable chez les femelles hétérozygotes en raison de l'inactivation d'un chromosome X [135]. Cependant, même si l'effet est plus fort chez les mâles, le chromosome X portant l'allèle bénéfique doit passer par les femelles, chez qui l'effet est délétère, et n'a qu'une probabilité de 50% d'être transmis au fils, ce qui permet un certain maintien de ces mutations dans la population [136, 137].

Outre le mécanisme passant par le chromosome X, des effets spécifiques aux sexes ont également été observés sur des loci de traits quantitatifs (QTL, *Quantitative Trait Loci* en anglais) autosomaux. Plusieurs mécanismes sont suggérés pour cette régulation spécifique au sexe, tels la présence d'un modificateur sur le chromosome sexuel qui va influencer le locus du trait [138], l'épissage alternatif du gène [139], l'empreinte parentale qui inhibe l'expression des allèles hérités du sexe opposé [140], et plusieurs autres mécanismes [138].

Les allèles SA ont différentes options pour se fixer dans la population, notamment lorsque l'effet avantageux dans un sexe est nettement supérieur à l'effet désavantageux dans l'autre sexe ou lorsque l'allèle a un effet de direction opposé entre les deux sexes. Toutefois, pour que cette deuxième option soit viable, dans le cas d'expression de gènes, chaque sexe devrait avoir acquis une région régulatrice distincte spécifique au sexe, faisant en sorte que la régulation diffère entre les sexes. Dans ce cas là, puisque deux régions doivent être coordonnées, la sélection agira plus lentement que pour les mutations SA agissant seule [136] et cela pourrait générer des événements de co-évolution (voir section 1.4.3).

1.2.2. Approches statistiques basées sur la différenciation populationnelle

Plusieurs méthodes statistiques ont été proposées pour détecter la présence de signature de sélection naturelle. Chaque approche s'appuie sur un type de statistique issue des données génétiques, tel que les fréquences alléliques, les déséquilibres de liaison et les statistiques démontrant les différences entre les populations [141].

Un type d'approche est celui qui se base sur les différences entre les populations. Deux populations qui ont divergé depuis peu devraient avoir des fréquences alléliques similaires. Il y a plusieurs raisons qui pourraient expliquer des différences dans la fréquences alléliques entre ces populations similaires, telles la pression de sélection et la dérive génétique. Par exemple, si l'environnement diffère entre ces deux populations, alors une mutation conférant un avantage évolutif pourrait entraîner une différence de fréquence entre les population. En effet, l'allèle fournissant un avantage dans un environnement se retrouvera en plus haute fréquence dans la population sous pression de sélection. Par exemple, la population tibétaine, qui vit en haute altitude, subit une pression sélective due à l'environnement avec un manque d'oxygène (hypoxie). Des signatures de sélection naturelle sont observées au niveau du gène *EPAS-1* (*Endothelial PAS domain-containing protein 1*) qui confère une meilleure valeur adaptative dans cet environnement hypoxique [142]. Les différences de fréquences alléliques entre populations peuvent être un indice sur la pression sélective dans une de ces populations. Cependant, les différences de fréquences peuvent également provenir de la dérive génétique, qui n'est pas liée à l'environnement, mais consiste en un processus aléatoire [143].

1.2.2.1. F_{ST}

Une des méthodes largement utilisée afin de comparer les populations est la statistique F_{ST} (Index de fixation) [143]. Elle permet de voir les différentiations génétiques entre des populations. Cette méthode regarde la variance des fréquences alléliques à l'intérieur d'une même population (θ_w) et celle entre deux populations (θ_b). Elle se base sur l'hypothèse suivante: les deux populations ont subit la même dérive génétique depuis leur divergence. Plusieurs versions de cette statistique existent [144, 145, 146], mais celle d'Hudson (Équation 1.2.3) [144, 147] est souvent choisie car elle est indépendante de la taille de la population, donc elle ne tend pas à surestimer la valeur de F_{ST} [144]. Voici la formule :

$$F_{ST} = 1 - \frac{\theta_w}{\theta_b} \tag{1.2.3}$$

La valeur F_{ST} varie entre 0 et 1. Une valeur proche de 0 signifie que les deux populations sont similaires du point de vue de la diversité génétique [143] ou qu'il existe une grande diversité à l'intérieur des populations étudiées. De plus, une valeur faible peut également indiquer la présence d'une sélection balancée ou d'une sélection positive dans les deux populations, menant à des fréquences similaires [148]. Une plus grande valeur de F_{ST}, par exemple

proche de 1, signifie qu'il y a une grande différence entre les deux populations ou qu'il y a peu de diversité au sein des populations étudiées, donc qu'il y a possiblement une pression sélective dans une des deux populations. La valeur basale attendue de F_{ST} varie selon les deux populations étudiées. Si les populations sont très éloignées, telles qu'une population de l'Asie avec une population d'Afrique, alors on s'attend à de hautes valeurs de F_{ST}. Par exemple, la population africaine Yoruba et la population asiatique du Japon ont une valeur F_{ST} moyenne de 0.19 [149]. Si les populations ont divergé depuis peu, alors de faibles valeurs sont attendues.

Cependant, cette méthode comporte plusieurs limites. Une d'entre elles est qu'elle n'indique pas quelle population est la plus diversifiée entre les deux, donc laquelle pourrait être sujette à la sélection [148, 150]. Afin d'identifier la population la plus diversifiée, la statistique des branches populationnelles (PBS) a été dérivée de la méthode F_{ST} (voir section 1.2.2.2).

Pour évaluer si la différence observée provient de sources autres que la pression de sélection, telles que les phénomènes démographiques ou de la dérive génétique, il est nécessaire de calculer les valeurs F_{ST} sur d'autres régions que celle d'intérêt. En effet, ces phénomènes influenceront tout le génome, tandis que la sélection ne touchera que la région concernée. En regardant d'autres régions ou même le génome entier, cela permettra d'établir une distribution nulle car on considérerait les autres régions comme un contrôle négatif [151]. Si les autres régions étudiées montrent également une grande diversité entre les deux populations, cela pourrait signifier que la différence observée ne serait pas d'origine de la pression de sélection. Si la région d'intérêt est un cas extrême comparé à la distribution nulle, alors cela pourrait être un indice qu'elle est sous pression de sélection.

1.2.2.2. Statistique des branches populationnelles

Une deuxième technique qui regarde la diversification entre les populations est la statistique des branches populationnelles (PBS) [150], dérivée de la statistique F_{ST}. Cette technique a été développée afin de pouvoir identifier la population qui présente le plus de différence et donc prédire quelle population de la comparaison aurait subi une pression sélective. Pour ce faire, une troisième population doit être incluse. Le choix des populations dépend de la population d'intérêt (P_A). La seconde population (P_B) doit être une population similaire à notre population d'intérêt, où la divergence est récente, potentiellement une

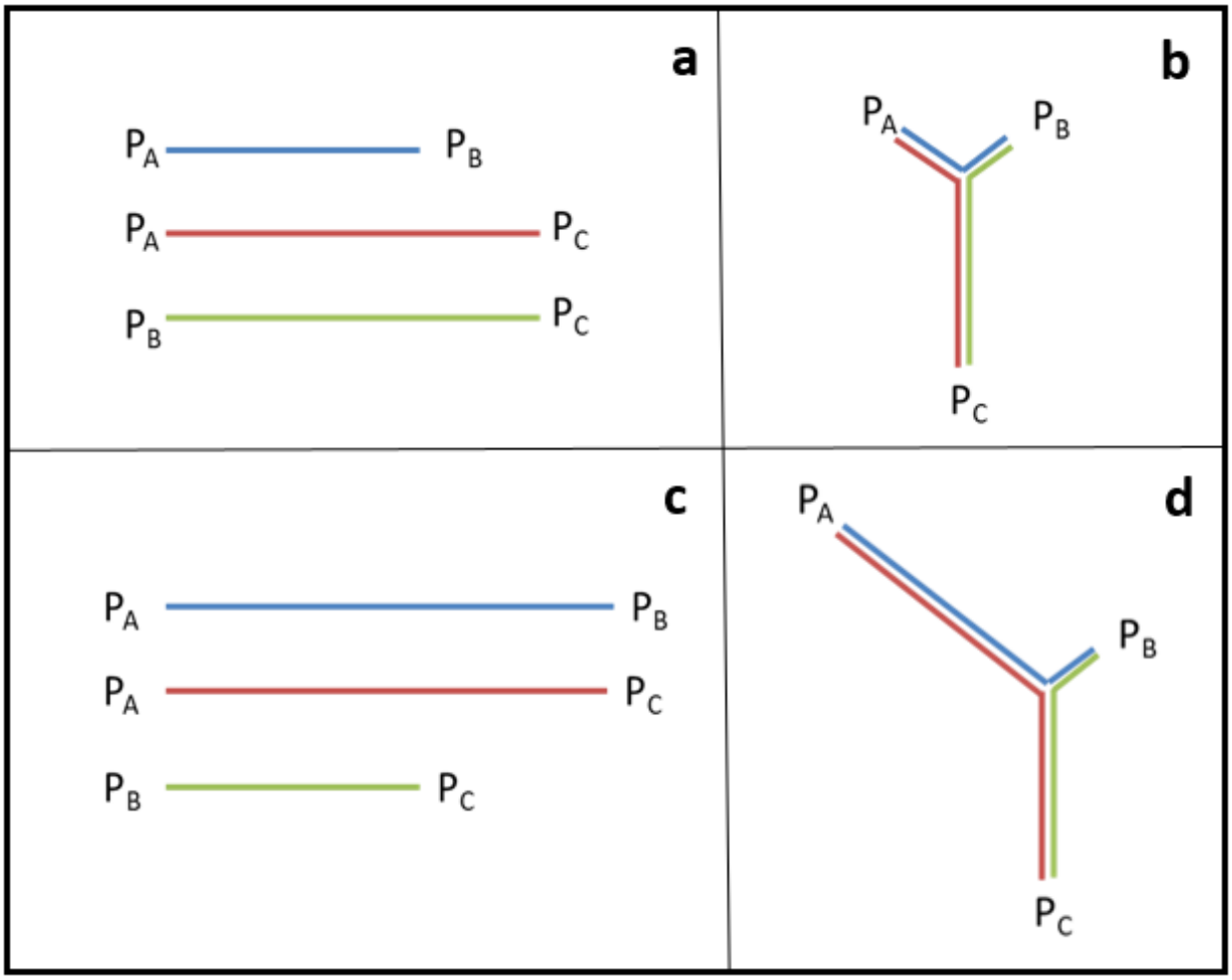

Fig. 1.6. Représentation des graphiques PBS sans et avec des pressions sélectives Représentation approximative de la statistique PBS quand il n'y a pas de pression de sélection (a,b) comparée à quand il y a une pression de sélection dans la population A (c,d). (a,c) Taille des branches entre les populations calculée avec l'équation 1.2.4, qui seront ensuite transformées en statistique PBS avec l'équation 1.2.5 afin de créer un arbre PBS (b,d). Population d'intérêt (P_A), Population similaire (P_B), Population mère (P_C).

population voisine géographiquement. La troisième (P_C) doit être une population mère dont la différence avec les deux premières population est similaire.

On calcule ensuite les statistiques de F_{ST} entre chaque paire de population. En mettant une troisième population, il est possible de voir celle qui se distingue des deux autres. Avec les valeurs F_{ST}, il est possible d'estimer le temps (T) depuis la divergence entre les deux populations (Équation 1.2.4).

$$T = -log(1 - F_{ST}) \tag{1.2.4}$$

Après avoir estimé le temps depuis la divergence, il sera possible de convertir en valeur *PBS* via la formule 1.2.5 et avoir une vue comparative globale des trois populations. La population étant la plus différenciée dans la région génomique étudiée aura une branche plus longue que les deux autres. Sans pression de sélection, on s'attendrait à ce que ce soit la population mère C qui présente les plus longues branches (Figure 1.6a,b). Cependant, lorsqu'il y a une pression de sélection sur la population A dans la région étudiée, sa branche sera plus longue que ce que l'on s'attendrait (Figure 1.6c,d).

$$PBS_A = \frac{T^{AB} + T^{AC} - T^{BC}}{2} \tag{1.2.5}$$

Ce test permet donc la détection d'un gène ou d'une région du gène sous sélection entre des populations vivant dans un environnement différent.

1.2.3. Approches par déséquilibre de liaison

1.2.3.1. iHS

Un allèle sous pression de sélection positive sera représenté dans la génération suivante plus souvent que ce qui est attendu. De plus, plus la pression de sélection est forte, plus son augmentation en fréquence de génération en génération sera rapide. Dû au déséquilibre de liaison avec les régions environnantes, l'effet de la pression évolutive se répercutera sur le bloc en LD avec l'allèle sous sélection. Avec la rapidité de la transmission, il y aura moins de recombinaisons qui se feront sur l'haplotype avec l'allèle favorable, ce qui génèrera de longs haplotypes autour de cet allèle [152, 153]. Dû à cette diminution de recombinaison, il y aura moins de diversité génétique sur les haplotypes avec l'allèle favorable [128, 133]. Cette théorie est à la base d'une statistique fréquemment utilisée, soit iHS (*integrated Haplotype Score*) [153], qui compare l'étendu des haplotypes entre l'allèle ancestral et l'allèle dérivé. Quand il y a une pression de sélection positive sur un allèle, les haplotypes centrés autour de sa position seront plus longs que ceux centrés autour de l'autre allèle, donnant une valeur positive quand la pression est sur l'allèle dérivé et négative pour l'allèle ancestral. Une valeur iHS absolue qui est supérieure à 2 correspond à une valeur p significative puisque, généralement, cela représente 5% des valeurs les plus grandes [153].

1.2.3.2. Déséquilibre de liaison sur longue distance

Deux mutations éloignées ont de forte chance d'avoir des événements de recombinaison entre elles, ce qui diminue leur LD. Cependant, certaines forces compensatoires peuvent créer des associations entre deux mutations [154], générant un déséquilibre de liaison sur longue distance (LRLD, *Long Range Linkage Disequilibrium* en anglais). Ces forces peuvent être de plusieurs origines.

Une première possibilité est la présence de métissage dans la population étudiée [154, 155, 156]. La plupart des populations ne proviennent pas d'une seule population uniforme, mais sont un mélange de plusieurs populations locales. Prenons deux populations de même taille et où les fréquences de deux mutations d'intérêt sont différentes, soit 0% pour les deux mutations dans une population et 100% pour la deuxième population. Lorsque calculé de manière individuelle sur chaque population, le déséquilibre de liaison est nul. Si on combine ces deux populations, les fréquences des mutations seront alors de 50%, sans la présence d'hétérozygote. Si on regarde le déséquilibre de liaison, celui-ci sera parfait, mais cela est dû seulement à cause d'un métissage de deux populations distinctes. Dans une population métissée récemment, le déséquilibre de liaison sera grand, mais diminuera au fil des générations grâce à la recombinaison. Il est possible de vérifier cette possibilité en regardant l'ancestralité des individus étudiés, tels qu'avec le logiciel de RFMix [157]. Ce logiciel permet d'inférer l'ancestralité de segments chromosomiques dans une population d'intérêt à partir d'haplotypes provenant de populations ancestrales. Il sera alors possible de vérifier si le LRLD observé entre deux mutations est causé par l'enrichissement de certaines ancestralités à ces loci.

Une deuxième possibilité est la dérive génétique (voir section 1.2.1.3), donc la présence de LRLD est due au hasard. De façon aléatoire, les deux mutations pourraient se retrouver transmises souvent ensemble, et donc, montreraient un déséquilibre de liaison.

Si l'origine de ce LRLD est causé par le métissage ou la dérive génétique, cela aura un impact sur l'ensemble du génome et pas seulement sur ces deux régions. Par conséquent, si l'on prend des paires de loci aléatoires qui présentent des similitudes avec la paire d'intérêt en termes de distance et de fréquence, et que l'on calcule leur LRLD pour créer une distribution nulle, la paire d'intérêt ne devrait pas présenter une association plus élevée par rapport aux autres paires de mutations aléatoires.

Une troisième possibilité est la co-évolution (voir section 1.4.3), souvent générée par de l'épistasie (voir section 1.4), c'est-à-dire d'une interaction non additive entre les deux gènes. Cette interaction peut maintenir le déséquilibre tant que la pression de sélection est présente [154, 155, 156, 158].

En général, la méthode de déséquilibre de liaison r^2 utilise les fréquences alléliques afin d'évaluer s'il y a un enrichissement d'une paire d'allèle. Une méthode alternative utilise à la place les fréquences génotypiques pour estimer le r^2 [159]. À la place de considérer les allèles sur le même haplotype, la statistique calcule la fréquence des combinaisons de génotypes. Étant donnée la distance, il sera difficile de savoir avec certitude si deux allèles sont sur le même haplotype, surtout s'ils sont séparés par le centromère, qui, jusqu'à récemment [160], n'avait pas été complètement séquencé. Dans le cas de pression de sélection, il est possible que l'appartenance à un même brin chromosomique n'ait pas d'importance. Chaque mutation aura son effet, les gènes produits seront en interaction causant la pression de sélection. L'utilisation des génotypes permet d'analyser si certaines combinaisons de génotypes sont plus fréquentes que ce qui serait attendu par hasard, ce qui pourrait indiquer la présence de forces de sélection qui maintiennent ces combinaisons associées, et donc une pression de co-évolution.

1.3. Expression génique

La relation entre le génotype et le phénotype a été un sujet de recherche depuis les expériences de Mendel. Comprendre cette relation est particulièrement importante en pharmacogénomique et en médecine de précision, où il est crucial de comprendre comment les variations génétiques peuvent influencer la réponse aux médicaments.

La transcription de ces gènes en ARN est un processus à plusieurs étapes contrôlé par des facteurs de transcription, des modifications épigénétiques et la structure de la chromatine. Après la transcription, l'ARN pré-messager contenant les introns doit subir des modifications post-transcriptionnelles pour être transporté et dégradé correctement. Finalement, l'ARN messager (ARNm) peut être traduit en protéines par des ribosomes pour produire des phénotypes spécifiques [161].

Le niveau d'expression des gènes est mesuré par la quantification de l'ARNm (voir section 1.3.3). L'étude du transcriptome permet d'identifier les gènes qui influencent un phénotype et

leur expression dans différents types cellulaires, permettant une meilleure compréhension du fonctionnement biologique des gènes et des relations entre eux [162, 163, 164]. Les mutations dans les exons et les introns peuvent perturber la production de protéines en modifiant l'acide aminé, la stabilité de l'ARNm et la régulation de l'épissage, ce qui peut entraîner une concentration de protéines variable ou des protéines défectueuses [165]. Il est donc important d'étudier comment une mutation affecte un gène pour comprendre le mécanisme qui mène au phénotype étudié.

1.3.1. Transcription

La régulation de la transcription chez les eucaryotes est très complexe et finement contrôlée. En plus des introns et des exons, la région du gène contient plusieurs autre éléments importants. La première région impliquée dans la transcription est la région promotrice, située en amont de l'extrémité 5' du gène, qui recrute la machinerie de transcription. La transcription débute par la liaison de l'enzyme d'ARN polymérase (ARN pol) à la région promotrice, avec l'aide de différentes protéines régulatrices de la transcription appelées facteurs de transcription (TF, *Transcription Factor* en anglais), qui se fixent sur le domaine C-terminal (CTD, *C-Terminal repeat Domain* en anglais) de l'ARN pol. Ce domaine C-terminal joue un rôle crucial en recrutant différents facteurs grâce à de la phosphorylation dynamique de ses résidus sérines [166].

Types d'ARN polymérase. Les différents types d'ARN pol synthétisent différents groupes de gènes. L'ARN pol I produit la majorité des ARN ribosomique (ARNr) nécessaires à la production des ribosomes. L'ARN pol II (Pol II) est responsable de la transcription des gènes codant pour des protéines en ARNm, ainsi que des molécules de micro-ARN. L'ARN pol III transcrit des gènes pour des petits ARN tels que les ARN de transfert et les ARNr 5S. Les plantes ont également l'ARN pol IV et l'ARN pol V qui produisent des petits ARN interférants afin de contrôler l'extinction de gènes [167, 168, 169].

1.3.1.1. Régulation de la transcriptomique

La transcription peut être régulée par plusieurs régions régulatrices séparées du gène. Ces régions régulatrices sont des éléments non-codants qui peuvent être catégorisées en trois types : les régions stimulatrices, répressives et isolatrices. Autres que les régions isolatrices, les autres régions régulatrices peuvent se trouver sur une région éloignée du promoteur.

Les régions stimulatrices sont de courtes séquences d'ADN qui contiennent des sites de liaison à des TF activateurs. Elles régulent ou activent l'initiation de la transcription en recrutant les TF nécessaire à l'assemblage de la machinerie de transcription à la région promotrice, soit via un contact direct ou via des signaux [170]. Les régions isolatrices, quant à elle, sont souvent trouvées entre les promoteurs et les régions stimulatrices, bloquant l'interaction entre ces régions. Elles peuvent également avoir une fonction de barrière en prévenant l'avancée de la chromatine condensée voisine et protèger l'expression des gènes des effets positifs ou négatifs de la chromatine [171]. Tout comme les stimulateurs, les répresseurs sont de courtes séquences d'ADN liant des TF, mais ceux-ci ont une activité inhibitrice. Lors de la liaison avec les TF, la transcription est bloquée. Il est important de noter que l'effet régulateur d'une séquence régulatrice n'est pas toujours répressif ou stimulatoire, mais dépend du contexte cellulaire [172].

Lors de la transcription, des modifications sont apportées aux ARN. Pour former un ARNm fonctionnel, trois étapes sont nécessaires : l'ajout d'une coiffe 5' et d'une queue poly-A en 3', ainsi que l'épissage pour enlever les introns et ne garder que les exons.

1.3.2. Épissage de l'ARN

Lors de la transcription du gène en ARN pré-messager, le produit contient à la fois des introns et des exons. Durant le processus de maturation de l'ARN, les introns seront enlevés et les exons seront joints. Ce processus se nomme l'épissage. Il y a deux principaux types d'épissage : l'épissage constitutif et l'épissage alternatif.

Épissage constitutif. L'épissage constitutif, ou l'épissage général, consiste à enlever les introns des ARN pré-messagers et à joindre les exons pour former un ARN mature. C'est le type d'épissage le plus courant dans les cellules, car les gènes avec plusieurs exons passent par ce processus.

Épissage alternatif. L'épissage alternatif permet l'inclusion d'introns ou l'exclusion d'exons, générant ainsi différentes combinaisons pour créer une diversité d'ARNm, ou isoformes, à partir d'un seul gène. Bien que moins fréquent que l'épissage constitutif, ce processus est très répandu chez les eucaryotes supérieurs [173, 174]. Certains isoformes ont des fonctions biologiques distinctes, comme la production de différentes protéines, tandis que d'autres peuvent être non-fonctionnels et considérés comme du bruit. Certains

isoformes pourraient également participer au processus de transcription de l'ARN, et donc, à la régulation de l'expression des gènes [173, 174].

Les mutations affectant la sélection des sites d'épissage, constitutif ou alternatif, sont courantes et peuvent être impliquées dans les maladies humaines. En effet, entre 15% et 50% des mutations liées à des maladies humaines affectent la sélection des sites d'épissage [175].

1.3.2.1. Mécanisme de l'épissage

Il y a deux étapes de base à l'épissage : l'assemblage du complexe du spliceosome et l'épissage de l'ARN pré-messager. Les exons qui se retrouvent dans l'ARNm matures sont entièrement définis par l'interaction entre les éléments cis et les facteurs trans. Tout comme pour la transcription, il existe des régions importantes pour l'épissage.

Le "code d'épissage" contient généralement les sites d'épissage 5' (5'SS, *5' Splice-Site* en anglais) et 3' (3'SS, *3' Splice-Site* en anglais) en amont et en aval d'un exon. En amont du site 3' se trouve également deux sites importants dans la reconnaissance des facteurs d'épissage, soit un site de branchement normalement situé entre 15 et 50 nucléotides, ainsi qu'une séquence polypyrimidine d'environ 15 à 20 nucléotides pyrimidiques de long. Dans le cas d'exons alternatifs, le site de branchement (BPS, *Branch Point Sequence* en anglais) peut même être situé plus loin, jusqu'à 400 nucléotides [176, 177]. Autre que les paires de nucléotides "GU" et "AG" pour les sites d'épissage 5' et 3' respectivement, il n'y a pas vraiment de séquences consensus entre les gènes pour les régions entourant ces sites. Cependant, ces régions sont importantes pour la liaison avec les composantes des petits ARN nucléaires (snRNA, *small nuclear RNA* en anglais) du complexe du spliceosome [178, 179].

Une grande partie de l'épissage chez les organismes complexes se fait pendant la transcription du gène. Pendant la transcription, l'ARN Pol II peut influencer les niveaux d'épissage alternatif en fonction des facteurs d'épissage recrutés sur son CTD et sa vitesse d'élongation, qui détermine la synchronisation de la présentation des sites d'épissage [180]. Par exemple, si l'élongation se produit rapidement et qu'un site d'épissage associé à un exon a une faible affinité avec les facteurs d'épissage, il y a une probabilité plus élevée que les sites d'épissage de cet exon ne soient pas reconnus, entraînant ainsi l'épissage de cet exon dans le produit final et conduisant à un événement d'épissage alternatif de l'exon. Dans le cas contraire, si

l'élongation est lente, il y aura plus de temps de se lier au site ayant une plus faible affinité, ce qui réduit la quantité d'épissage de cet exon. En général, les sites autour des exons alternatifs ont une affinité plus faible que ceux autour des exons constitutifs [181]. Outre la transcription, l'épissage alternatif est également contrôlé par la structure de la chromatine, les ARN non-codants et la dégradation due à des codons de terminaison prématurés [173, 182, 183].

1.3.2.2. Éléments régulateurs de l'épissage

Les séquences des sites du "code d'épissage" sont courtes et peuvent être présentes à plusieurs endroits dans le génome. Les régions introniques peuvent contenir plusieurs séquences de sites d'épissage "leurres". Ces sites présentent des séquences similaires aux sites d'épissage véritables et sont souvent accompagnés de "pseudoexons" ayant une taille et une force d'interaction de sites d'épissage similaires à celles des exons. Malgré la présence de ces "leurres", le processus d'épissage est rarement erroné, ce qui suggère que d'autres éléments sont importants pour la régulation de l'épissage [184, 185].

Parmi ces éléments importants, on retrouve de nombreux éléments régulateurs en cis qui vont recruter des facteurs d'épissage agissant en trans. La régulation indirecte, ou régulation en trans, se produit quand l'effet de la mutation agit en cis sur un deuxième gène, tel qu'un facteur de transcription, et donc, ne se retrouve pas nécessairement dans la région proche du gène [186].

Les éléments cis incluent des activateurs d'épissage exonique (ESE, *Exonic Splicing Enhancer* en anglais) et intronique (ISE, *Intronic Splicing Enhancer* en anglais), s'ils sont localisés dans les exons ou les introns respectivement, qui sont liés à des facteurs trans qui stimulent l'épissage, tels que les membres de la famille des protéines riche en sérine/arginine (protéine SR, *Serine/Arginine-rich family of nuclear phosphoproteins* en anglais). À l'inverse, les inhibiteurs d'épissage exonique (ESS, *Exonic Splicing Silencer* en anglais) et intronique (ISS, *Intronic Splicing Silencer* en anglais) sont liés à des facteurs qui inhibent l'épissage, tels que les ribonucléoprotéines nucléaires hétérogènes [187, 188]. C'est la collaboration additive entre les inhibiteurs et les activateurs qui va inhiber ou promouvoir l'assemblage du spliceosome sur les sites d'épissage avec de faibles affinités. En général, les activateurs jouent un rôle important pour l'épissage consécutif des exons, tandis que les inhibiteurs sont relativement plus importants dans le contrôle de l'épissage alternatif [185]. Presque tous les exons

contiennent une séquence ESE interne. Les ESE vont recruter des membres de la famille des protéines activatrices, qui vont faciliter les interactions avec les protéines du spliceosome. Les ESS vont lier des répresseurs d'épissage et auront plusieurs mécanismes inhibiteurs possibles, tels que bloquer l'interaction entre des complexes du spliceosome, nécessaire à sa formation [185, 189]. Concernant les régulateurs introniques, les ISE et ISS se regroupent souvent près des sites d'épissage 5' ou 3', souvent à moins de 300 nucléotides d'un exon alternatif [190].

Le rôle régulateur d'une séquence dépend du contexte. Par exemple, son effet dépend de sa localisation. Une même séquence peut avoir un effet activateur si elle est dans l'exon, mais un effet inhibiteur si elle est située dans l'intron. Son activité peut également varier selon le gène, c'est-à-dire qu'elle peut être régulatrice pour un gène, mais sans effet sur un autre [185]. De plus, la composition du spliceosome diffère entre les types cellulaires, ce qui rend la régulation de l'épissage spécifique à chaque tissu [191].

Protéine SR. La famille des protéines SR est très importante dans le processus d'épissage. Cette famille contient des domaines riches en sérine et arginine (Domaine RS, *Serine/Arginine rich domain* en anglais), ainsi que des motifs de reconnaissance de l'ARN. Tout au long du processus d'épissage, la phosphorylation et la déphosphosphorylation des protéines SR ou de certains facteurs d'épissage sont des événements très dynamiques qui régulent les interactions protéine-protéine lors de la formation du spliceosome, les interactions protéines-ARN faisant intervenir les domaines RS, ainsi que dans les étapes catalytiques de l'épissage [192, 193]. Plusieurs kinases peuvent participer à cette dynamique, telles que la kinase dépendante à l'AMPc (PKA) [194, 195].

1.3.2.3. Formation du spliceosome

Le processus de l'épissage peut être résumé par la figure 1.7. L'épissage est catalysé par un complexe ribonucléoprotéique appelé spliceosome, qui est assemblé principalement sur le CTD de l'ARN polymérase. Le spliceosome est composé d'environ une centaine de protéines centrales et de cinq petits ARN nucléaires (U1, U2, U4/U6 et U5) qui forment de petites particules ribonucléoprotéiques (snRNP, *small nuclear RiboNucleoProtein* en anglais) en interagissant avec les protéines. Il y a formation de différents complexes.

Le premier complexe (Complexe précoce E) permet la reconnaissance du "code d'épissage". À ce stade, la famille des protéines SR permet de stabiliser les liaisons en liant un motif ESE dans les exons à combiner.

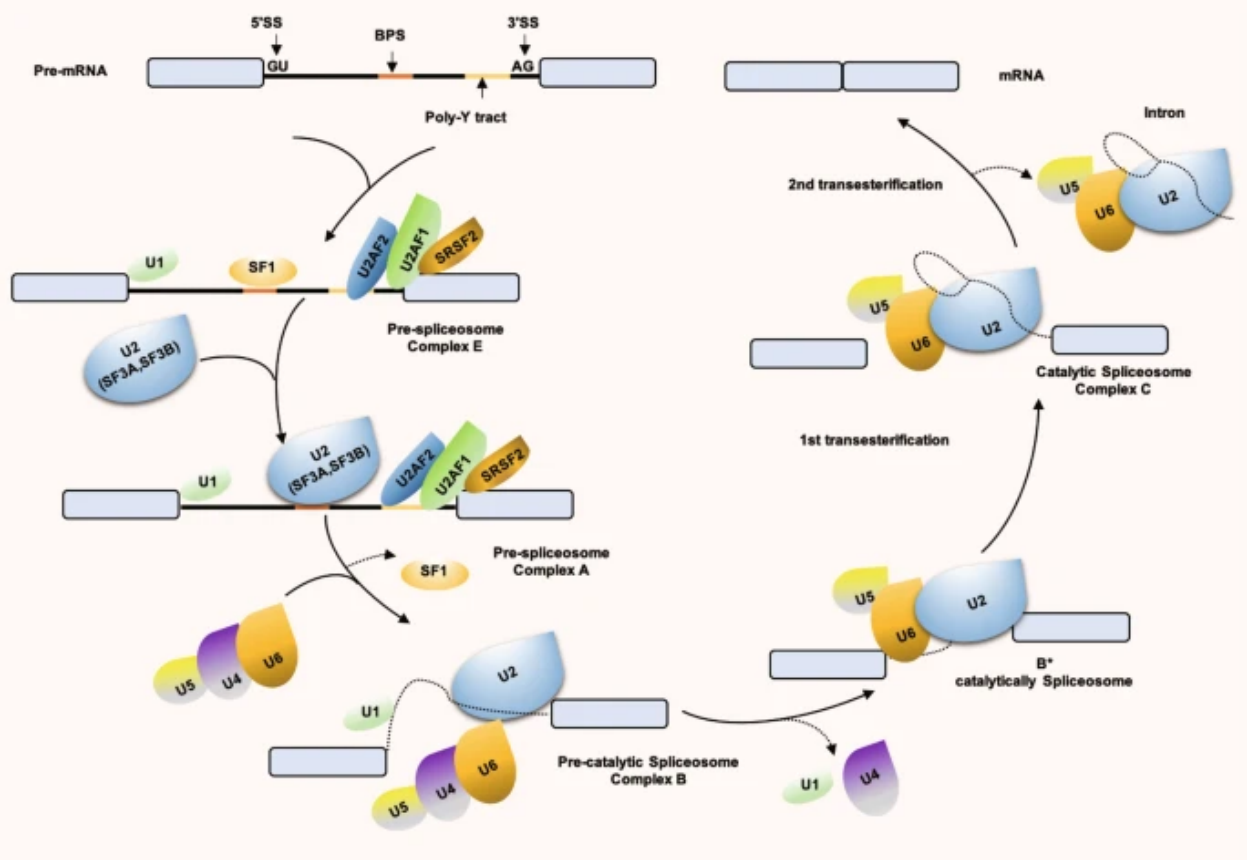

Fig. 1.7. Assemblage du spliceosome
Cheminement simplifié des étapes de l'épissage. **Complexe E** : Reconnaissance et liaison de 5'SS par le snRNP U1, liaison de SF1 à BPS, liaison de U2AF2 à la séquence polypyrimidine et liaison de U2AF1 à 3'SS. **Complexe A** : Le snRNP U2 se lie au BPS, l'unité U5/U4/U6 est également recruté. **Complexe B** : Combinaison des snRNP U4 et U6 par un appariement complémentaire de leurs composantes ARN. **Complexe B+** : Réaction de trans-estérification. **Complexe C** : Réaction de trans-estérification, connections des exons, dégradation de l'intron, formation de l'ARNm mature. SF1 : Facteur d'épissage 1, BPS : Séquence du site de branchement, SS : Site d'épissage, snRNP : Petit RNP nucléaires. Source : Zhang et al. [196]

Le complexe suivant (Complexe pré-spliceosome A) permet l'appariement de la région d'épissage 3' et du site de branchement en hydrolysant l'ATP. La formation de l'appariement des sites d'épissage se produit durant la formation de ce complexe et le mécanisme dépend de la longueur de l'intron. En général, pour les introns courts, le complexe se forme tout au long de l'intron. Ce mécanisme est appelé *intron definition* et c'est le mode prédominant chez les levures. Cependant, chez les organismes plus complexes, les exons sont généralement courts et les introns longs. Le complexe se formera alors à travers l'exon à conserver, ce qui permettra de joindre plusieurs petites unités d'exon au lieu d'une très longue unité à travers

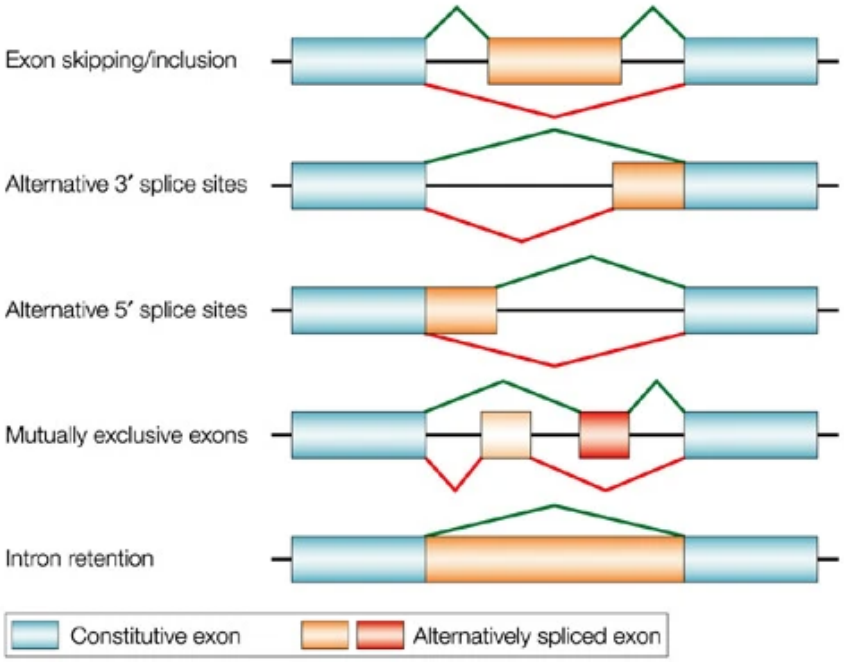

Fig. 1.8. Types d'épissage alternatif fréquents
Cinq types d'épissage alternatif souvent observés. Dans chaque cas, les chemins de l'épissage alternatif sont représentés en vert et en rouge, à l'exception du dernier cas qui correspond à l'absence d'épissage. Source : Cartegni et al. [200]

l'intron [197]. Ce mécanisme est appelé *exon definition*. C'est durant la formation de ce complexe que l'appariement des sites d'épissage se passe et donc le choix d'un événement d'épissage alternatif [198].

Les complexes suivants (B et C) sont catalytiques et procèdent à l'épissage de l'intron et le jumelage des exons.

1.3.2.4. Génération d'événements d'épissage

Il existe cinq principaux types d'événements d'épissage alternatif (Figure 1.8). L'épissage d'un exon, qui est le plus fréquent chez les mammifères, et la rétention d'un intron sont respectivement quand un exon est enlevé ou un intron est conservé dans le transcrit mature. L'exclusion mutuelle d'exons (MXE, *Mutually Exclusive Exons* en anglais) est quand les transcrits d'un gène contiennent un exon ou un autre, mais pas les deux. Un site d'épissage alternatif, soit en 5' soit en 3', est quand l'épissage de l'intron se fait plus loin dans l'intron [199].

Les différents types d'événement n'ont pas la même probabilité de donner des protéines fonctionnelles. En effet, l'épissage d'exon a plus tendance à former des isoformes alternatifs viables, tandis que la rétention d'intron, qui est la forme la plus présente chez les plantes, mène plus souvent à la dégradation de l'ARN [201].

1.3.2.5. Origine des événements d'épissage alternatif

Plusieurs causes peuvent être à l'origine d'événement d'épissage alternatif [202]. En général, cela peut arriver quand des mutations surviennent aux sites d'épissage ou ceux importants pour l'assemblage du spliceosome.

Un premier phénomène est par la duplication d'exon, qui est la principale source pour les événements MXE. Une copie pourrait conserver ses fonctions ancestrales tandis que la nouvelle pourrait mener à de nouvelles fonctions.

Un deuxième phénomène est la transition, c'est-à-dire qu'un exon constitutif devient un exon alternatif. Cela peut arriver par la mutation dans les séquences consensus ou régulatrices, ce qui diminue son affinité avec les facteurs d'épissage.

Un troisième est l'exonation qui arrive quand des sites d'épissage sont générés dans de courts rétrotransposons ou par la perte des éléments régulateurs inhibiteurs dans les longs rétrotransposons. Par exemple, un grand nombre d'exons spécifiques à une lignée phylogénétique sont survenus à partir d'éléments *Alu*, soient des rétrotransposons, dans le génome des primates [202] et jusqu'à 33 exons dérivés d'*Alu* chez l'humain ont été trouvés comme contribuant à des isoformes protéiques, mais dont les évidences de fonctions sont manquantes [203, 204].

1.3.2.6. L'importance de l'épissage alternatif dans l'évolution

Le mécanisme d'épissage médie divers processus importants dans le vivant, pouvant être impliqué dans le développement de tissus, la différenciation d'espèce et l'évolution du génome [173, 174]. L'épissage alternatif a été suggéré comme étant l'un des principaux moteurs de l'évolution et de la formation des caractéristiques spécifiques des espèces. En effet, depuis les années 70, il a été observé que des espèces qui divergent au niveau phénotypique possèdent pourtant des séquences d'ADN très similaires [205], telles que l'humain et le chimpanzé qui partagent 98.8% de similarité dans leur génome [206]. Cela peut suggérer que les séquences régulatrices jouent un rôle dominant afin de causer des différences phénotypiques. De plus,

l'épissage n'a pas la même utilité entre les espèces. Il a été observé que les plantes utilisent ces événements principalement pour la régulation de gènes en réponse aux stress, tandis que les animaux les utilisent principalement pour générer des protéomes spécifiques aux tissus [207].

Il a été observé que les mutations générant l'épissage sont souvent négligées par les pressions de sélection. En effet, lorsqu'une forme alternative du gène apparaît, elle sera à faible fréquence et n'aura pas nécessairement des fonctions. Cela fait en sorte que les pressions de sélection ne la toucheront pas nécessairement, permettant une évolution sous neutralité. Avec le temps, la forme alternative pourrait gagner des mutations, et donc, potentiellement acquérir des nouvelles fonctions pouvant être bénéfiques [181, 208].

Par exemple, les chauves-souris détectent les chaleurs nocives via le canal d'ion sensitif à la chaleur (TRPV1), qui s'active lorsque les températures dépassent 43°C [209]. Une mutation causant un événement d'épissage alternatif conduit à la production d'un isoforme du canal qui s'active à partir d'un seuil de 30°C. Chez la plupart des espèces de chauves-souris, cet isoforme alternatif n'offre pas d'avantage sélectif et est donc présent à faible fréquence. Cependant, chez les chauves-souris vampiriques, cet isoforme alternatif est plus fréquent dans les fibres nerveuses du visage. Cela leur permet de détecter les sources de chaleur infra-rouge nécessaire à la localisation des vaisseaux sanguins de leurs proies [209]. Cette adaptation spécifique des chauves-souris vampiriques démontre que l'épissage alternatif peut jouer un rôle crucial dans l'adaptation des espèces, et sa régulation peut être spécifique à certains types cellulaires ou organes.

On peut donc dire qu'avec l'aide de l'épissage alternatif, l'évolution peut parvenir à trouver de nouvelles solutions sans nécessairement éliminer les anciennes.

1.3.3. Méthodologies pour quantifier l'ARN

Afin d'identifier les variants génétiques qui influencent les niveaux d'expression (eQTL, *expression Quantitative Trait Loci* en anglais) ou d'épissage (sQTL, *splicing Quantitative Trait Loci* en anglais), il est nécessaire de mesurer les niveaux d'expression des gènes et de séquencer le génome des individus étudiés. Il existe différentes méthodes pour obtenir ces mesures.

Puce à ADN. La première technologie permettant de quantifier les niveaux d'expression des gènes est la puce à ADN, développée dans les années 90. Cette technologie a été conçue pour détecter et quantifier simultanément l'expression de milliers de gènes en utilisant un processus d'hybridation de l'ARN sur une puce contenant des brins d'ADN [210].

Séquençage deuxième génération. Contrairement à la puce à ADN, qui permet de quantifier uniquement une liste prédéfinie de gènes, le séquençage d'ARN (RNA-Seq, *RNA-Sequencing* en anglais) permet de séquencer et de quantifier l'ensemble du transcriptome, y compris les transcrits non-codants, les différents isoformes d'un gène et les différentes modifications post-transcriptionnelles [211]. De plus, cette techniquement permet de quantifier l'ARN même lorsque la concentration d'ARN est faible, permettant d'effectuer des analyses à partir d'une seule cellule, soit du *single-cell RNA-Seq* [212].

La principale limitation de cette technologie réside dans la difficulté à reconstruire précisément les transcrits [211]. Étant donné que les fragments de lecture sont courts, souvent entre 50 et 300 paires de base, il peut y avoir des alignements sur des régions similaires au gène d'origine, ce qui peut fausser la quantification.

Séquençage de troisième génération. Afin de pallier à la limitation majeure du séquençage de deuxième génération, des avancées technologiques ont permis le développement du séquençage d'ARN à longs fragments de lecture, où les fragments peuvent atteindre des longueurs supérieures à 100'000 paires de bases. Cette approche présente l'avantage potentiel de couvrir un transcrit complet, ce qui réduit les erreurs liées à la reconstruction des transcrits [213].

Variabilité du profil transcriptomique. Le profil transcriptomique d'un organisme peut être modifié en réponse à des changements environnementaux [161, 214]. Étant donné que les données du transcriptome sont obtenues à un moment spécifique, elles ne capturent que le profil à ce moment précis dans le temps. Par conséquent, le profil transcriptomique d'un même individu peut varier entre deux extractions de matériel biologique. Cette variation temporelle peut constituer une limitation lors de l'analyse génétique, car elle peut introduire du bruit non génétique dans les données.

Cependant, dans d'autres contextes, cette variation temporelle peut être utilisée pour comparer le profil transcriptomique d'un individu dans deux conditions différentes, afin

d'identifier les gènes influencés par ces conditions [161]. Dans les études portant sur l'effet génétique de manière plus globale, l'inclusion d'un nombre suffisant d'individus permet de capturer l'effet moyen des mutations d'intérêt sur l'expression génique.

1.3.4. Manipulation des données de séquençage d'ARN

Le séquençage d'ARN permet d'obtenir des fichiers contenant les séquences des fragments de lecture. Cependant, ces fichiers ne fournissent pas d'informations sur les gènes associés à ces séquences, il est donc nécessaire de les aligner sur le génome ou le transcriptome de référence. Une fois l'alignement réalisé, il est possible d'analyser les niveaux d'expression des transcrits ou d'évaluer les niveaux d'épissage alternatif.

Alignement et contrôle de qualité. Les premières étapes sont l'étape de l'alignement des fragments de lecture et l'étape du contrôle de qualité de l'échantillon. En général, on s'attend à ce que seulement entre 70% et 90% des fragments s'alignent sur le génome humain [215]. L'alignement permet de déterminer les coordonnées du fragment de lecture dans le génome de référence. Des biais peuvent également survenir à différentes étapes depuis le prélèvement, tels que la dégradation de l'ARN, nécessitant ainsi des contrôles de qualité [216, 217].

1.3.4.1. Quantification des niveaux d'expression

Une fois les données alignées et un contrôle de qualité effectué, diverses manipulations peuvent être réalisées. L'une des manipulations courantes consiste à quantifier les niveaux d'expression des gènes ou des isoformes.

Grâce à des outils bio-informatiques, tels que RSEM [218] et kallisto [219], il est possible d'estimer le niveau d'expression de chaque gène en évaluant l'abondance des fragments de lecture qui s'alignent sur ces gènes. Cependant, avec le séquençage de deuxième génération, un gène plus long aura généralement un plus grand nombre de fragments de lecture qu'un gène plus court [220, 221]. Cela ne signifie pas nécessairement qu'il est davantage exprimé dans les cellules. De plus, il peut exister des variations d'abondance des transcrits entre les individus en raison des différences de couverture ou de profondeur de séquençage. Pour atténuer ces biais, des méthodes de normalisation des données doivent être utilisées.

Plusieurs méthodes de normalisation sont couramment utilisées, telles que la normalisation des lectures par million de kilobases (RPKM, *Reads Per Kilobase Million* en anglais),

la normalisation des transcrits par million (TPM, *Transcripts Per Kilobase Million* en anglais), ainsi qu'une normalisation en deux étapes avec la moyenne tronquée des valeurs M (TMM, *Trimmed Mean of M values* en anglais) [222] et "voom" [223]. Les normalisations RPKM et TPM permettent de normaliser selon la taille du gène et la profondeur de la bibliothèque, permettant ainsi de mesurer l'abondance des gènes et de comparer les niveaux entre les individus [220, 221]. Quant à la troisième méthode, elle normalise et transforme les données entre les échantillons et entre les gènes d'un échantillon, améliorant la fiabilité des comparaisons entre les gènes et les échantillons [221, 222].

Après ces normalisations, les données seront prêtes à être utilisées dans des analyses comparatives.

1.3.4.2. Quantification des niveaux d'épissage alternatif

Un gène peut avoir différents isoformes qui peuvent avoir des profils d'expression distincts. L'étude des isoformes peut être abordée à travers différentes approches.

La première consiste en la reconstruction des transcrits, réalisée par des outils tels que MISO [224], RSEM [218] et kallisto [219]. Ces méthodes utilisent des modèles statistiques pour estimer la fréquence ou l'abondance relative de chaque transcrit. Elles sont efficaces lorsque les niveaux d'expression sont élevés, qu'il y a une bonne couverture de séquençage et que les événements d'épissage ne se chevauchent pas entre les transcrits [225].

La deuxième approche consiste à comparer l'utilisation différentielle des caractéristiques du gène, telles que les exons, entre différentes conditions. Cette approche est utilisée par des outils tels que DEXSeq [226] et edgeR [227]. Cependant, cette approche ne permet pas de quantifier directement les isoformes et se limite généralement aux informations associées aux jonctions étudiées. De plus, elle ne permet pas de prédire le type d'épissage, car elle se contente de comparer les comptes des caractéristiques entre des conditions expérimentales [228].

La troisième approche consiste à utiliser les informations sur les jonctions pour inférer les événements d'épissage. Cette approche est utilisée par des outils tels que MAJIQ [229], LeafCutter [230] et ASpli [231]. Elle utilise les fragments de lecture qui couvrent au moins deux exons pour quantifier la fréquence des jonctions d'épissage. Une métrique couramment utilisée dans cette approche est la valeur PSI (*Percent Spliced In* ou ψ), qui représente le pourcentage de fragments de lecture incluant un exon spécifique ou un site d'épissage. Tout

comme la deuxième approche, celle-ci ne permet pas de quantifier directement les isoformes et se limite généralement aux informations associées aux jonctions étudiées [231]. De plus, la précision de la valeur PSI dépend de la couverture de séquençage de l'événement.

1.3.4.3. Corrections des données

Lors de la quantification des niveaux d'expression, il existe plusieurs facteurs autres que les facteurs génétiques qui peuvent l'influencer les résultats.

Un des biais possible est un biais technique. Les expérimentations peuvent être effectuées par différentes personnes, à différents moments et parfois même dans différents laboratoires. Cela entraîne des conditions expérimentales qui diffèrent entre les échantillons, ce qui crée un biais technique (ou effet de *batchs*). Ces variations introduisent du bruit qui peut masquer ou générer des signaux indésirables. Contrairement aux facteurs liés aux individus, tels que l'âge et le sexe, il n'est pas possible d'attribuer directement une valeur à ce type de biais. Par conséquent, diverses techniques ont été développées pour estimer ces facteurs cachés.

Parmi les techniques les plus utilisées, on trouve l'analyse en composantes principales (PCA), l'analyse de variables substitutives (SVA, *Surrogate Variable Analysis* en anglais) [232] et l'estimation probabiliste des résidus d'expression (PEER, *Probabilistic Estimation of Expression Residuals* en anglais) [233]. L'approche PCA pour l'expression est similaire à celle utilisée pour capturer l'ethnicité génétique. Elle vise à capturer la variation observée dans les données d'expression. L'approche SVA est similaire à la PCA, mais elle permet de protéger une variable d'intérêt. Elle est donc idéale pour des analyses visant à comparer les niveaux d'expression entre différentes conditions, car elle élimine les autres sources de bruit. La dernière technique mentionnée, PEER, utilise une approche bayésienne pour capturer les variations. Le logiciel ne cherchera pas à éliminer toutes les sources de variation, mais essayera de décomposer les variations observables en facteurs d'intérêt ou en facteurs confondants.

1.4. Épistasie

La plupart des disciplines de recherche présentées dans les sections précédentes ne prennent pas pleinement en compte les interactions entre les mutations, qui jouent pourtant un rôle crucial dans ces domaines. Cependant, en négligeant ces interactions, plusieurs associations peuvent être manquées, telles que la présence de co-évolution entre des gènes

(voir section 1.4.3), mais également générer le problème de l'héritabilité manquante. En effet, ce problème survient lorsque, malgré l'identification de nombreuses mutations génétiques associées à des phénotypes, elles ne parviennent pas à expliquer l'intégralité de la variabilité phénotypique observée [156, 234]. Plusieurs raisons peuvent expliquer cette situation, telles que la pénétrance incomplète du phénotype, ainsi que les interactions avec l'environnement ou entre les gènes, c'est-à-dire l'épistasie [235]. Dans cette section, nous explorerons plus en détails le concept d'épistasie, son impact sur la génétique des populations et ses implications dans différents domaines de recherche.

1.4.1. Définition

Le concept d'épistasie remonte à plus d'un siècle et a été introduit par William Bateson en 1909 pour décrire les interactions entre les gènes. Le terme épistasie se traduit par "*se tenir dessus*" et désigne le phénomène où l'action d'un locus masque les effets d'un autre locus. Le locus qui masque les effets est appelé "épistatique", tandis que le locus dont les effets sont masqués est appelé "hypostatique" [236, 237].

Aujourd'hui, il est possible de séparer la définition d'épistasie en plusieurs catégories. Cependant, les termes utilisés pour les définitions ne sont pas unanimes entre les auteurs. Dans cette thèse, je présente les termes définis par P.C. Phillips [156].

Le premier type d'épistasie est l'**épistasie fonctionnelle**, aussi appelé "interaction protéine-protéine". Cette définition est strictement fonctionnelle sans lien direct avec la génétique. Ce concept fait référence aux interactions moléculaires que les protéines, ou d'autres éléments génétiques, ont entre elles, que ce soit des interactions directes ou qu'elles opèrent sur une même voie signalétique. Un exemple mentionné précédemment est l'interaction entre les gènes *Adcy9* et $CETP^{tg}$ chez la souris [89]. Dans cette étude, aucune mutation génétique n'a été prise en compte. Au lieu de cela, les niveaux d'expression des protéines ont été directement modifiés, ce qui a permis d'observer des différences d'effet sur divers phénotypes. Ainsi, cette interaction entre les gènes *Adcy9* et $CETP^{tg}$ serait considérée comme un exemple d'épistasie fonctionnelle.

Un deuxième type d'épistasie est l'**épistasie compositionnelle**, un terme introduit par William Bateson en 1909. Il représente la définition traditionnelle de l'épistasie, où un allèle à un locus "A" influence l'effet d'un allèle à un locus "B", et le phénotype résultant dépend de

la combinaison des génotypes. Pour étudier ce type d'épistasie, il est nécessaire de modifier directement les positions génétiques d'intérêt tout en maintenant le même contexte génétique (*genetic background* en anglais). Cette approche permet de mesurer les effets de substitutions alléliques dans un contexte génétique fixe. Étant donné qu'il est essentiel de maintenir le même contexte génétique, ce type d'étude n'est généralement pas réalisé dans des populations naturelles, mais plutôt dans des expériences utilisant des modèles animaux ou cellulaires. Ainsi, l'étude de l'épistasie compositionnelle nécessite une approche expérimentale contrôlée où les génotypes sont ciblés et modifiés pour étudier les interactions spécifiques entre les allèles à différents loci.

Le troisième type d'épistasie est l'**épistasie statistique**. Cette définition a été introduite par Ronald Fisher et concerne la déviation du modèle additif [238]. En d'autres mots, l'interaction entre deux mutations ne correspond pas à l'effet attendu si le modèle était purement additif (Équation 1.4.1). Les analyses basées sur ce concept peuvent être réalisées à l'aide de bases de données génétiques et d'informations phénotypiques, ce qui en fait le modèle le plus couramment utilisé en bio-informatique pour évaluer des phénotypes quantitatifs [239].

$$P_{AB} \neq P_A + P_B \tag{1.4.1}$$

Dans l'équation 1.4.1, $P_{\{A,B\}}$ représente l'effet du locus "A" ou "B" sur le phénotype, tandis que P_{AB} représente l'effet de la combinaison des loci sur le phénotype. Contrairement à l'épistasie compositionnelle, l'épistasie statistique ne nécessite pas un contexte génétique fixe et peut donc être utilisée pour des analyses dans une population naturelle. Cette approche examine l'interaction avec un contexte génétique "moyen" [156]. Cela implique l'utilisation de grandes bases de données et la correction par des covariables.

1.4.1.1. Types d'épistasie

L'interaction entre deux mutations peut avoir plusieurs effets. En l'absence d'épistasie, la valeur du phénotype sera simplement la somme des effets des deux loci (Figure 1.9-Sans épistasie).

Un premier effet est appelé épistasie positive ou effet synergique (Figure 1.9-Épistasie positive). Cela se produit quand l'effet des deux allèles combinés est amplifié lorsqu'ils se trouvent dans une certaine combinaison. Si l'effet est positif, cette combinaison augmentera

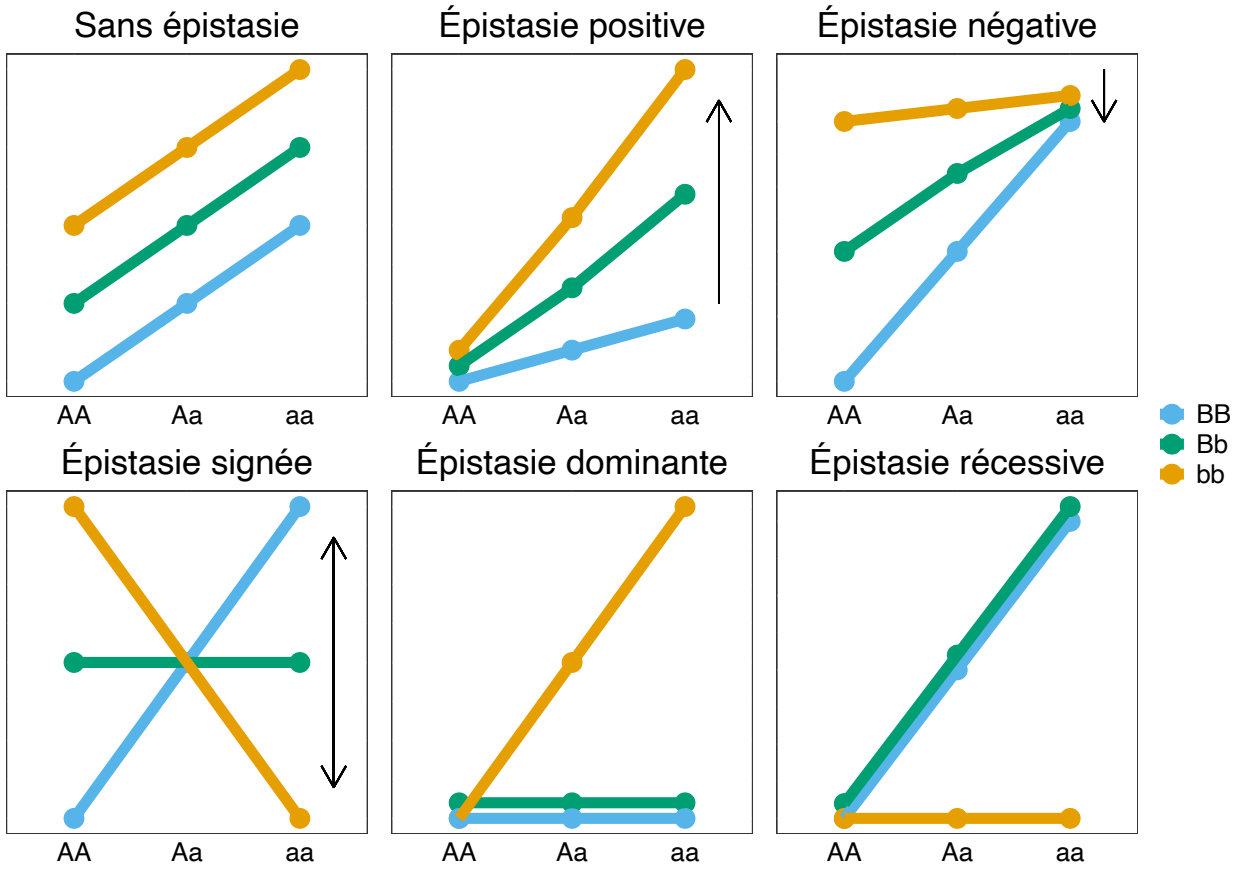

Fig. 1.9. Différents types d'épistasie
L'effet sur un phénotype selon la combinaison des génotypes des loci "A" et "B". Dans ces graphiques, le locus "B" est le locus épistatique et le locus "A" est le locus hypostatique. La flèche flèche indique la direction de l'effet de l'interaction.

davantage l'effet, et vice-versa si l'effet est négatif [156, 240]. Par exemple (Figure 1.9-Épistasie positive), l'effet du locus "A" est fortement amplifié et non additif selon le génotype du locus "B".

Un deuxième effet est appelé épistasie négative ou effet antagonique (Figure 1.9-Épistasie négative). Contrairement à l'épistasie positive, l'effet est atténué par rapport à ce qui est attendu, pouvant même aller jusqu'à l'absence totale d'effet [156, 240].

Un troisième effet est appelé l'épistasie signée (Figure 1.9-Épistasie signée) [241]. Cela signifie que la direction de l'effet du locus "A" varie en fonction du locus "B". Par exemple, si les allèles individuels diminuent le succès de reproduction, mais que, dans une certaine combinaison, ils l'augmentent, cela est considéré comme une épistasie signée [156]. Si l'inversion d'effet est similaire des deux côtés, cela peut même masquer complètement le signal du locus hypostatique lorsqu'il est considéré seul dans une analyse de GWAS.

Tout comme dans les modèles d'hérédité de Mendel, il peut exister de l'épistasie récessive (Figure 1.9-Épistasie récessive) ou dominante (Figure 1.9-Épistasie dominante).

Dans le cas de la récessive, tant que le locus épistatique "B" n'est pas homozygote récessif, le locus hypostatique "A" pourra exercer son effet sur le phénotype. Cependant, lorsque le locus "B" devient homozygote récessif, le phénotype sera entièrement déterminé par le locus "B".

Dans le cas de l'épistasie dominante, tant qu'il existe un allèle majeur du locus épistatique "B", l'effet sera contrôlé uniquement par le locus "B". Cependant, lorsque le locus "B" devient homozygote récessif, le phénotype sera modulé par le locus hypostatique "A".

1.4.2. Mécanismes biologiques de l'épistasie

Plusieurs mécanismes biologiques peuvent engendrer une relation épistasique entre deux gènes [242], en voici quelques-uns.

Interaction protéine-protéine. Un changement dans l'une des protéines peut modifier sa reconnaissance par l'autre protéine, ce qui peut entraîner une épistasie positive ou négative.

Redondance fonctionnelle. Deux gènes ou voies métaboliques peuvent effectuer des fonctions moléculaires communes ou similaires. Ainsi, la perte d'un gène n'entraîne que des effets modestes sur la fonction, car la perte du gène est compensée par l'autre

Contrainte physique ou chimique. Lorsqu'un trait a une valeur maximale ou minimale, la combinaison de mutations peut entraîner une effet supérieur ou inférieur à ces valeurs respectivement. Par conséquent, l'effet observé ne sera pas linéaire.

Il convient de noter que ces exemples ne couvrent pas tous les mécanismes possibles, mais ils illustrent certains des processus couramment observés dans les relations épistasiques entre les gènes.

1.4.3. Co-évolution génétique

En 1931, Sewall Wright, l'un des pères de la génétique des populations, a proposé sa "*théorie de l'équilibre changeant*" [243] comme mécanisme de sélection agissant sur les combinaisons de gènes. Il a présenté deux possibilités pour expliquer comment plusieurs gènes peuvent interagir pour produire un phénotype, pouvant entraîner des pressions de sélection co-évolutive entre les gènes.

La première possibilité est que l'effet combiné des allèles à différents loci soit égal à la somme des effets individuels, ce qui correspond au modèle additif. La seconde possibilité est que l'effet combiné des allèles ait un effet supérieur ou inférieur à leur effet additif, ce qui correspond au modèle non-additif ou l'épistasie. Wright a reconnu que l'épistasie pouvait complexifier le processus évolutif [244].

Prenons un exemple simple avec un organisme haploïde, c'est-à-dire qu'il possède une seule copie de chaque chromosome, et considérons deux loci, "A" et "B". À l'état initial, la valeur adaptative de la combinaison "AB" est de 1. Lorsque les mutations se produisent à ces loci, cela peut entraîner une différence au niveau de la valeur adaptative (*fitness*) de l'individu, ce qui crée une pression de sélection représentée par le coefficient de sélection s. Supposons que les mutations aux deux positions confèrent un avantage, qui se traduit par un coefficient de sélection s de 0.1, ce qui signifie que la valeur adaptative de la combinaison "ab" est de 1.1.

Dans un modèle additif, les combinaisons "aB" et "Ab" auront une valeur adaptative comprise entre 1 et 1.1, et la pression de sélection favorisera l'augmentation de la fréquence des deux mutations afin que les allèles "a" et "b" soient plus fréquents dans la population.

Cependant, dans le cas d'un modèle avec interaction, il est possible que les combinaisons "Ab" et "aB" confèrent un désavantage sélectif, c'est-à-dire une valeur adaptative inférieure à 1. Cela signifie que la population conservera la combinaison "AB" qui a une valeur adaptative de 1, même si la combinaison "ab" a une valeur adaptative supérieur. Dans cette situation, l'épistasie agit comme une contrainte évolutive, limitant la capacité des mutations à se propager dans la population [244].

La visualisation à l'aide de seulement deux loci chez les haploïdes est simple et facile à interpréter. Cependant, cela devient plus complexe chez les organismes diploïdes ou lorsqu'il y a des interactions avec plusieurs loci. C'est pour aider à visualiser ce concept que Wright a élaboré sa "théorie de l'équilibre changeant" [243]. Il a utilisé la métaphore du paysage adaptatif pour illustrer ce concept (voir la figure 1.10). Ce paysage présente des creux où la valeur adaptative est la plus faible, ainsi que des sommets où cette valeur est la plus élevée.

Selon cette métaphore, une population est représentée à une certaine position sur la carte d'un paysage adaptatif. À partir de cette population, plusieurs populations distinctes et isolées les unes des autres seront générées. Des mutations apparaissent de manière indépendante

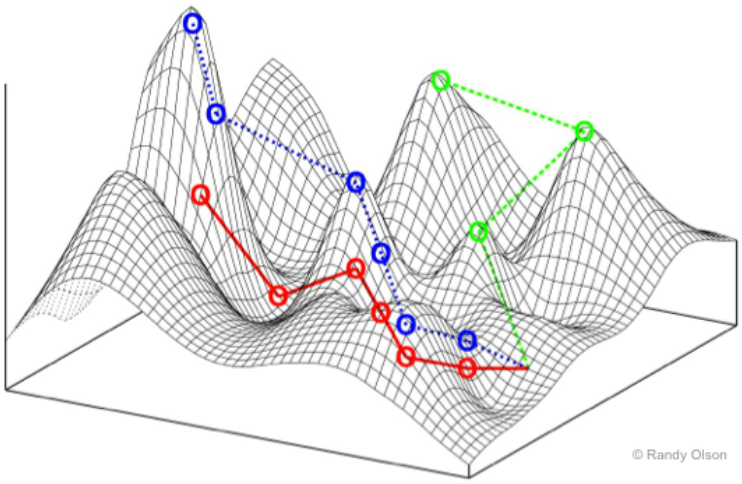

Fig. 1.10. Paysage adaptatif sur deux dimensions.
L'axe vertical représente la valeur de sélection. Les axes horizontaux représentent les variables en interaction, telles que les fréquences des génotypes ou les valeurs des phénotypes. Les sommets sont les endroits où la valeur de sélection est le plus élevé. Les différentes lignes représentent les différentes possibilité d'évolution qu'une population pourrait suivre à travers les vallées des valeurs de sélection. (Source : *NK Fitness Landscape* par Randy Olson)

dans ces populations, ce qui entraîne des changements dans les valeurs adaptatives. Au fil du temps et des influences environnementales propres à chaque population, les mutations entraînent des déplacements de la population dans le paysage, soit vers des creux, soit vers des sommets. Entre deux sommets, il y a inévitablement un creux où la valeur adaptative est plus faible. Une fois qu'une population atteint un sommet, il devient difficile de passer d'un sommet à un autre uniquement par dérive génétique et génération de mutations.

Cependant, si une pression de sélection survient, combinée à la dérive génétique, à la migration et au métissage entre populations présentant des fréquences alléliques diffèrentes, ces forces peuvent être suffisamment puissantes pour permettre un changement de sommet [244, 245]. Ainsi, la combinaison spécifique de variables joue un rôle dans la survie dans un environnement, ce qui correspond à la théorie de la co-évolution. Si les variables sont des mutations génétiques, la combinaison de génotypes la plus avantageuse correspond à un sommet, et ces génotypes seraient observés ensemble plus fréquemment que prévu, entraînant un déséquilibre de liaison potentiellement sur une longue distance (voir section 1.2.1.4).

Plusieurs exemples de co-évolution ont déjà été identifiés. Parmi ceux-ci, un exemple a été observé lors du séquençage du génome du chimpanzé [246]. Lors de la comparaison avec le génome humain, des chercheurs ont remarqué que certains allèles de référence chez le chimpanzé correspondaient à des allèles associés à des maladies chez l'humain, tels que la substitution *Asn29Thr* dans le gène *PRSS1* associée à la pancréatite héréditaire [246]. Cependant, ces maladies ne sont pas présentes chez les chimpanzés. Une hypothèse a été émise que la tolérance de ces allèles délétères pourrait être due à une interaction épistasique compensatoire. Cela signifie qu'un autre locus aurait un effet d'atténuation sur l'allèle pathogène, réduisant ainsi son impact sur le phénotype.

1.4.4. Effet sur la transcriptomique

Tout comme pour les mutations individuelles, l'effet des interactions épistasiques entre des mutations sur un phénotype peut être médié par l'expression des gènes. L'influence de ces mutations sur la variation de l'expression génique peut contribuer à l'identification des réseaux de régulation dans lesquels les gènes d'intérêt interagissent [239].

Dans le cas d'interaction épistasique de deux loci qui modulent l'expression (eQTL) ou l'épissage (sQTL), il y a trois types d'interaction possibles : le *cis-cis*, le *cis-trans* et le *trans-trans*. L'interaction *cis-trans* est relativement simple à comprendre, car elle implique que le *trans*-QTL du premier locus régule l'effet *cis*-QTL du deuxième locus. Dans le cas de l'interaction *cis-cis*, elle peut être due à un déséquilibre de liaison. En ce qui concerne l'interaction *trans-trans*, sa complexité réside dans le grand nombre d'interactions possibles [247].

Dans une étude menée par Becker et al. [247], portant sur 440 transcrits, il a été observé que 15% d'entre eux étaient régulés par une interaction entre deux loci, ce qui est supérieur au nombre attendu sous l'hypothèse nulle, soit que 5% des transcrits soient associés par chance. Étant donné que les phénotypes sont régulés par l'expression des gènes, l'épistasie peut jouer un rôle crucial dans la compréhension et la régulation des phénotypes complexes [248]. Cependant, dans cette étude, la plupart des loci en interaction n'avaient pas d'effet significatif lorsqu'ils étaient considérés individuellement. Cela signifie qu'ils pourraient être négligés lors des analyses d'association classiques, qui ne prennent en compte que les effets individuels des loci.

1.4.5. Associations phénotypiques

Il est relativement simple d'évaluer une association linéaire entre la génétique et un phénotype. Cependant, lorsque des interactions sont impliquées, la complexité augmente considérablement. Dans les analyses de GWAS, généralement entre 500'000 et un million de variants génétiques sont examinés. Cependant, pour étudier les interactions, le nombre de tests augmente de façon exponentielle pour chaque mutation à considérer, ce qui réduit considérablement la puissance statistique et nécessite une quantité de temps de calcul et de mémoire importantes.

Plusieurs approches sont possibles pour réduire le nombre de tests à effectuer [105, 249]. Une approche consiste à ne conserver que les variants qui présentent un certain niveau d'association dans un modèle linéaire. Cependant, cette approche ne permettra pas de détecter les variants qui ne montrent un effet qu'en présence d'interaction. Une autre possibilité est de filtrer les variants selon un contexte biologique, tel qu'une famille de protéines ou une voie signalétique. Le but de ces méthodes est de réduire le nombre de tests d'interaction à réaliser, mais cela limite la découverte aux fonctions connues des gènes et ne permet pas la découverte de nouvelles fonctions.

Chez l'humain, plusieurs interactions ont déjà été identifiées pour des traits polygéniques [249, 250], et certaines de ces connaissances sont utilisées dans le développement de traitements [251].

1.4.6. Application à la médecine de précision

La découverte d'interactions épistasiques permet de mieux comprendre les gènes et peut également être utilisée en clinique. Par exemple, dans le traitement des cancers du sein chez les porteurs de mutations dans les gènes *BRCA*, qui sont associées à un risque accru de développer ce type de cancer, une approche basée sur l'épistasie est utilisée [5]. Il a été découvert que l'inhibition de la protéine PARP n'affecte pas les cellules saines, mais entraîne la mort des cellules porteuses de mutations dans les gènes *BRCA*. Cette découverte a conduit au développement d'inhibiteurs de PARP pour le traitement des cancers du sein associés à ces mutations. Cet exemple démontre que l'identification et la compréhension des interactions épistasiques permettent de développer des traitements qui exploitent un phénotype favorable dans des groupes spécifiques de personnes en fonction de leur génétique.

1.5. Questions de recherche et introduction des projets

La médecine évolutionniste est un domaine de recherche qui explore les interactions entre l'évolution humaine et la santé.

Hypothèse globale. En appliquant les principes de la génétique des populations à la pharmacogénomique, il est possible d'élargir notre compréhension de l'évolution des variations génétiques associées à la pharmacogénomique et comment elles peuvent influencer la réponse individuelle aux médicaments.

L'**objectif global** de cette thèse est d'étudier les gènes liés à la pharmacogénomique en utilisant des analyses de génétique des populations, de transcriptomique et d'étude d'association phénotypique.

Cette thèse sera divisée en deux parties. La première partie, présentée dans les chapitres 2 et 3, se concentrera sur l'étude de la relation entre le gène *CETP*, dont la protéine est ciblée par le médicament dalcetrapib, et le gène *ADCY9*, dont une mutation a été identifiée comme modulant la réponse au dalcetrapib. La deuxième partie, présentée dans le chapitre 4, utilisera des méthodologies utilisées dans la première partie en se concentrant sur la superfamille des *CYP450*, famille de gènes impliquée dans le domaine de la pharmacogénomique et de la médecine de précision.

1.5.1. Projet : Étude post-pharmacogénomique du dalcetrapib

Le mécanisme sous-jacent derrière l'association entre la mutation rs1967309 dans le gène *ADCY9* modulant la réponse au dalcetrapib et la protéine ciblée, soit CETP, demeure inexpliquée.

Hypothèse du chapitre 2. L'étude des pressions de sélection nous permettra de mieux comprendre le lien entre les différents gènes étudiés, en s'appuyant sur des analyses de génétique des populations, de transcriptomique et d'études d'association phénotypique.

Objectifs spécifiques du chapitre 2 :

- Identifier les signatures de pression évolutive dans le gène *ADCY9* avec des analyses de génétique des populations;
- Analyser les liens évolutifs du gène *ADCY9* avec le gène *CETP* avec des analyses de génétique des populations;

- Étudier l'effet de l'interaction épistasique entre les gènes *ADCY9* et *CETP* sur la transcriptomique et sur les phénotypes.

Dans le chapitre 2, nous avons identifié de nombreux liens appuyant la présence d'une relation entre ces deux gènes. Grâce à une approche basée sur le LRLD, nous avons identifié la présence d'une interaction démontrant une co-évolution et qui montre des différences entre les sexes, suggérant une pression de sélection spécifique aux sexes.

Suite aux analyses présentées au chapitre 2, nous avons émis une nouvelle hypothèse en lien au mécanisme reliant la fonction des deux gènes. La mutation identifiée sous pression de sélection dans le gène *CETP*, soit la mutation rs158477, se trouve à moins de 200 paires de base de l'exon 9 de *CETP*, exon qui est épissé dans le deuxième isoforme de *CETP*.

Hypothèse du chapitre 3. La régulation de l'épissage alternatif de l'exon 9 de *CETP* pourrait être impliquée dans la pression de sélection identifiée.

Objectifs spécifiques du chapitre 3 :

- Caractériser la régulation génétique des isoformes du gène *CETP* avec des analyses d'expression et d'épissage alternatif;
- Étudier l'effet de la variation des proportions des isoformes de *CETP* sur des phénotypes via des analyses de randomisation mendélienne;
- Évaluer l'effet de l'interaction épistasique entre les gènes *ADCY9* et *CETP* sur les niveaux d'épissage alternatif.

Dans le chapitre 3, nous avons fait état des limitations dans l'évaluation de l'expression de *CETP* au niveau du gène et nous avons présenté les avantages de considérer les isoformes. Grâce à des analyses de randomisation mendélienne, nous avons identifié que le gène *CETP* pourrait être important dans le bon fonctionnement des glandes thyroïdiennes et/ou pituitaires, et nous avons identifié de nouvelles associations causales avec des phénotypes potentiellement liés à la signature de sélection identifiée au chapitre 2. Également en continuation avec les résultats présentés au chapitre 2, nous avons identifié une relation épistasique sur l'épissage alternatif de l'exon 9 impliquant la mutation rs158477 dans le gène *CETP* et le locus rs1967309 dans le gène *ADCY9*.

1.5.2. Projet : Caractérisation de la sous-famille des *CYP3A*

La superfamille des *CYP450* est importante dans la survie des individus dans différents environnements, puisque leur principale fonction réside dans la détoxification du corps en éliminant les molécules provenant de l'environnement externe. Beaucoup de ces sous-familles proviennent d'événements de duplication et se retrouvent proches dans le génome. Des hypothèses ont été émises que plusieurs sous-familles des CYP seraient sujetes à des pressions évolutives impliquant les interactions entre les gènes.

Hypothèse du chapitre 4. La diversité génétique des gènes *CYP450* a été influencée par des pressions de sélection, et certaines de ses sous-familles ont co-évolué ensemble.

Objectif du chapitre 4 :

- Identifier les pressions de sélection à travers les sous-familles de la superfamille des *CYP450* avec des analyses de génétique des populations;
- Identifier les liens co-évolutifs entre les membres des sous-familles de *CYP450* en utilisant la méthodologie utilisée dans le chapitre 2;
- Caractériser les liens co-évolutifs identifiés avec des analyses de transcriptomique et de randomisation mendélienne.

Dans le chapitre 4, nous avons identifié des signatures de potentiels événements de co-évolution dans les sous-familles des *CYP3A* et des *CYP4F* impliquant la régulation d'expression des gènes. De plus, nous avons identifié que la co-évolution dans la famille des CYP3A dans la population africaine pourrait avoir un lien avec le compte des réticulocytes.

1.5.3. Bases de données

Dans le cadre de cette thèse, quatre bases de données ont été utilisées.

1.5.3.1. Projet des 1000 Génomes

Le projet des 1000 génomes (1000G, *The 1000 Genomes Project* en anglais) [252] a été utilisé pour les analyses présentées aux chapitres 2 et 4. Ce projet a pour but de permettre une meilleure compréhension des variants génétiques à travers les populations humaines. Il contient le séquençage du génome de 2504 individus de 26 sous-populations et cinq grandes populations, soit l'Afrique, l'Europe, l'Asie de l'Est, l'Asie du Sud et les Amériques latines.

Les données proviennent de la troisième phase du projet et ont été alignées sur la version de référence GRCh37.

À partir d'un sous-ensemble de ces échantillons, soit 465 échantillons provenant principalement des populations européennes, mais également d'individus de la population africaine, le projet GEUVADIS (*GEnetic EUropean VAriation in DISease*) a été formé [253]. Ce projet comporte des données de séquençage d'ARN provenant de lignée cellulaire lymphoblastoide (LCL, *Lymphoblastoid Cell Line* en anglais).

Les données de ce projet sont fréquemment utilisées par la communauté biomédicale pour cette grande diversité génétique qui permet d'étudier les variants qui sont spécifiques à des populations ou qui pourraient être sous pression de sélection dans certaines populations.

1.5.3.2. CARTaGENE

La base de données de CARTaGENE (CaG) [254] est une biobanque québécoise d'environ 43 000 individus, contenant principalement des individus de descendance européenne. Cette base de données contient des informations génétiques, phénotypiques, mais également des informations détaillées sur leur santé, leur mode de vie, leurs antécédents médicaux et leur environnement. Pour un sous-ensemble de 911 participants, nous avons également des données de séquençage d'ARN dans le sang, et le séquençage du génome entier est présentement en court pour 2184 participants. Elle a été utilisée dans les analyses du chapitre 2.

1.5.3.3. Biobanque du Royaume-Uni

La base de données du Royaume-Uni (UKb, *UK biobank* en anglais) [115] est une des plus grandes bases de données, avec plus de 500 000 individus, principalement de descendance européenne. La base de données contient des données génétiques, mais également des informations détaillées sur les participants, notamment des données démographiques, des antécédents médicaux, des résultats d'examens médicaux et des questionnaires de santé. C'est également une base de donnée longitudinale, signifiant que les informations de plusieurs participants sont mises à jour, incluant les causes de décès s'il y a lieu.

1.5.3.4. GTEx

Le projet d'expression génotype-tissu (GTEx, *Genotype-Tissue Expression* en anglais) [116] a été mis sur pied afin de pouvoir étudier la relation entre les variants génétiques et leur impact à travers divers tissus. Nous avons utilisé la version 8 (v8) de ce jeu de données, qui

contient des échantillons biologiques provenant de 54 tissus de 948 individus, dont 834 avec des données de génotypage, également principalement de descendance européenne. Même si les informations sont limitées, elle contient également certaines informations phénotypiques, incluant certaines maladies, le dossier de santé du patient et la cause de mort après une autopsie. Les donneurs sont des individus décédés et le prélèvement des échantillons a été fait dans un fenêtre de 24h après le décès du donneur. Cette base de données a été utilisée dans tous les chapitres de cette thèse.

Chapitre 2

A sex-specific evolutionary interaction between *ADCY9* and *CETP*

Contributions à ce chapitre

Mes contributions à l'article inclut dans ce chapitre sont les suivantes en tant que première auteure :

- Formulation des hypothèses de recherche
- Planification et réalisation des analyses en génétique des populations, ainsi que l'interprétation de leurs résultats
  - Distributions des fréquences génotypiques
  - Estimation de l'ancestralité génétique avec RFMix
  - Graphiques iHS
  - Génération de l'approche de déséquilibre de liaison sur longue distance (LRLD) afin d'identifier des paires de mutations en co-évolution
  - Génération d'une distribution nulle pour nos analyses de LRLD
  - Stratification des analyses par sexe
  - Établissement d'une collaboration internationnale
- Analyses de transcriptomique
  - Traitement des données d'expression (filtrage, normalisation, calcul des facteurs cachés PEER)
  - Analyses de régression linéaire
- Analyses phénotypiques

    - Effet épistatique entre les mutations sur les événements cardiovasculaires dans la base de donnée de GTEx
- Rédaction de l'intégralité de l'article (introduction, méthodes, résultats, discussion)
- Réalisation de toutes les figures
- Révisions, réponses aux réviseurs et analyses supplémentaires pour parution finale de l'article

Ce travail n'aurait pas été possible sans l'aide des personnes suivantes :

- Julie Hussin
    - Supervision du projet
    - Aide à la conceptualisation du projet
    - Aide à la rédaction de l'article
    - Révisions et corrections de l'article
- Jean-Christophe Grenier
    - Aide à l'obtention des bases de données (1000G, CARTaGENE, GTEx)
    - Traitement de certaines données
    - Aide à la relecture de l'article
- Marie-Pierre Dubé
    - Co-supervision du projet
    - Aide à la conceptualisation du projet
    - Fourniture de diverses ressources (*UK biobank*)
- Jean-Claude Tardif
    - Aide à la conceptualisation du projet
    - Acquisition de ressources biologiques
- Éric Rhéaume et Rocio Sanchez
    - Réalisation des analyses biologiques
    - Fourniture des données d'expression suite au Kock-Down d'*ADCY9*
- Marc-André Legault, Amina Barhdadi et Yassamin Feroz Zada
    - Traitement des données dans la base de données du *UK biobank*
- Samira Asgari, Yang Luo, Leonid Lecca, Megan Murray et Soumya Raychaudhuri
    - Fourniture des données nécessaires pour la réplication dans une deuxième cohorte péruvienne

- Holly Trochet
  - Rédaction des premiers scripts pour l'utilisation de RFMix

A sex-specific evolutionary interaction between *ADCY9* and *CETP*

by

Isabel Gamache[1,2], Marc-André Legault[1,2,3], Jean-Christophe Grenier[1], Rocio Sanchez[1], Eric Rhéaume[1,2], Samira Asgari[4,5], Amina Barhdadi[1,3], Yassamin Feroz Zada[3], Holly Trochet[1,2], Yang Luo[4,5], Leonid Lecca[6,7], Megan Murray[4], Soumya Raychaudhuri[4,5,8,9,10], Jean-Claude Tardif[1,2], Marie-Pierre Dubé[1,2,3], and Julie G. Hussin[1,2]

([1]) Montreal Heart Institute, Montreal, Québec, Canada

([2]) Faculty of Medicine, Université de Montréal, Montreal, Quebec, Canada

([3]) Université de Montréal Beaulieu-Saucier Pharmacogenomics Centre, Montreal, Canada

([4]) Center for Data Sciences, Brigham and Women's Hospital, Harvard Medical School, United States

([5]) Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, United States

([6]) Socios En Salud, Peru

([7]) Harvard Medical School, United States

([8]) Centre for Genetics and Genomics Versus Arthritis, Manchester Academic Health Science Centre, University of Manchester, United Kingdom

([9]) Department of Biomedical Informatics, Harvard Medical School, United States

([10]) Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, United States

This article was published in *eLIFE* : Gamache, I., Legault, M. A., Grenier, J. C., Sanchez, R., Rhéaume, E., Asgari, S., Barhdadi, A., Zada, Y. F., Trochet, H., Luo, Y., Lecca, L., Murray, M., Raychaudhuri, S., Tardif, J. C., Dubé, M. P., & Hussin, J. (2021). A sex-specific evolutionary interaction between *ADCY9* and *CETP*. *eLife*, 10, e69198. https://doi.org/10.7554/eLife.69198

1. Abstract

Pharmacogenomic studies have revealed associations between rs1967309 in the adenylyl cyclase type 9 (*ADCY9*) gene and clinical responses to the cholesteryl ester transfer protein (CETP) modulator dalcetrapib, however, the mechanism behind this interaction is still unknown. Here, we characterized selective signals at the locus associated with the pharmacogenomic response in human populations and we show that rs1967309 region exhibits signatures of positive selection in several human populations. Furthermore, we identified a variant in *CETP*, rs158477, which is in long-range linkage disequilibrium with rs1967309 in the Peruvian population. The signal is mainly seen in males, a sex-specific result that is replicated in the LIMAA cohort of over 3,400 Peruvians. Analyses of RNA-seq data further suggest an epistatic interaction on *CETP* expression levels between the two SNPs in multiple tissues, which also differs between males and females. We also detected interaction effects of the two SNPs with sex on cardiovascular phenotypes in the UK Biobank, in line with the sex-specific genotype associations found in Peruvians at these loci. We propose that *ADCY9* and *CETP* coevolved during recent human evolution due to sex-specific selection, which points towards a biological link between dalcetrapib's pharmacogene *ADCY9* and its therapeutic target CETP.

Keywords: Population genetics, Pharmacogenomics, Transcriptomics, Long-range linkage disequilibrium, Phenotypic association analysis

2. Introduction

Coronary artery disease (CAD) is the leading cause of mortality worldwide. It is a complex disease caused by the accumulation of cholesterol-loaded plaques that block blood flow in the coronary arteries. The cholesteryl ester transfer protein (CETP) mediates the exchange of cholesterol esters and triglycerides between high-density lipoproteins (HDL) and lower density lipoproteins [36, 44]. Dalcetrapib is a CETP modulator that did not reduce cardiovascular event rates in the overall dal-OUTCOMES trial of patients with recent acute coronary syndrome [77]. However, pharmacogenomic analyses revealed that genotypes at rs1967309 in the *ADCY9* gene, coding for the ninth isoform of adenylate cyclase, modulated clinical responses to dalcetrapib [24]. Individuals who carried the AA genotype at rs1967309 in *ADCY9* had less cardiovascular events, reduced atherosclerosis progression, and enhanced

cholesterol efflux from macrophages when treated with dalcetrapib compared to placebo [24, 255]. In contrast, those with the GG genotype had the opposite effects from dalcetrapib. Furthermore, a protective effect against the formation of atherosclerotic lesions was seen only in the absence of both Adcy9 and CETP in mice [89], suggesting an interaction between the two genes. However, the underlying mechanisms linking *CETP* and *ADCY9*, located 50 Mb apart on chromosome 16, as well as the relevance of the rs1967309 non-coding genetic variant are still unclear.

Identification of selection pressure on a genetic variant can help shed light on its importance. Adaptation to different environments often leads to a rise in frequency of variants, by favoring survival and/or reproduction fitness. An example is the lactase gene (LCT) [256, 257, 258, 259, 260], where a positively selected intronic variant in MCM6 leads to an escape from epigenetic inactivation of LCT and facilitates lactase persistence after weaning [261]. Results of genomic studies for phenotypes such as adaptation to high-altitude hypoxia in Tibetans [150], fatty acid metabolism in Inuits [262] or response to pathogens across populations [263] have also been confirmed by functional studies [264, 265, 266, 267, 268]. Thus, population and regulatory genomics can be leveraged to unveil the effect of genetic mutations at a single non-coding locus and reveal the biological mechanisms of adaptation.

When two or more loci interact during adaptation, a genomic scan will likely be underpowered to pinpoint the genetic determinants. In this study, we took a multi-step approach on the *ADCY9* and *CETP* candidate genes to specifically study their interaction (Figure 2.1). We used a joint evolutionary analysis to evaluate the potential signatures of selection in these genes (Step 1), which revealed positive selection pressures acting on *ADCY9*. Sex-specific genetic associations between the two genes are discovered in Peruvians (Step 2), a population in which natural selection for high-altitude was previously found on genes related to cardiovascular health [269]. Furthermore, our knock-down experiments and analyses of large-scale transcriptomics (Step 3) as well as available phenome-wide resources (Step 4) bring further evidence of a sex-specific epistatic interaction between *ADCY9* and *CETP*.

3. Methods

3.1. Key Resources Table

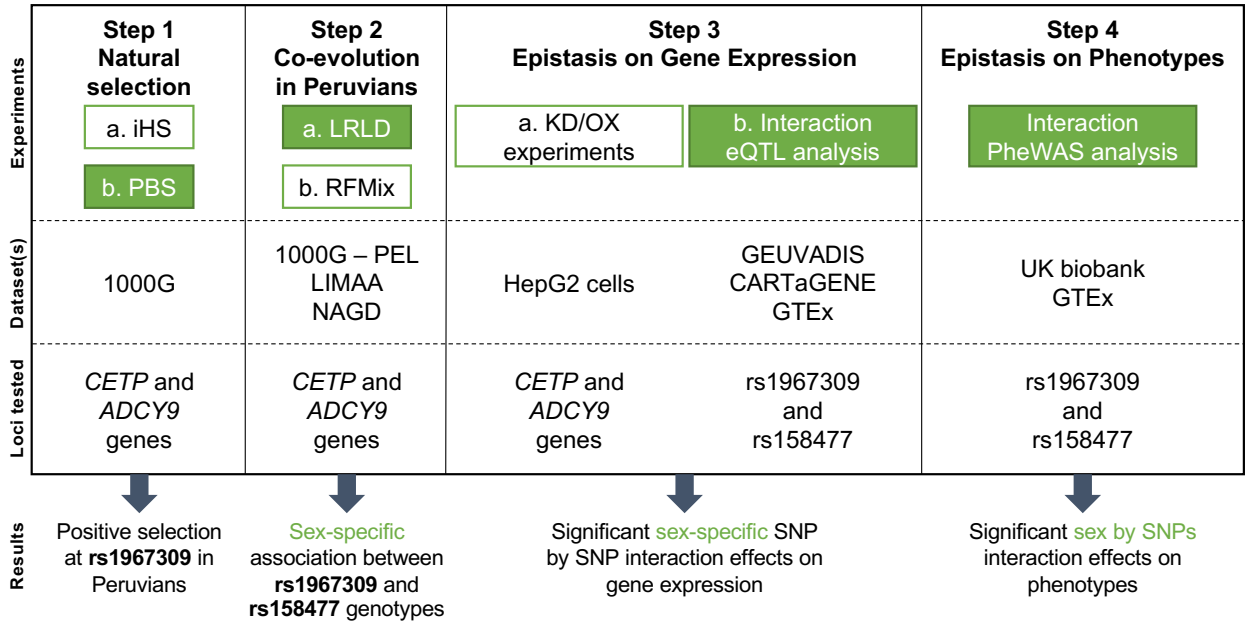

Fig. 2.1. Flowchart of experimental design and main results.
The four main steps of the analyses conducted in this study are reported along with the datasets used for each step and the genetic loci on which the analyses are performed. Green colored boxes represent analyses for which sex is considered.
KD = Knock-down
OX = Overexpression

Table 2.1. Key Resources Table

Reagent type (species) or resource	**Designation**	**Source or reference**	**Identifiers**	**Additional information**
gene (Homo Sapiens)	*CETP*	GenBank	HGNC:1869	
gene (Homo Sapiens)	*ADCY9*	GenBank	HGNC:240	
cell line (Homo Sapiens)	HepG2	ATCC	RRID:CVCL_0027 Hepatoblastoma	
recombinant DNA reagent	pEZ-M46-AC9	plasmid GeneCopoeiaTM	EX-H0609-M46	Methods section
recombinant DNA reagent	pEZ-M50-*CETP* plasmid	GeneCopoeiaTM	EX-C0070-M50	Methods section
antibody	anti-*CETP* (rabbit monoclonal)	Abcam	#ab157183	(1:1000) in 3% BSA, TBS, tween 20 0.5%, O/N 4°C
antibody	Goat anti-rabbit antibody (goat polyclonal)	Abcam	RRID:AB_955447	(1:10 000) in 3% BSA 1h at room temperature
sequence-based reagent	Human *CETP*_F	IDT Technologies	PCR primers	CTACCTGTCTTTCCATAA
sequence-based reagent	Human *CETP*_R	IDT Technologies	PCR primers	CATGATGTTAGAGATGAC
sequence-based reagent	Human *ADCY9*_F	IDT Technologies	PCR primers	CTGAGGTTCAAGAACATCC
sequence-based reagent	Human *ADCY9*_R	IDT Technologies	PCR primers	TGATTAATGGGCGGCTTA
sequence-based reagent	Silencer Select siRNA against human *ADCY9*	Ambion	Cat. #4390826 ID 1039	CCUGAUGAAAGAUUACUUUtt
sequence-based reagent	Silencer Select siRNA against human *CETP*	Ambion	Cat. #4392420 ID 2933	GGACAGAUCUGCAAAGAGAtt
sequence-based reagent	Negative Control siRNA	Ambion	Cat. #4390844	
commercial assay or kit	Lipofectamine RNAiMAX reagent	Invitrogen	Cat. #13778	
commercial assay or kit	Lipofectamine 2000 reagent	Invitrogen	Cat. #11668-019	
commercial assay or kit	RNeasy Plus Mini Kit	Qiagen	Cat. #74136	
commercial assay or kit	High-Capacity cDNA Reverse Transcription Kit	Applied Biosystems	Cat. #4368814	
commercial assay or kit	Agilent RNA 6000 Nano Kit for Bioanalyzer 2100 System	Agilent Technologies	Cat. #5067-1511	
commercial assay or kit	SYBR-Green reaction mix	BioRad	Cat. #1725274	
commercial assay or kit	Amicon Ultra 0.5 ml 10 kDa cutoff units	Millipore Sigma	Cat. #UFC501096	
commercial assay or kit	Western Lightning ECL Pro	Perkin Elmer	Cat. #NEL122001EA	
commercial assay or kit	TGX Stain-Free FastCast Acrylamide 10%	BioRad	Cat #1610183	
software, algorithm	TrimGalore!	DOI : 10.14806/ej.17.1.200	RRID:SCR_011847	
software, algorithm	STAR (v.2.6.1a)	DOI : 10.1093/bioinformatics/bts635	RRID:SCR_019993	
software, algorithm	RSEM (v.1.3.1)	DOI : 10.1186/1471-2105-12-323	RRID:SCR_013027	
software, algorithm	R statistical software (v.3.6.0/v.3.6.1)	https://www.r-project.org/	RRID:SCR_001905	
software, algorithm	FlashPCA2	DOI : 10.1093/bioinformatics/btx299	RRID:SCR_021680	
software, algorithm	Vcftools (v.0.1.17)	DOI : 10.1093/bioinformatics/btr330	RRID:SCR_001235	
software, algorithm	RFMix (v.2.03)	DOI : 10.1016/j.ajhg.2013.06.020		
software, algorithm	PEER	DOI : 10.1038/nprot.2011.457	RRID:SCR_009326	
software, algorithm	pyGenClean (v.1.8.3)	DOI : 10.1093/bioinformatics/btt261		
software, algorithm	SAS (v.9.4)	https://www.sas.com/en_us/ software/stat.html	RRID:SCR_008567	
software, algorithm	EPO pipeline (version e59)	DOI : 10.1093/database/bav096		

Reagent type (species) or resource	**Designation**	**Source or reference**	**Identifiers**	**Additional information**
software, algorithm	Bcftools (v.1.9)	DOI : 10.1093/bioinformatics/btr509	RRID:SCR_005227	
software, algorithm	GenotypeHarmonizer (v.1.4.20)	DOI : 10.1186/1756-0500-7-901		
software, algorithm	Hapbin (v.1.3.0)	DOI : 10.1093/molbev/msv172		
software, algorithm	SHAPEIT2 (r.837)	DOI : 10.1038/nmeth.1785		
software, algorithm	PBWT	DOI : 10.1093/bioinformatics/btu014		
software, algorithm	Beacon designer software (v.8) (Premier Biosoft)	http://www.premierbiosoft.com/qOligo /Oligo.jsp?PID=1		
Other	1000 Genomes project	DOI : 10.1038/nature15393	RRID : SCR_006828	
Other	LIMAA	DOI : 10.1038/s41467-019-11664-1	dbGAP : phs002025.v1.p1	dbgap project #26882
Other	Native American genetic dataset	DOI : 10.1038/nature11258		
Other	GEUVADIS	DOI : 10.1038/nature12531	RRID:SCR_000684	
Other	GTEx (v8)	DOI : 10.1038/ng.2653	RRID:SCR_013042	dbgap project #19088
Other	CARTaGENE biobank	DOI : 10.1093/ije/dys160	RRID:SCR_010614	CAG project number 406713
Other	UK biobank	DOI : 10.1371/journal.pmed.1001779	RRID:SCR_012815	UKB project #15357 and #20168
Other	Sanger Imputation Server	DOI : 10.3389/fgene.2019.00034		

3.2. Population Genetics Datasets

The whole-genome sequencing data from the 1000 Genomes project (1000G) Phase III dataset (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/) was filtered to exclude INDELs and CNVs so that we kept only biallelic SNPs. This database has genomic variants of 2,504 individuals across five ancestral populations: Africans (AFR, n = 661), Europeans (EUR, n = 503), East Asians (EAS, n = 504), South Asians (SAS, n = 489) and Americans (AMR, n = 347) [252]. The replication dataset, LIMAA, has been previously published [270, 271] and was accessed through dbGaP [phs002025.v1.p1, dbgap project #26882]. This cohort was genotyped with a customized Affymetric LIMAAray containing markers optimized for Peruvian-specific rare and coding variants. We excluded related individuals as reported previously [270], resulting in a final dataset of 3,509 Peruvians. We also identified fine-scale population structure in this cohort and a more homogeneous subsample of 3,243 individuals (1,302 females and 1,941 males) in this cohort was kept for analysis (Table 2.2, Supplementary text 9). The Native American genetic dataset (NAGD) contains 2,351 individuals from Native descendants from the data from a previously published study [272]. Individuals were separated by their linguistic families identified by Reich and colleagues [272]. NAGD came under the Hg18 coordinates, so a lift over was performed to transfer to the Hg19 genome coordinates. Pre-processing details for these datasets are described in Supplementary text 9.

Cohort/Subpopulation	Abbreviation	Ethnicity	Sample size (% female)	Age	Reference
1000G – Peruvian	PEL*	Peruvian	85 (52%)	NA	[252]
LIMAA/Peruvian	LIMAA	Peruvian	3,243 (40%)	29.6 ± 13.8	[270, 271]
Native Amerind/Andean	NAGD/AND	Amerind/Peruvian	88 (40%)	NA	[272]
GEUVADIS	GEUVADIS*	European descent	287 (54%)	NA	[253]
CARTaGENE	CaG	European descent	728 (51%)	53.6 ± 8.7	[254]
GTEx	GTEx	European descent	699 (34%)	52.6 ± 13.1	[116]
UK biobank	UKb*	European descent	413,138 (54%)	56.8 ± 8.0	[115]

Table 2.2. Cohort information. Sample sizes are reported after quality control steps.

3.3. eQTL Datasets

We used several datasets (Table 2.2) for which we had both RNA-seq data and genotyping. First, the GEUVADIS dataset [253] for 1000G individuals was used (available at https://www.internationalgenome.org/data-portal/data-collection/geuvadis). A total of

287 non-duplicated European samples (CEU, GBR, FIN, TSI) were kept for analysis. Second, the Genotype-Tissue Expression v8 (GTEx) [116] was accessed through dbGaP (phs000424.v8.p2, dbgap project #19088) and contains gene expression across 54 tissues and 948 donors, genetic and phenotypic information. Phenotype analyses are described in Supplementary text 9. The cohort contains mainly of European descent (84.6%), aged between 20 and 79 years old. Analyses were done on 699 individuals, 66% of males and 34% of females (Supplementary figure 2.25a). Third, we used the data from the CARTaGENE biobank [254] (CAG project number 406713) which includes 728 RNA-seq whole-blood samples with genotype data, from individuals from Quebec (Canada) aged between 36 and 72 years old (Supplementary figure 2.25b). Genotyping and RNA-seq data processing pipelines for these datasets are detailed in Supplementary text 9. To quantify *ADCY9* gene expression, we removed the isoform transcript ENST00000574721.1 (*ADCY9*-205 from the Hg38) from the Gene Transfer Format (GTF) file because it is a "retained intron" and accumulates genomic noise (Supplementary text 9), masking true signals for *ADCY9*. To take into account hidden factors, we calculated PEER factors [273] on the normalized expressions, on all samples and stratified by sex (sPEER factors). To detect eQTL effects, we performed a two-sided linear regression on *ADCY9* and *CETP* expressions using R (v.3.6.0) (https://www.r-project.org/) with the formula $lm(p \sim rs1967309 * rs158477 + Covariates)$ for evaluating the interaction effect, $lm(p \sim rs1967309 + rs158477 + Covariates)$ for the main effect of the SNPs and $lm(p \sim rs1967309 * rs158477 * sex + Covariates)$ for evaluating the three-way interaction effect. Under the additive model, each SNP is coded by the number of non-reference alleles (G for rs1967309 and A for rs158477), under the genotypic model, dummy coding is used with homozygous reference genotype set as reference. The covariates include the first 5 Principal Components (PCs), age (except for GEUVADIS, information not available), sex, as well as PEER factors. We tested the robustness of our results to the inclusion of different numbers of PEER factors in the models and we report them all for GEUVADIS, CARTaGENE and GTEx (Supplementary figures 2.22-2.24). Reported values in the text are for five PEER factors in GEUVADIS, ten PEER factors in CARTaGENE, 25 sPEER for skin sun exposed in male and 10 sPEER for artery tibial in female in GTEx. Covariates specific to each cohort are reported in Supplementary text 9.

3.4. UK biobank processing and selected phenotypes

The UK biobank [115] contains 487,392 genotyped individuals from the UK still enrolled as of August 20th 2020, imputed using the Haplotype Reference Consortium as the main reference panel, and accessed through project #15357 and UKB project #20168. Additional genetic quality control was done using pyGenClean (v.1.8.3) [274]. Variants or individuals with more than 2% missing genotypes were filtered out. Individuals with discrepancies between the self-reported and genetic sex or with aneuploidies were removed from the analysis. We considered only individuals of European ancestry based on PCs, as it is the largest population in the UK Biobank, and because ancestry can be a confounder of the genetic effect on phenotypes. We used the PCs from UK Biobank to define a region in PC space using individuals identified as "white British ancestry" as a reference population. Using the kinship estimates from the UK Biobank, we randomly removed individuals from kinship pairs where the coefficient was higher than 0.0884 (corresponding to a 3rd degree relationship). The resulting post QC dataset included 413,138 individuals. For the reported phenotypes, the date of baseline visit was between 2006 and 2010. The latest available hospitalization records discharge date was June 30th 2020 and the latest date in the death registries was February 14th 2018. We used algorithmically-defined cardiovascular outcomes based on combinations of operation procedure codes (OPCS) and hospitalization or death record codes (ICD9/ICD10). A description of the tested continuous variables can be found in Supplementary table 2.4. We used age at recruitment defined in variable #21022 and sex in variable #31. We ignored self-reported events for cardiovascular outcomes as preliminary analyses suggested they were less precise than hospitalization and death records.

In association models, each SNP analyzed is coded by the number of non-reference alleles, G for rs1967309 and A for rs158477. SNP rs1967309 was also coded as a genotypic variable, to allow for non-additive effects. For continuous traits (Supplementary table 2.4) in the UK Biobank, general two-sided linear models (GLM) were performed using SAS software (v.9.4). A GLM model was first performed using the covariates age, sex and PCs 1 to 10. The externally studentized residuals were used to determine the outliers, which were removed. The normality assumption was confirmed by visual inspection of residuals for most of the outcomes, except birthwt and sleep. For biomarkers and cardiovascular endpoints, regression analyses were done in R (v.3.6.1). Linear regression analyses were conducted on standardized

outcomes and logistic regression was used for cardiovascular outcomes. Marginal effects were calculated using margins package in R. In both cases, models were adjusted for age at baseline and top 10 PCs, as well as sex when not stratified. In models assessing two-way (rs1967309 by rs158477) or three-way (rs1967309 by rs158477 by sex) interactions, we used a 2 d.f. likelihood ratio test for the genotypic dummy variables' interaction terms (genotypic model) (Supplementary text 9).

3.5. RNA-sequencing of *ADCY9*-knocked-down HepG2 cell line

The human liver hepatocellular HepG2 cell line was obtained from ATCC, a cell line derived from the liver tissue of a 15-year-old male donor [275]. Our cells tested negatively for mycoplasma contamination and have a morphology and expression profile concordant with this cell type. Cells were cultured in EMEM Minimum essential Medium Eagle's, supplemented with 10% fetal bovine serum (Wisent Inc). 250 000 cells in 2 ml of medium in a six-well plate were transfected using 12.5 pmol of Silencer Select siRNA against human ADCY9 (Ambion cat #4390826 ID 1039), Silencer Select siRNA against *CETP* (Ambion cat 4392420 ID 2933) or Negative Control siRNA (Ambion cat #4390844) with 5 μl of Lipofectamine RNAiMAX reagent (Invitrogen cat #13778) in 500 μl Opti-MEM I reduced serum medium (Invitrogen cat #31985) for 72h (Supplementary table 2.5, Supplementary text 9). The experiment was repeated five times at different cell culture passages. Total RNA was extracted from transfected HepG2 cells using RNeasy Plus Mini Kit (Qiagen cat #74136) in accordance with the manufacturer's recommendation. Preparation of sequencing library and sequencing was performed at the McGill University Innovation Center. Briefly, ribosomal RNA was depleted using NEBNext rRNA depletion kit. Sequencing was performed using Illumina NovaSeq 6000 S2 paired end 100 bp sequencing lanes. Basic QC analysis of the 10 samples was performed by the Canadian Centre for Computational Genomics (C3G). To process the RNA-seq samples, we first performed read trimming and quality clipping using TrimGalore! [276] (https://github.com/FelixKrueger/TrimGalore), we aligned the trimmed reads on the Hg38 reference genome using STAR (v.2.6.1a) and we ran RSEM (v.1.3.1) on the transcriptome aligned libraries. Prior to normalization with limma and voom, we filtered out genes which had less than 6 reads in more than 5 samples. For *ADCY9* and *CETP* gene-level differential expression analyses, we compared the mean of each group of replicates

with a t-test for paired samples. The transcriptome-wide differential expression analysis was done using limma, on all genes having an average of at least 10 reads across samples from a condition. Samples were paired in the experiment design. The multiple testing was taken into account by correcting the p-values with the `qvalue` R package (v.4.0.0) [277], to obtain transcriptome-wide FDR values.

3.6. Overexpression of *ADCY9* and *CETP* genes in HepG2 cell line

For *ADCY9* and *CETP* overexpression experiments, 500 000 cells in 2 ml of medium in a six-well plate were transfected using 1 ug of pEZ-M46-AC9 or pEZ-M50-*CETP* plasmids (GeneCopoeiaTM) with 5 μl of Lipofectamine 2000 reagent (Invitrogen cat #11668-019) for 72h. Total RNA was extracted from transfected HepG2 cells using RNeasy Plus Mini Kit (Qiagen cat #74136) in accordance with the manufacturer's recommendation (Supplementary table 2.5, Supplementary text 9).

3.7. Natural selection analyses

We used the integrated Haplotype Statistic (iHS) [153] and the population branch statistic (PBS) [150] to look for selective signatures. The iHS values were computed for the each 1000G population. An absolute value of iHS above 2 is considered to be a genome wide significant signal [153]. Prior to iHS computation, ancestral alleles were retrieved from 6 primates using the EPO pipeline (version e59) [278] and the filtered 1000 Genomes vcf files were converted to change the reference allele as ancestral allele using bcftools [279] with the fixref plugin. The hapbin program (v.1.3.0) [280] was then used to compute iHS using per population-specific genetic maps computed by Adam Auton on the 1000G OMNI dataset (ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20130507_omni_recombination_rates). When the genetic map was not available for a subpopulation, the genetic map from the closest sub-population was selected according to their global FST value computed on the phase 3 dataset.

We scanned the *ADCY9* and *CETP* genes using the population branch statistic (PBS), using 1000G sub-populations data. PBS summarizes a three-way comparison of allele frequencies between two closely related populations, and an outgroup. The grouping we focused on was PEL/MXL/CHB, with PEL being the focal population to test if allele frequencies

are especially differentiated from those in the other populations. The CHB population was chosen as an outgroup to represent a Eurasian population that share common ancestors in the past with the American populations, after the out-of-Africa event. Using PJL (South Asia) and CEU (Europe) as an outgroup, or CLM as a closely related population (instead of MXL) yield highly similar results. To calculate FST for each pair of population in our tree, we used vcftools [281] by subpopulation. We calculated normalized PBS values as in [269], which adjusts values for positions with large branches in all populations, for the whole genome. We use this distribution to define an empirical threshold for significance based on the 95th percentile of all PBS values genome-wide for each of the three populations.

3.8. Long-range linkage disequilibrium

Long-range linkage disequilibrium (LRLD) was calculated using the function geno-r2 of vcftools (v.0.1.17) which uses the genotype frequencies. LRLD was evaluated in all subpopulations from 1000 Genomes Project Phase III, in LIMAA and NAGD, for all biallelic SNPs in *ADCY9* (chr16:4,012,650-4,166,186 in Hg19 genome reference) and *CETP* (chr16:56,995,835-57,017,756 in Hg19 genome reference). We analyzed loci from the phased VCF files that had a MAF of at least 5% and a missing genotype of at most 10%, in order to retain a maximum of SNPs in NAGD which has higher missing rates than the others. We extracted the 99th percentile of all pairs of comparisons between *ADCY9* and *CETP* genes to use as a threshold for empirical significance and we refer to these as *ADCY9*/*CETP* empirical p-values. In LIMAA, we also evaluated the genotypic association using a X^2 test with four degrees of freedom (X^2_4) using a permutation test, as reported in [159] (Supplementary text 9).

Furthermore, for both cohorts, we created a distribution of LRLD values for random pairs of SNPs across the genome to obtain a genome-wide null distribution of LRLD to evaluate how unusual the genotypic association between our candidate SNPs (rs1967309-rs158477) is while taking into account the cohort-specific background genomic noise/admixture and allele frequencies. We extracted 3,513 pairs of SNPs that match rs1967309 and rs158477 in terms of MAF, physical distance (in base pairs) and genetic distance (in centiMorgans (cM), based on the PEL genetic map) between them in both cohorts (Supplementary text 9), and report genome-wide empirical p-values based on this distribution. For the analyses of LRLD between *ADCY9* and *CETP* stratified by sex, we considered the same set of SNP pairs that

we used for the full cohorts, but separated the dataset by sex before calculating the LRLD values. To evaluate how likely the differences observed in LRLD between sex are, we also performed permutations of the sex labels across individuals to create a null distribution of sex specific effects (Supplementary text 9).

3.9. Local ancestry inference

To evaluate local ancestry in the PEL subpopulation and in the LIMAA cohort, we constructed a reference panel using the phased haplotypes from 1000 Genomes (YRI, CEU, CHB) and the phased haplotypes of NAGD (Northern American, Central American and Andean) (Supplementary text 9). We kept overlapping positions between all datasets, and when SNPs had the exact same genetic position, we kept the SNP with the highest variance in allele frequencies across all reference populations (Supplementary text 9). We ran RFMix (v.2.03) [157] (with the option 'reanalyze-reference' and for 25 iterations) on all phased chromosomes. We estimated the whole genome average proportion of each ancestry using a weighted mean of the chromosome specific proportions given by RFMix based on the chromosome size in cM. For comparing the overall Andean enrichment inferred by RFMix between rs1967309/rs158477 genotype categories, we used a two-sided Wilcoxon-t-test. To evaluate the Andean local ancestry enrichment specifically at *ADCY9* and *CETP*, we computed the genome-wide 95th percentile for proportion of Andean attribution for all intervals given by RFMix.

3.10. Code and source data

Numerical summary data represented as a graph in main figures, as well as the code to reproduce figures and analyses, can be found here. Raw RNA sequencing data for knocked down experiments in hepatocyte HepG2 cells are deposited the data on NCBI Gene Expression Omnibus, accession number GSE174640.

4. Results

4.1. Signatures of selection at rs1967309 in *ADCY9* in human populations

The genetic variant rs1967309 is located in intron 2 of *ADCY9*, in a region of high linkage disequilibrium (LD), in all subpopulations in the 1000 Genomes Project (1000G), and harbors heterogeneous genotype frequencies across human populations (Figure 2.2a). Its intronic location makes it difficult to assess its functional relevance but exploring selective signals around intronic SNPs in human populations can shed light on their importance. In African populations (AFR), the major genotype is AA, which is the homozygous genotype for the ancestral allele, whereas in Europeans (EUR), AA is the minor genotype. The frequency of the AA genotype is slightly higher in Asia (EAS, SAS) and America (AMR) compared to that in Europe, becoming the most frequent genotype in the Peruvian population (PEL). Using the integrated haplotype score (iHS) [153] (Step 1a, Figure 2.1), a statistic that enables the detection of evidence for recent strong positive selection (typically when $|iHS| > 2$), we observed that several SNPs in the LD block around rs1967309 exhibit selective signatures in non-African populations ($|iHS_{SAS}|= 2.66$, $|iHS_{EUR}|= 2.31$), whereas no signal is seen in this LD block in African populations (Figure 2.2b, Supplementary figure 2.16, Supplementary text 9). Our analyses suggest that this locus in *ADCY9* has been the target of recent positive selection in several human populations, with multiple, possibly independent, selective signals detectable around rs1967309. However, recent positive selection as measured by iHS does not seem to explain the notable increase in frequency for the A allele in the PEL population (f_A=0.77), compared to the European (f_A=0.41), Asian (f_A=0.44), and other American populations (f_A=0.54 in AMR without PEL).

To test whether the difference between PEL and other AMR allele frequencies at rs1967309 is significant, we used the population branch statistic (PBS) (Step 1b, Figure 2.1). This statistic has been developed to locate selection signals by summarizing differentiation between populations using a three-way comparison of allele frequencies between a specific group, a closely related population, and an outgroup [150]. It has been shown to increase power to detect incomplete selective sweeps on standing variation. Applying this statistic to investigate rs1967309 allele frequency in PEL, we used Mexicans (MXL) as a closely

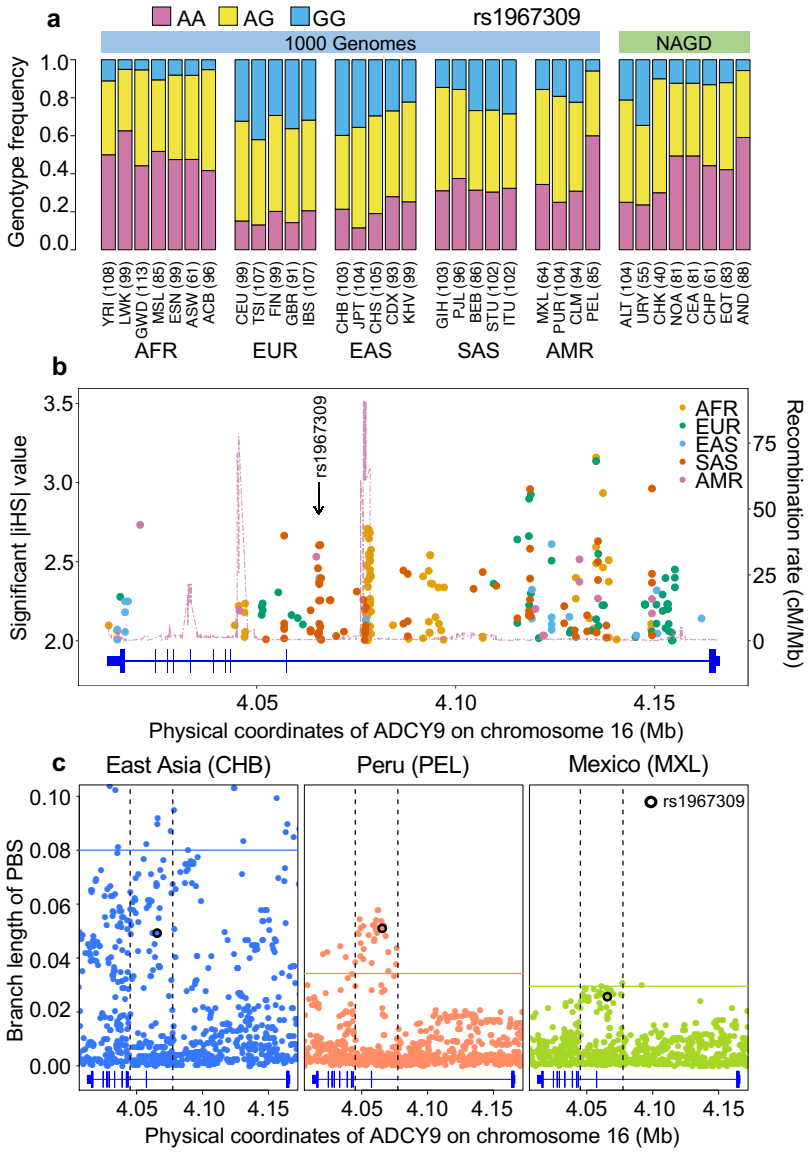

Fig. 2.2. Natural selection signature at rs1967309 in *ADCY9*.
(a) Genotype frequency distribution of rs1967309 in populations from the 1000 Genomes (1000G) Project and in Native Americans (NAGD). (b) Significant iHS values (absolute values above 2) for 1000G continental populations and recombination rates from AMR-1000G population-specific genetic maps, in the *ADCY9* gene. c) PBS values in the *ADCY9* gene, in CHB (outgroup, left panel), PEL (middle panel) and MXL (right panel). Horizontal lines represent the 95th percentile PBS value genome-wide for each population. Vertical dotted black lines define the LD block around rs1967309 (black circle) from 1000G population-specific genetic maps. Gene plots for *ADCY9* showing location of its exons are presented in blue below each plot. Abbreviations: Altaic from Mongolia and Russia: ALT; Uralic Yukaghir from Russia: URY; Chukchi Kamchatkan from Russia: CHK; Northern American from Canada, Guatemala and Mexico: NOA; Central American from Costal Rica and Mexico: CEA; Chibchan Paezan from Argentina, Bolivia, Colombia, Costa Rica and Mexico: CHP; Equatorial Tucanoan from Argentina, Brazil, Colombia, Gualana and Paraguay: EQT; Andean from Bolivia, Chile, Colombia and Peru: AND. For 1000G populations, abbreviations can be found here https://www.internationalgenome.org/category/population/.

related group and a Chinese population (CHB) as the outgroup (Methods). Over the entire genome, the CHB branches are greater than PEL and MXL branches ($mean_{CHB}$=0.020, $mean_{MXL}$=0.008, $mean_{PEL}$=0.009), which reflects the expectation under genetic drift. However, the estimated PEL branch length at rs1967309 (Figure 2.2c), which reflects differentiation since the split from the MXL population ($PBS_{PEL,rs1967309}$=0.051, empirical p-value = 0.014), surpasses the CHB branch length ($PBS_{CHB,rs1967309}$=0.049, empirical p-value > 0.05), which reflects differentiation since the split between Asian and American populations, whereas no such effect is seen in MXL ($PBS_{MXL,rs1967309}$=0.026, empirical p-value > 0.05), or for any other AMR populations. Furthermore, the PEL branch lengths at several SNPs in this LD block (Figure 2.2c) are in the top 5% of all PEL branch lengths across the whole genome ($PBS_{PEL,95th}$ = 0.031), whereas these increased branch lengths are not observed outside of the LD block (Figure 2.2c). These results are robust to the choice of the outgroup and the closely related AMR population (Methods).

The increase in frequency of the A allele at rs1967309 is also seen in genotype data from Native American populations [272], with Andeans showing genotype frequencies highly similar to PEL (f_A=0.77, Figure 2.2a). The PEL population has a large Andean ancestry (Methods, Supplementary figure 2.17a,b) and almost no African ancestry, strongly suggesting that the increase in AA genotype arose in the Andean population and not from admixture with Africans. The PEL individuals that harbor the AA genotype for rs1967309 do not exhibit a larger genome-wide Andean ancestry than non-AA individuals (p-value=0.30, Mann-Whitney U test). Overall, these results suggest that the ancestral allele A at rs1967309, after dropping in frequency following the out-of-Africa event, has increased in frequency in the Andean population and has been preferentially retained in the Peruvian population's genetic makeup, potentially because of natural selection.

4.2. Evidence for co-evolution between *ADCY9* and *CETP* in Peru

The pharmacogenetic link between *ADCY9* and the *CETP* modulator dalcetrapib raises the question of whether there is a genetic relationship between rs1967309 in *ADCY9* and *CETP*, both located on chromosome 16. Such a relationship can be revealed by analyzing patterns of long-range linkage disequilibrium (LRLD) [158, 159], in order to detect whether specific combinations of alleles (or genotypes) at two loci are particularly overrepresented.

To do so, we calculated the genotyped-based linkage disequilibrium (r^2) (Step 2a, Figure 2.1) between rs1967309 and each SNP in *CETP* with minor allele frequency (MAF) above 0.05. In the Peruvian population, there are four SNPs, (including 2 in perfect LD in PEL) that exhibit r^2 values with rs1967309 that are in the top 1% of r^2 values (Figure 2.3a) computed for all 37,802 pairs of SNPs in *ADCY9* and *CETP* genes with MAF>0.05 (Methods). Despite the r^2 values themselves being low ($r^2_{rs158477}$=0.080, $r^2_{rs158480;rs158617}$=0.089, $r^2_{rs12447620}$=0.090), these values are highly unexpected for these two genes situated 50 Mb apart (*ADCY9*/*CETP* empirical p-value<0.006, Supplementary table 2.3) and thus correspond to a significant LRLD signal. This signal is not seen in other 1000G populations (Supplementary table 2.3). We also computed r^2 between the four identified SNPs' genotypes and all *ADCY9* SNPs with MAF above 0.05 (Figure 2.3b). The distribution of r^2 values for the rs158477 *CETP* SNP shows a clear bell-shaped pattern around rs1967309 in *ADCY9*, which strongly suggests the rs1967309-rs158477 genetic association detected is not simply a statistical fluke, while the signal in the region for the other SNPs is less conclusive. The SNP rs158477 in *CETP* is also the only one that has a PEL branch length value higher than the 95^{th} percentile, also higher than the CHB branch length value ($PBS_{PEL,rs158477}$= 0.062, Supplementary figure 2.18a), in line with the observation at rs1967309. Strikingly, this *CETP* SNP's genotype frequency distribution across the 1000G and Native American populations resembles that of rs1967309 in *ADCY9* (Figure 2.3c).

Given that the Peruvian population is admixed [282], particular enrichment of genome segments for a specific ancestry, if present, would lead to inflated LRLD between these segments [155, 283, 284, 285], we thus performed several admixture-related analyses (Step 2b, Figure 2.1). No significant enrichment is seen at either locus and significant LRLD is also seen in the Andean source population (Figure 2.4a,b, Supplementary text 9). Furthermore, we see no enrichment of Andean ancestry in individuals harboring the overrepresented combination of genotypes, AA at rs1967309 + GG at rs158477, compared to other combinations (p-value=0.18, Mann-Whitney U test). These results show that admixture patterns in PEL cannot be solely responsible for the association found between rs1967309 and rs158477. Finally, using a genome-wide null distribution which allows to capture the LRLD distribution expected under the admixture levels present in this sample (Supplementary text 9), we show that the r^2 value between the two SNPs is higher than expected given their allele

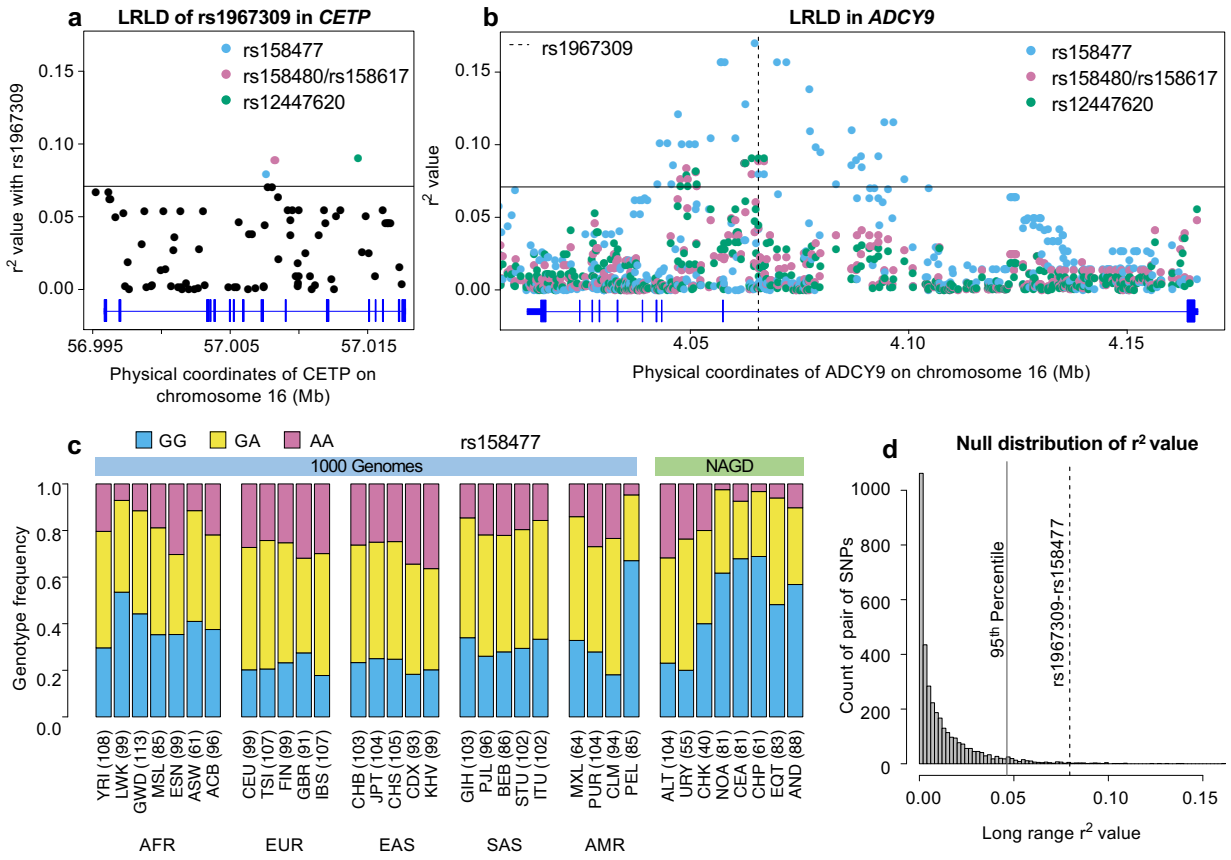

Fig. 2.3. Long-range linkage disequilibrium between rs1967309 and rs158477 in Peruvians from Lima, Peru.

(a) Genotype correlation (r^2) between rs1967309 and all SNPs with MAF>5% in *CETP*, for the PEL population. (b) Genotype correlation between the 3 loci identified in (a) to be in the 99th percentile and all SNPs with MAF>5% in *ADCY9*. The dotted line indicates the position of rs1967309. The horizontal lines in (a,b) represent the threshold for the 99th percentile of all comparisons of SNPs (MAF>5%) between *ADCY9* and *CETP*. Figure 2.4 presents the same plots for Andeans and in the replication cohort (LIMAA) and Figure 2.5 compares the r^2 values between PEL and LIMAA (c) Genotype frequency distribution of rs158477 in 1000G and Native American populations. (d) Genomic distribution of r^2 values from 3,513 pairs of SNPs separated by between 50-60 Mb and 61±10 cM away across all Peruvian chromosomes from the PEL sample, compared to the rs1967309-rs158477 r^2 value (dotted grey line) (genome-wide empirical p-value=0.01). The vertical black line shows the threshold for the 95th percentile threshold of all pairs. Gene plots showing location of exons for *CETP* (a) and *ADCY9* (b) are presented in blue below each plot. Abbreviations: Altaic from Mongolia and Russia: ALT; Uralic Yukaghir from Russia: URY; Chukchi Kamchatkan from Russia: CHK; Northern American from Canada, Guatemala and Mexico: NOA; Central American from Costal Rica and Mexico: CEA; Chibchan Paezan from Argentina, Bolivia, Colombia, Costa Rica and Mexico: CHP; Equatorial Tucanoan from Argentina, Brazil, Colombia, Gualana and Paraguay: EQT; Andean from Bolivia, Chile, Colombia and Peru: AND. For 1000G populations, abbreviations can be found here https://www.internationalgenome.org/category/population/

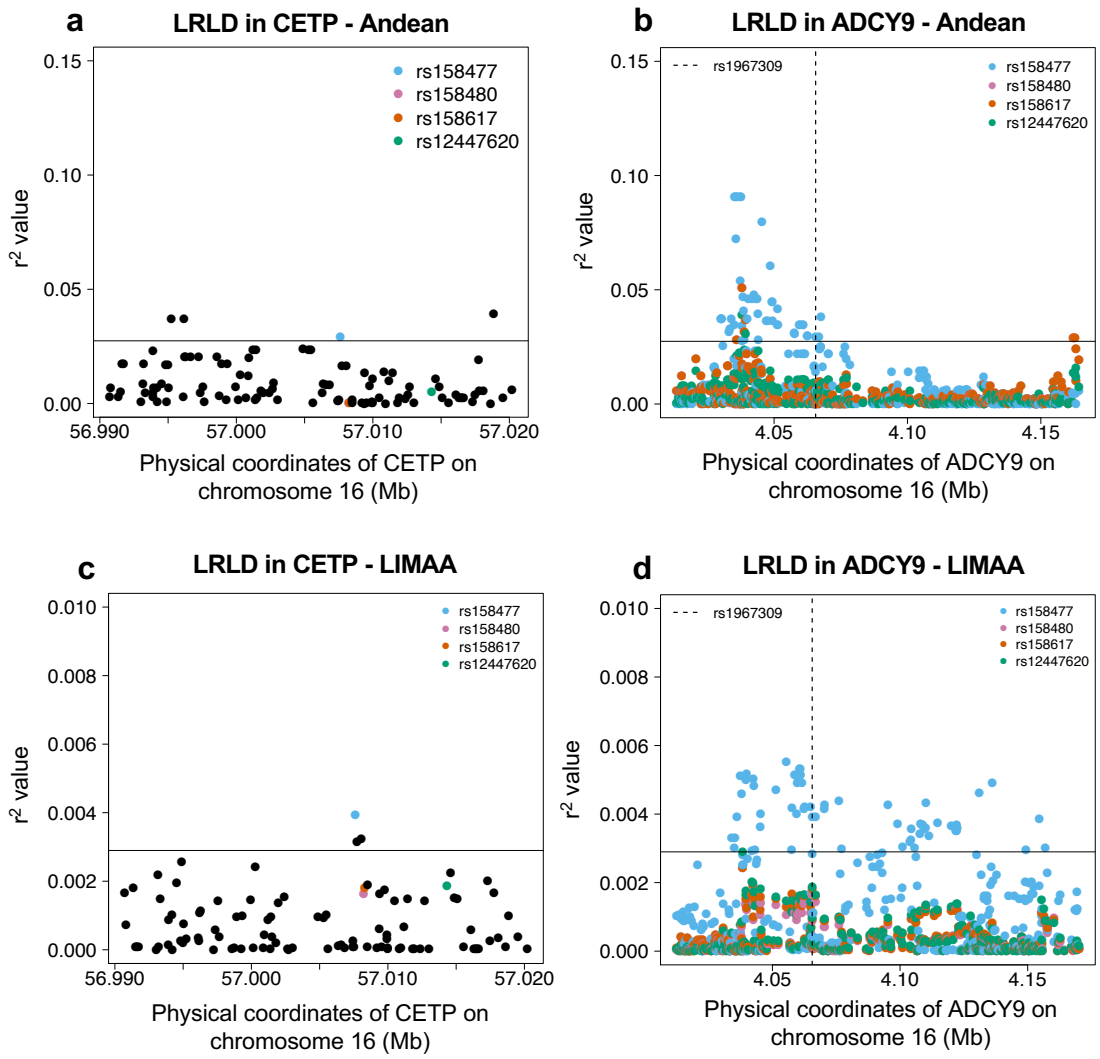

Fig. 2.4. Long-range linkage disequilibrium in the Andean population from the Native Population (n=88) (a,b) and in the LIMAA cohort (n=3243) (c,d).
(a,c) Genotype correlation (r^2) between rs1967309 and all SNPs with MAF>5% in *CETP*. (b,d) Genotype correlation between the 3 loci identified in Figure 2.3a to be in the 99th percentile and all SNPs with MAF>5% in *ADCY9*. The dotted line indicates the position of rs1967309. The horizontal lines represent the threshold for the 99th percentile of all comparisons of SNPs (MAF>5%) between *ADCY9* and *CETP*.

frequencies and the physical distance between them (genome-wide empirical p-value=0.01, Figure 2.3d). Taken together, these findings strongly suggest that the AA/GG combination is being transmitted to the next generation more often (ie. is likely selectively favored) which reveals a signature of co-evolution between *ADCY9* and *CETP* at these loci.

Still, such a LRLD signal can be due to a small sample size [284]. To confirm independently the association between genotypes at rs1967309 of *ADCY9* and rs158477 of *CETP*,

we used the LIMAA cohort [270, 271], a large cohort of 3,509 Peruvian individuals with genotype information, to replicate our finding. The ancestry distribution, as measured by RFMix (Methods) is similar between the two cohorts (Supplementary figure 2.17a,b), however, the LIMAA cohort population structure shows additional subgroups compared to the 1000G PEL population sample (Supplementary figure 2.17c-e): to limit confounders, we excluded individuals coming from these subgroups (Supplementary text 9). In this dataset (N=3,243), the pair of SNPs rs1967309-rs158477 is the only pairs identified in PEL who shows evidence for LRLD, with an r^2 value in the top 1% of all pairs of SNPs in *ADCY9* and *CETP* (*ADCY9*/*CETP* empirical p-value=0.003, Figures 2.4c,d and 2.5, Supplementary table 2.3). The r^2 test used above is powerful to detect allelic associations, but the net association measured will be very small if selection acts on a specific genotype combination rather than on alleles. In that scenario, and when power allows it, the genotypic association is better assessed by with a X^2 distributed test statistic (with four degrees of freedom, X^2_4) comparing the observed and expected genotype combination counts [159]. The test confirmed the association in LIMAA (X^2_4=82.0, permutation p-value <0.001, genome-wide empirical p-value=0.0003, Supplementary text 9). The association discovered between rs1967309 and rs158477 is thus generalizable to the Peruvian population and not limited to the 1000G PEL sample.

4.3. Sex-specific long-range linkage disequilibrium signal

Because the allele frequencies at rs1967309 were suggestively different between males and females (Figure 2.7), we performed sex-stratified PBS analyses, which suggested that the LD block around rs1967309 is differentiated between sexes in the Peruvians (Figure 2.8, Supplementary text 9). We therefore explored further the effect of sex on the LRLD association found between rs1967309 and rs158477 and performed sex-stratified LRLD analyses. These analyses revealed that the correlation between rs1967309 and rs158477 is only seen in males in PEL (Figure 2.6a,b, Supplementary figure 2.19a,b, Supplementary table 2.3): the r^2 value rose to 0.348 in males (*ADCY9*/*CETP* empirical p-value=8.23 x 10^{-5}, genome-wide empirical p-value<2.85 x 10^{-4}, N=41) and became non-significant in females (*ADCY9*/*CETP* empirical p-value=0.78, genome-wide empirical p-value=0.80, N=44). In the Andean population, the association between rs1967309 and rs158477 is not significant when we stratified

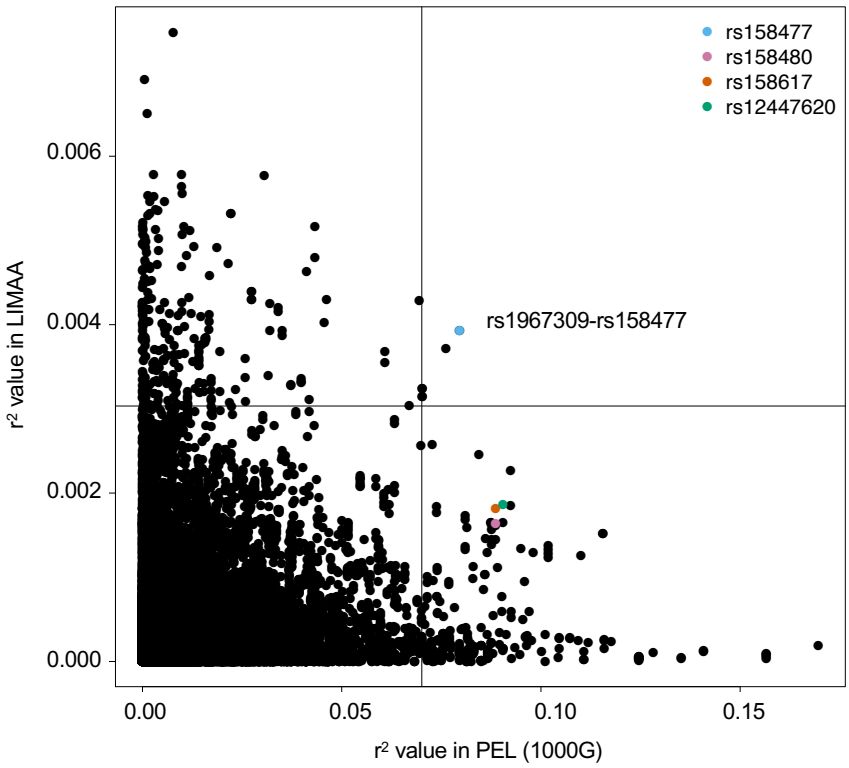

Fig. 2.5. Comparison of genotype correlation between Peruvian from 1000G and from the LIMAA cohort.
Comparison of genotype correlation (r^2) between all SNPs in *ADCY9* and *CETP* with MAF>5% in the Peruvian population (PEL) in 1000G (x axis) and LIMAA cohort (y axis). Colored dots represent the value for SNPs higher than the 99th percentile with rs1967309 in PEL identified in Figure2.3a. Black lines represent the 99th percentile in both populations.

by sex (Supplementary table 2.3), but we still see significant association signals with rs158477 at other SNPs in *ADCY9* LD block in both sexes (Figure 2.9). The LRLD result in PEL cannot be explained by differences of Andean ancestry proportion between males and females (p-value=0.27, Mann-Whitney U test). A permutation analysis that shuffled the sex labels of samples established that the observed difference between the sexes is larger than what we expect by chance (p-value=0.002, Supplementary figure 2.19c, Supplementary text 9). In the LIMAA cohort, we replicate this sex-specific result (Figure 2.6c,d, Supplementary table 2.3) where the r^2 test is significant in males (*ADCY9*/*CETP* empirical p-value=0.003, N=1,941) but not in females (*ADCY9*/*CETP* empirical p-value=0.52, N=1,302). The genotypic X_4^2 test confirms the association between *ADCY9* and *CETP* is present in males ($X_4^2 = 56.6$, permutation p-value=0.001, genome-wide empirical p-value=0.002, Supplementary text 9),

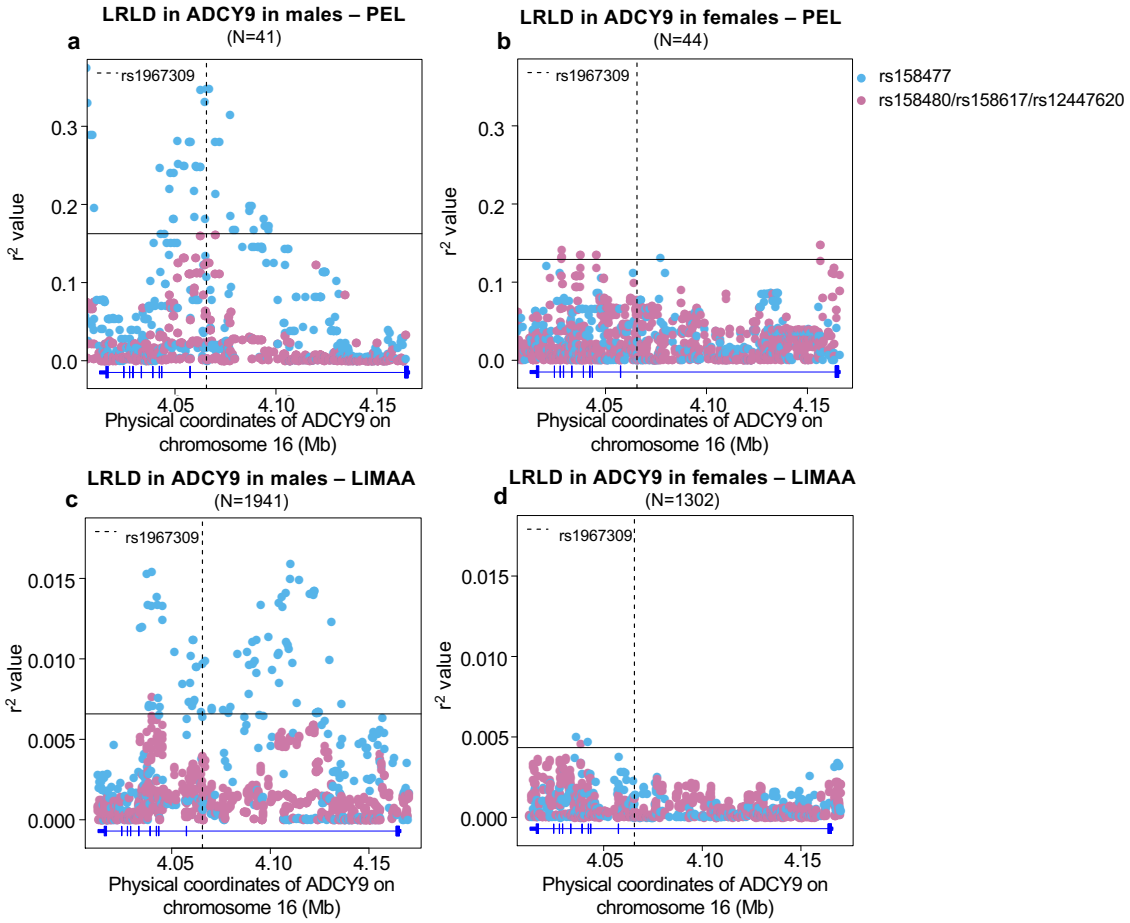

Fig. 2.6. Sex-specific long-range linkage disequilibrium.
Genotype correlation between the loci identified in *CETP* in Figure 2.3a and all SNPs with MAF>5% in *ADCY9* for (a,b) the PEL population and (c,d) LIMAA cohort in males (a,c) and in females (b,d). Genotype frequencies per sex are shown in Figure 2.7 and sex-specific PBS values in Figure 2.8. The horizontal line shows the threshold for the 99th percentile of all comparisons of SNPs (MAF>5%) between *ADCY9* and *CETP*. The vertical dotted line represents the position of rs1967309. Blue dots represent the rs158477 SNPs and pink represents the other three SNPs identified in Figure 2.3a (rs158480, rs158617 and rs12447620), which are in near-perfect LD. Figure 2.9 shows the same analysis in Andeans from NAGD. Gene plots for *ADCY9* showing location of its exons are presented in blue below each plot.

revealing an excess of rs1967309-AA + rs158477-GG. This is also the genotype combination driving the LRLD in PEL. In females, the test also shows a weaker but significant effect ($X^2_4 = 37.0$, permutation p-value = 0.017, genome-wide empirical p-value = 0.001) driven by an excess of a different genotype combination, rs1967309-AA + rs158477-AA, which is, however, not replicated in PEL possibly because of lack of power (Supplementary text 9).

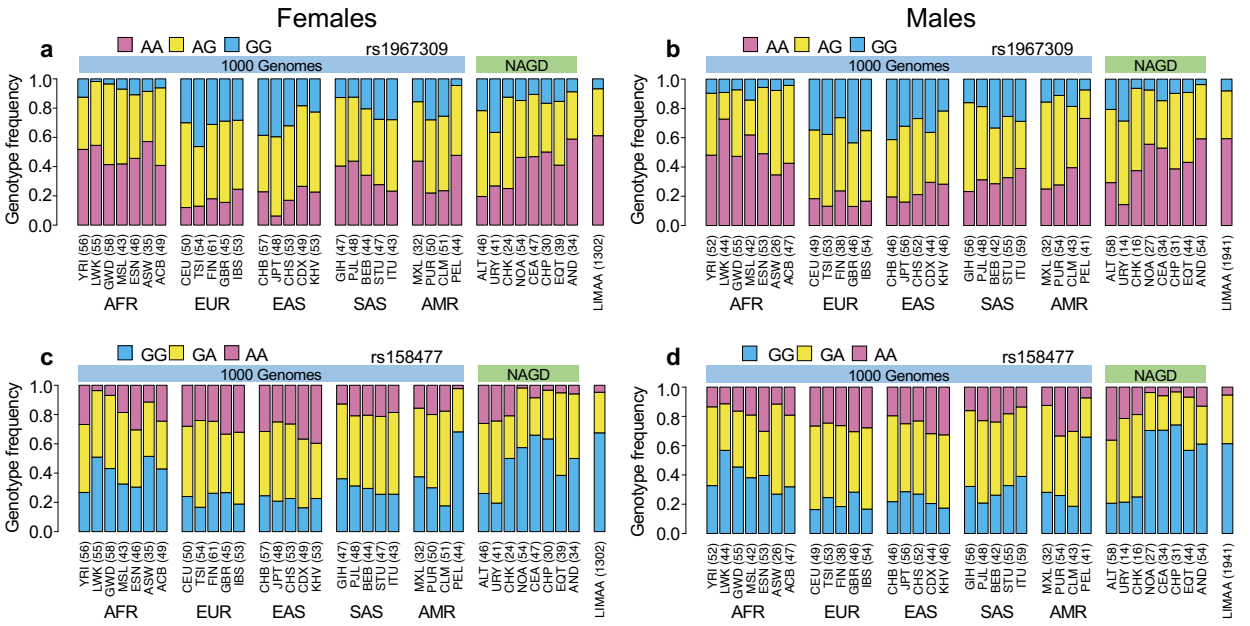

Fig. 2.7. Genotype frequency distribution per sex.
Genotype frequency distribution of rs1967309 in *ADCY9* (a,b) and rs158477 in *CETP* (c,d) in populations from the 1000 Genomes (1000G) Project, in Native Americans (NAGD) and LIMAA cohorts, in females (a,c) and males (b,d). Abbreviations: Altaic from Mongolia and Russia: ALT; Uralic Yukaghir from Russia: URY; Chukchi Kamchatkan from Russia: CHK; Northern American from Canada, Guatemala and Mexico: NOA; Central American from Costal Rica and Mexico: CEA; Chibchan Paezan from Argentina, Bolivia, Colombia, Costa Rica and Mexico: CHP; Equatorial Tucanoan from Argentina, Brazil, Colombia, Gualana and Paraguay: EQT; Andean from Bolivia, Chile, Colombia and Peru: AND. For 1000G populations, abbreviations can be found here https://www.internationalgenome.org/category/population/.

4.4. Epistatic effects on *CETP* gene expression

LRLD between variants can suggest the existence of gene-gene interactions, especially if they are functional variants [284]. In order to be under selection, mutations typically need to modulate a phenotype or an endophenotype, such as gene expression. We have shown previously [89] that CETP and Adcy9 interact in mice to modulate several phenotypes, including atherosclerotic lesion development. To test whether these genes interact in humans, we knocked down (KD) *ADCY9* in hepatocyte HepG2 cells (Step 3a, Figure 2.1) and performed RNA sequencing on five KD biological replicates and five control replicates, to evaluate the impact of decreased *ADCY9* expression on the transcriptome. We confirmed the KD was successful as *ADCY9* expression is reduced in the KD replicates (Figure 2.10a), which represents a drastic drop in expression compared to the whole transcriptome changes

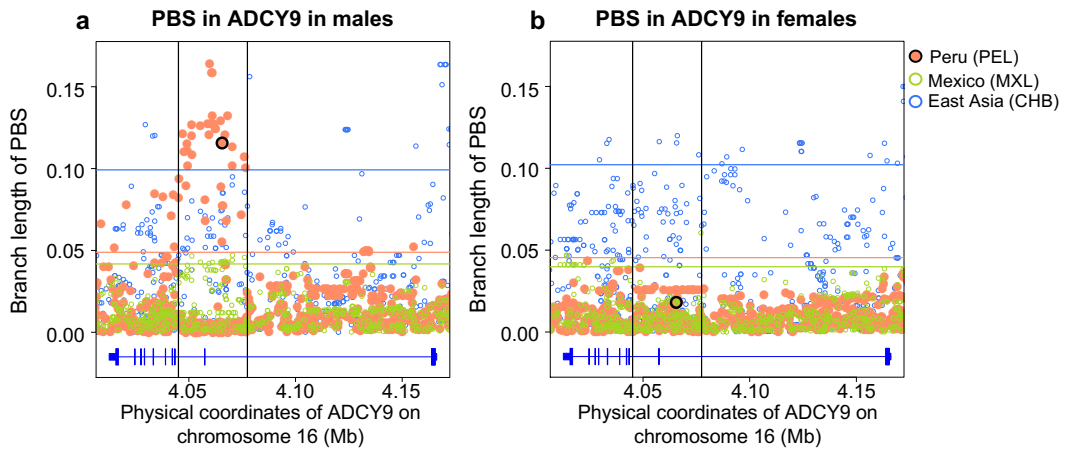

Fig. 2.8. PBS values in the *ADCY9* per sex, comparing the CHB (outgroup), MXL and PEL.
Horizontal lines represent the 95th percentile PBS value of the chromosome 16 for each population for each sex. Vertical black lines represent the LD block around rs1967309 (shown as a black circle for PEL). Gene plots for *ADCY9* showing location of its exons are presented in blue below each plot.

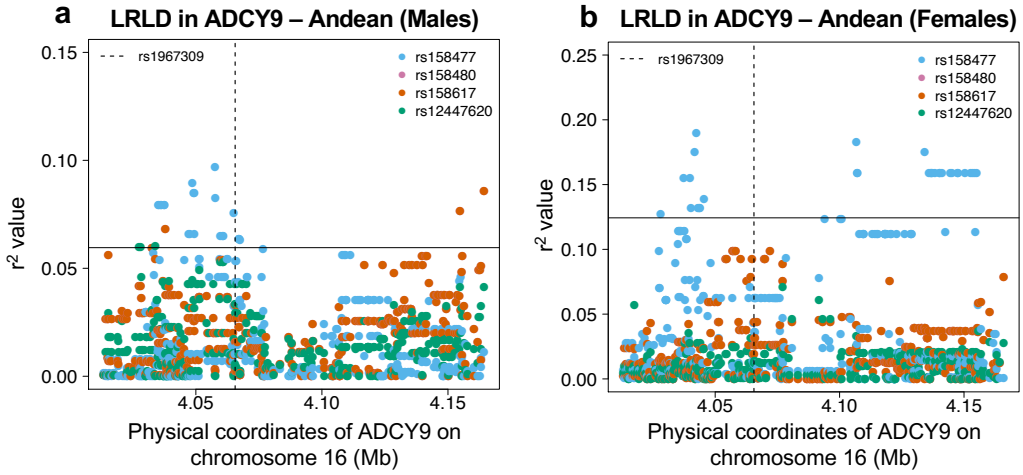

Fig. 2.9. Sex-specific long-range linkage disequilibrium in the Andean population (NAGD). Genotype correlation between the loci identified in *CETP* in Figure 2.3a and all SNPs with MAF>5% in *ADCY9* for the Andean population, in males (N=54) and in females (N=34). The horizontal line shows the threshold for the 95th percentile of all comparisons of SNPs (MAF>5%) between *ADCY9* and *CETP*. The vertical dotted line represents the position of rs1967309.

(False Discovery Rate [FDR] = 4.07 x 10^{-14}, Methods). We also observed that *CETP* expression was increased in *ADCY9*-KD samples compared to controls (Figure 2.10a), an increase that is also transcriptome-wide significant (FDR=1.97 x 10^{-7}, β = 1.257). This increased expression was validated by qPCR, and western blot also showed increased CETP protein

product (Methods, Figure 2.11a,b, Supplementary text 9), but its overexpression did not significatively modulate *CETP* expression (Figure 2.11c). Knocking down or overexpressing *CETP* did not impact *ADCY9* expression on qPCR (Figure 2.11 d,e). These experiments demonstrate an interaction between *ADCY9* and *CETP* at the gene expression level and raised the hypothesis that *ADCY9* potentially modulates the expression of *CETP* through a genetic effect mediated by rs1967309.

To test for potential interaction effects between rs1967309 and *CETP*, we used RNA-seq data from diverse projects in humans: the GEUVADIS project [253], the Genotype-Tissue Expression (GTEx v8) project [116] and CARTaGENE (CaG) [254] (Step 3b, Figure 2.1). When looking across tissues in GTEx, *ADCY9* and *CETP* expressions negatively correlate in almost all the tissues (Supplementary figure 2.21, Supplementary text 9), which is consistent with the effect observed during the *ADCY9*-KD experiment, showing increased expression of *CETP* expression when *ADCY9* is lowly expressed (Figures 2.10a and 2.11a,b). We evaluated the effects of the SNPs on expression levels of *ADCY9* and *CETP* by modelling both SNPs as continuous variables (additive model) (Methods). The *CETP* SNP rs158477 was reported as an expression quantitative trait locus (eQTL) in GTEx v7 and, in our models, shows evidence of being a cis eQTL of *CETP* in several other tissues (Supplementary text 9), although not reaching genome-wide significance. To test specifically for an epistatic effect between rs1967309 and rs158477 on *CETP* expression, we included an interaction term in eQTL models (Methods). We note here that we are testing for association for this specific pair of SNPs only, and that effects across tissues are not independent, such that we set our significance threshold at p-value=0.05. This analysis revealed a significant interaction effect (p-value=0.03, β= -0.22) between the two SNPs on *CETP* expression in GEUVADIS lymphoblastoid cell lines (Figure 2.10b, Supplementary figure 2.22a). In rs1967309 AA individuals, copies of the rs158477 A allele increased *CETP* expression by 0.46 (95% CI 0.26-0.86) on average. In rs1967309 AG individuals, copies of the rs158477 A allele increased *CETP* expression by 0.24 (95% CI 0.06-0.43) on average and the effect was null in rs1967309 GG individuals ($p - value_{GG}$=0.58). This suggests that the effect of rs158477 on *CETP* expression changes depending on genotypes of rs1967309. The interaction is also significant in several GTEx tissues, most of which are brain tissues, like hippocampus, hypothalamus and substantia nigra, but also in skin, although we note that the significance of the interaction

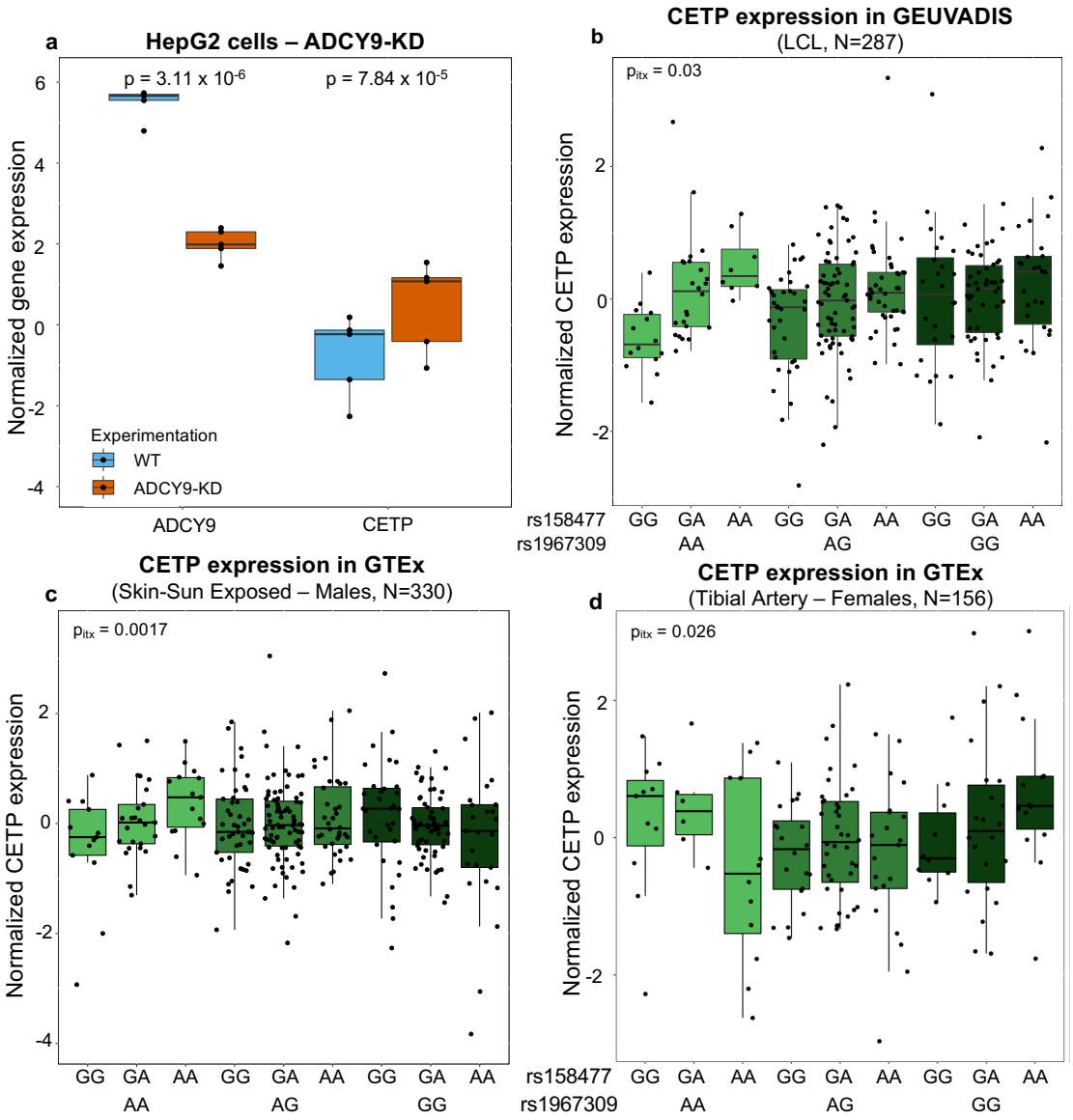

Fig. 2.10. Effect of *ADCY9* on *CETP* expression.
(a) Normalized expression of *ADCY9* or *CETP* genes depending on wild type (WT) and *ADCY9*-KD in HepG2 cells from RNA sequencing on five biological replicates in each group. P-values were obtained from a two-sided Wilcoxon paired test. qPCR and western blot results in HepG2 are presented in Figure 2.11. (b,c,d) *CETP* expression depending on the combination of rs1967309 and rs158477 genotypes in (b) GEUVADIS (p-value=0.03, ß= -0.22, N=287), (c) GTEx-Skin Sun Exposed in males (p-value=0.0017, ß= -0.32, N=330) and in (d) GTEx-Tibial artery in females (p-value=0.026, ß= 0.38, N=156), for individuals of European descent according to principal component analysis. P-values reported were obtained from a two-way interaction of a linear regression model for the maximum number of PEER/sPEER factors considered. Figure 2.12 show the interaction p-values depending on number of PEER/sPEER factors included in the linear models.

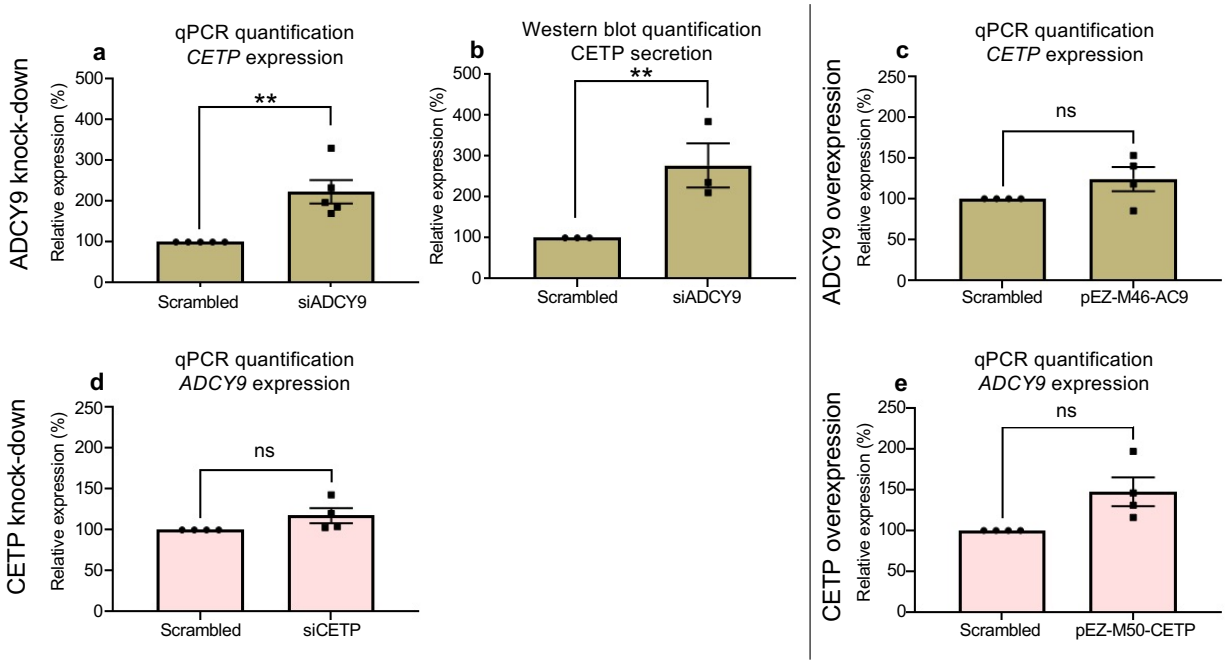

Fig. 2.11. *ADCY9*/*CETP* interaction in HepG2 cells.
(a) Relative mRNA expression of *CETP* of HepG2 cells 72h post-transfection with siRNA against human *ADCY9* (si1039). qPCR assay was normalized with PGK1 and HBS1L genes, n= 5 independent experiments, (p-value=0.0026 from t-test). (b) Quantification of CETP protein by Western blot assay, 200 ml of cell media (concentrated with Amicon ultra 0.5 ml 10 kDA units) from cells transfected with siRNA against human *ADCY9* (si1039), were separated on 10% TGX-acrylamide gel and transferred to PVDF membrane. CETP protein expression was determined using a primary antibody rabbit monoclonal anti-CETP (Abcam, ab157183) 1:1000 (3% BSA, TBS, Tween 20 0.5%) O/N 4oC, followed by HRP-conjugated secondary antibody goat anti-rabbit 1:10 000 (3% BSA) 1h RT. Figureb represents densitometry analysis of n=3 experiments, p-value=0.0029 from t-test. (c,e) Relative mRNA expression of (c) *CETP* and (e) *ADCY9* genes in HepG2 cells post-transfection with pEZ-M50-*CETP* (overexpression of *CETP*) or pEZ-M46-*ADCY9* (overexpression of *ADCY9*) plasmids. qPCR assay was normalized with PGK1 and HBS1L genes, n=4 independent experiments. (d) Relative mRNA expression of *ADCY9* of HepG2 cells 72h post-transfection with siRNA against human *CETP*. qPCR assay was normalized with PGK1 and HBS1L genes, n= 4 independent experiments.

depends on the number of PEER factors included in the model (Supplementary figure 2.23). These factors are needed to correct for unknown biases in the data, but also potentially lead to decreased power to detect interaction effects [286]. In CaG whole blood samples, the interaction effect using additive genetic effect at rs1967309 was not significant, similarly to results from GTEx in whole blood samples. However, given the larger size of the dataset, we evaluated a genotypic encoding for the rs1967309 SNP in which the interaction effect is significant (p-value=0.008, Supplementary figure 2.22b) in whole blood, suggesting that

rs1967309 could be modulating rs158477 eQTL effect, in this tissue at least, with a genotype-specific effect. We highlight that the sample sizes of current transcriptomic resources do not allow to detect interaction effects at genome-wide significance, however the likelihood of finding interaction effects between our two SNPs on *CETP* expression in three independent datasets is unlikely to happen by chance alone, providing evidence for a functional genetic interaction.

Given the sex-specific results reported above, we stratified our interaction eQTL analyses by sex. We observed that the interaction effect on *CETP* expression in CaG whole blood samples (N_{male}=359) is restricted to male individuals, and, despite low power due to smaller sample size in GEUVADIS, the interaction is also only suggestive in males (Supplementary figure 2.22c,d). In GTEx, most well-powered tissues that showed a significant effect in the sex-combined analyses also harbor male-specific interactions (Supplementary figure 2.24). For instance, GTEx skin male samples (N_{male}=330) show the most significant male-specific interaction effects, with the directions of effects replicating the sex-combined result in GEUVADIS (an increase of *CETP* expression for each rs158477 A allele in rs1967309 AA individuals) albeit with an observable reversal of the direction in rs1967309 GG individuals (decrease of *CETP* expression with additional rs158477 A alleles) (Figures 2.10c and 2.12a). However, significant effects in females are detected in tissues not previously seen as significant for the interaction in the sex-combined analysis, in the tibial artery (Figure 2.10d, Supplementary figure 2.12) and the heart atrial appendage (Supplementary figure 2.24). For tissues with evidence of sex-specific effects in stratified analyses, we also tested the effect of an interaction between sex, rs158477 and rs1967309 (Methods) on *CETP* expression: the three-way interaction is only significant for tibial artery (Figure 2.12).

4.5. Epistatic effects on phenotypes

The interaction effect of rs1967309 and rs158477 on *CETP* expression in several tissues, found in multiple independent RNA-seq datasets, coupled with the detection of LRLD between these SNPs in the Peruvian population suggest that selection may act jointly on these loci, specifically in Peruvians or Andeans. These populations are well known for their adaptation to life in high altitude, where the oxygen pressure is lower and where the human

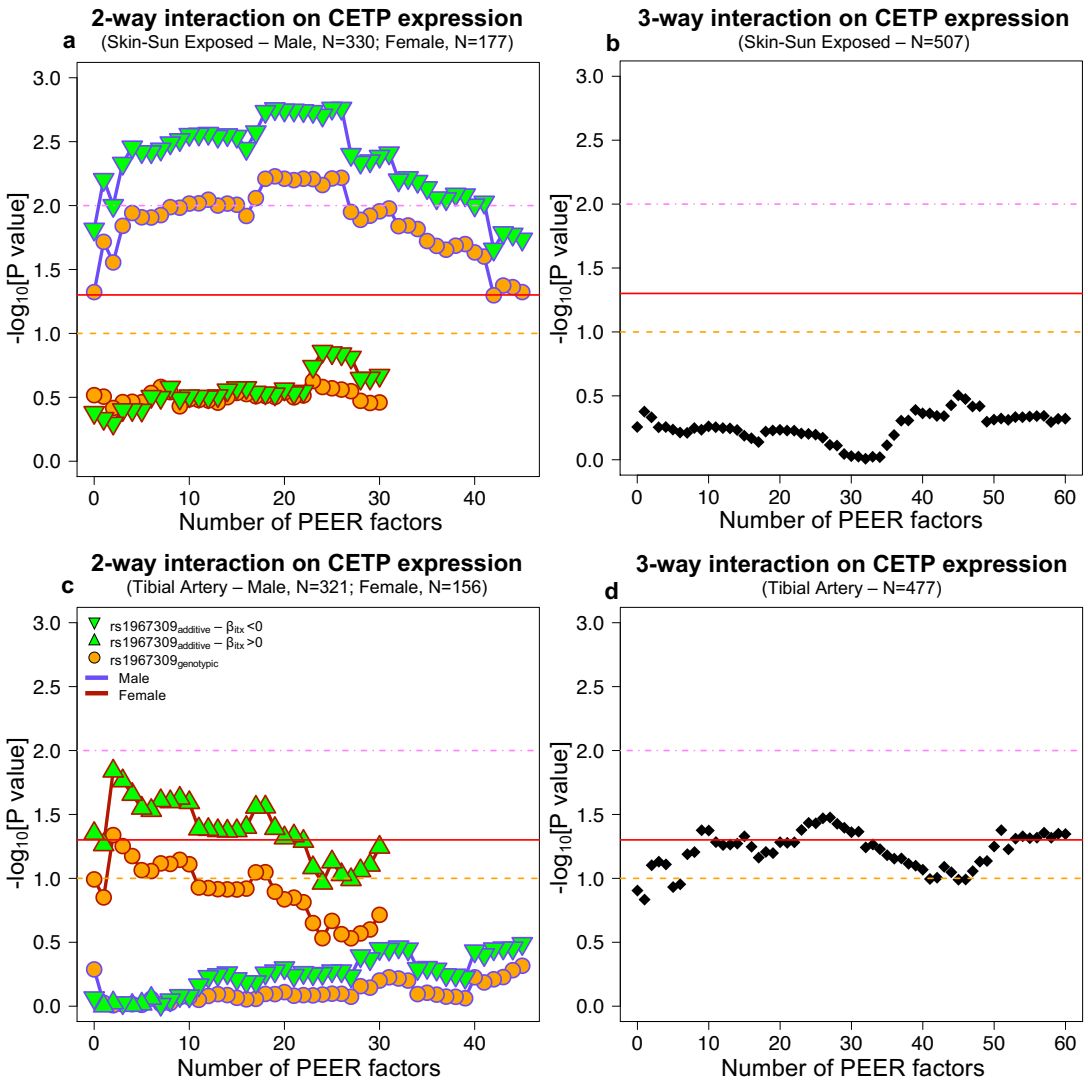

Fig. 2.12. Interaction effect p-values on *CETP* expression depending by the number of PEER factors in Skin-sun exposed (a,b) and Tibial artery (c,d) in GTEx.
For the two-way interaction (rs1967309*rs158477) (a,c), rs158477 is codded as additive (GG=0, GA=1, AA=2). In the additive model (green triangle), rs1967309 is codded as additive (AA=0, AG=1, GG=2). For the genotypic model (orange circle), rs1967309 was codded as a genotypic variable and p-values were obtained from a likelihood ratio test comparing models with and without the interaction term between the SNPs. The color of lines linking each value represents the sex. For the three-way interaction (rs1967309*rs158477*sex), both SNPs were codded as additive, and p-values were obtained from a linear regression model in R. P-values are presented on a -$\log_{10}$ scale. The orange, red and pink lines represent p-values of 0.1, 0.05 and 0.01 respectively.

body is subjected to hypoxia [287, 288, 289, 290]. High altitude hypoxia impacts individuals' health in many ways, such as increased ventilation, decreased arterial pressure, and alterations of the energy metabolism in cardiac and skeletal muscle [291, 292]. To test which

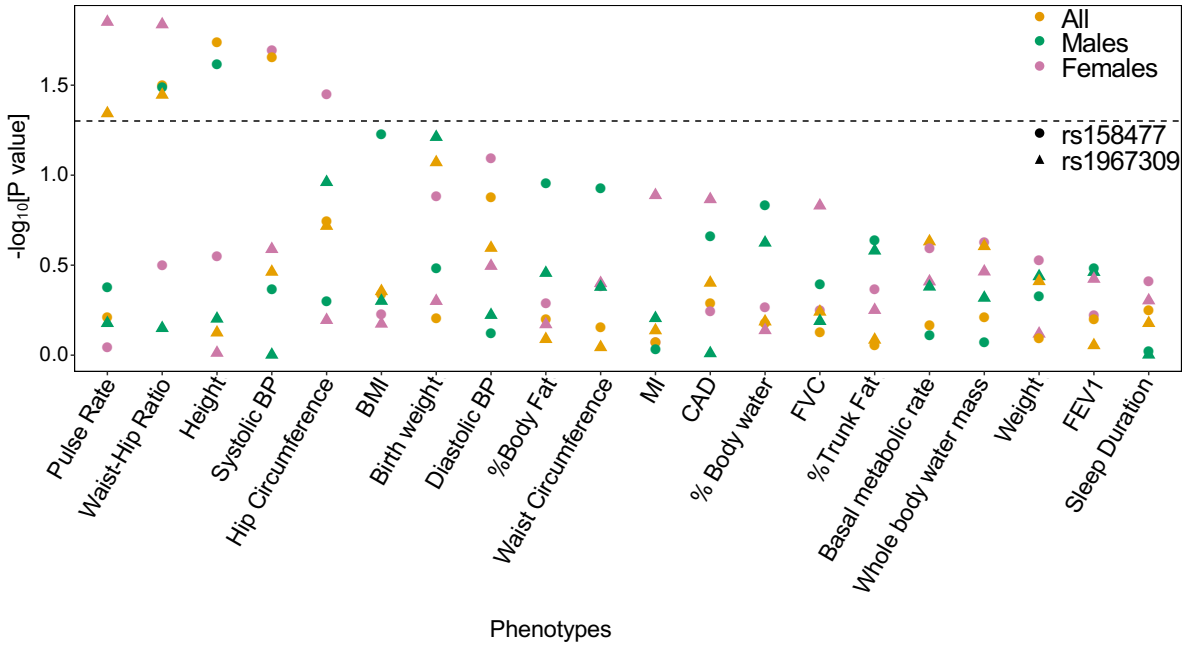

Fig. 2.13. Single SNP effects of rs1967309 and rs158477 on phenotypes in the UK biobank. Significance of the marginal effect of rs1967309 and rs158477, both codded as additive, on several physiological traits, energy metabolism and cardiovascular outcomes, overall and stratified by sex in the UK biobank. The dotted line represents the p-value at 0.05. See Supplementary table 2.4 for the list of abbreviations.

phenotype(s) may explain the putative coevolution signal discovered (Step 4, Figure 2.1), we investigated the impact of the interaction between rs1967309 and rs158477 on several physiological traits, energy metabolism and cardiovascular outcomes using the UK Biobank and GTEx cohort (Figure 2.13, Supplementary table 2.4). The UK Biobank has electronic medical records and GTEx has cause of death and variables from medical questionnaires [116]. The interaction term was found to be nominally significant (p-value$<$0.05) for forced vital capacity (FVC), forced expiratory volume in 1-second (FEV1) and whole-body water mass, and suggestive (p-value$<$0.10) for the basal metabolic rate, all driven by the effects in females (Figure 2.14a). For CAD, the interaction is suggestive (p-value$<$0.10) and, in this case, driven by males (Figure 2.14a).

Among the biomarkers studied (Supplementary table 2.4), only lipoprotein(a) [Lp(a)] is suggestive in males (p-value=0.08) for an interaction between rs1967309 and rs158477, with the same direction of effect as that for CAD (Figure 2.14). Again, given the differences observed between the sexes, we tested the effect of an interaction between sex, rs158477 and rs1967309 (genotypic coding, Methods) on biomarkers, and only Lp(a) was nominally

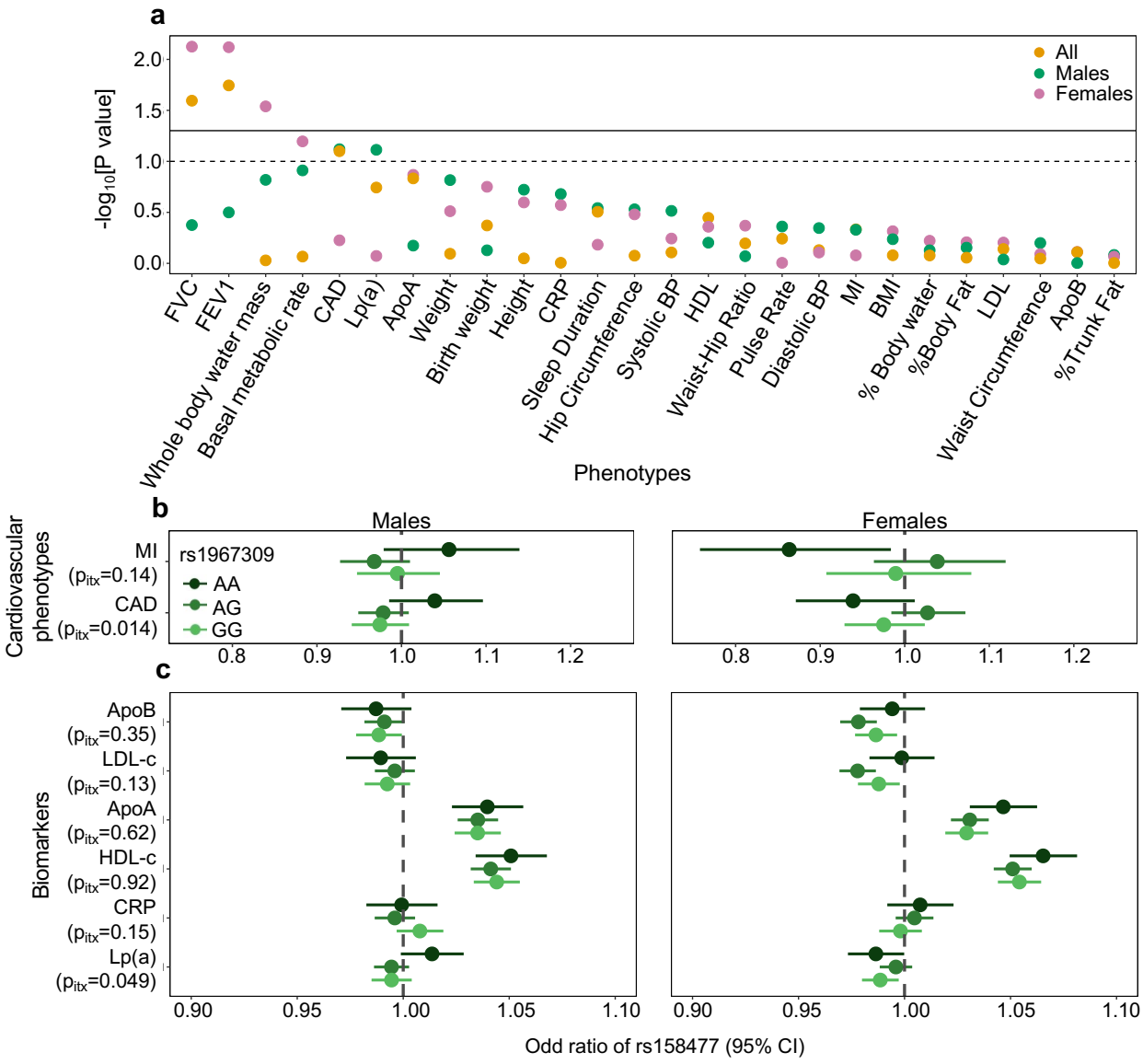

Fig. 2.14. Epistatic association of rs1967309 and rs158477 on phenotypes in the UK biobank.

(a) Significance of the interaction effect between rs1967309 and rs158477 on several physiological traits, energy metabolism and cardiovascular outcomes overall and stratified by sex in the UK biobank. Horizontal lines represent the p-value thresholds at 0.05 (plain) and 0.10 (dotted). Single-SNP p-values are shown in Figure 2.13. (b,c) Sex-stratified effects of rs158477 on (b) cardiovascular phenotypes and (c) biomarkers depending on the genotype of rs1967309 (genotypic encoding). The p-values pitx reported come from a likelihood ratio test comparing models with and without the three-way interaction term between the two SNPs and sex. Sex-combined results using GTEx cardiovascular phenotype data are shown in Figure 2.15. See Supplementary table 2.4 for the list of abbreviations.

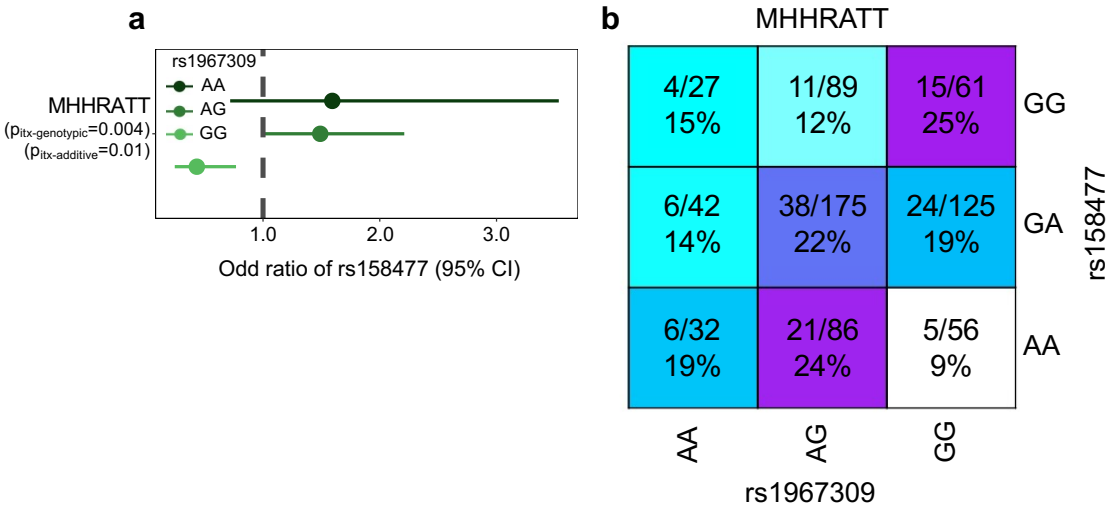

Fig. 2.15. Epistatic association of rs1967309 and rs158477 on cardiovascular disease in GTEx.
(a) Effect of the rs158477 SNP on the cardiovascular phenotype (n=693, cas=120, control=563) depending on the genotype of rs1967309 in GTEx. For both models, rs158477 was codded as additive (GG=0, GA=1, AA=2). For the additive model, rs1967309 was codded as additive (AA=0, AG=1, GG=2). P-value of the interaction (pitx) was obtained using a linear regression in R. For the genotypic model, rs1967309 was codded as a genotypic variable and p-values were obtained from a likelihood ratio test comparing models with and without the interaction term between the SNPs. (b) Proportion of cases for each genotype combinations between rs1967309 and rs158477. The numerator indicates the number of cases and the denominator the number total of individuals (cases+controls). Darker colors show higher proportions of cases.

significant in a three-way interaction (p-value=0.049). The pattern is similar to the results for CAD, ie. a change in the effect of rs158477 depending on the genotype of rs1967309 in males, with the effect for AA females in the opposite direction compared to males (Figure 2.14b). These concordant results between CAD and Lp(a) support that the putative interaction effect between the loci under study on phenotypes involves sex as a modifier.

5. Discussion

In this study, we used population genetics, transcriptomics and interaction analyses in biobanks to study the link between *ADCY9* and *CETP*. Our study revealed selective signatures in *ADCY9* and a significant genotypic association between *ADCY9* and *CETP* in two Peruvian cohorts, specifically between rs1967309 and rs158477, which was also seen in the Native population of the Andes. The interaction between the two SNPs was found to be nominally significant for respiratory and cardiovascular disease outcomes (Figures 2.14 and

2.15). Additionally, a nominally significant epistatic interaction was seen on *CETP* expression in many tissues, including the hippocampus and hypothalamus in the brain. Despite brain tissues not displaying the highest *CETP* expression levels, CETP that is synthesized and secreted in the brain could play an important role in the transport and the redistribution of lipids within the central nervous system [45, 293] and has been associated with Alzheimer's disease risk [50, 294]. These findings reinforce the fact that the SNPs are likely functionally interacting, but extrapolating on the specific phenotypes under selection from these results is not straight forward. Identifying the phenotype and environmental pressures that may have caused the selection signal is complicated by the fact that the UK Biobank participants, on which the marginally significant associations have been found, do not live in the same environment as Peruvians. In Andeans from Peru, selection in response to hypoxia at high altitude was proposed to have effects on the cardiovascular system [269]. The hippocampus functions are perturbed at high altitude (eg. deterioration of memory [295, 296]), whereas the hypothalamus regulates the autonomic nervous system (ANS) and controls the heart and respiratory rates [297, 298], phenotypes which are affected by hypoxia at high altitude [299, 300]. Furthermore, high altitude-induced hypoxia [301, 302] and cardiovascular system disturbances [303, 304] have been shown to be associated in several studies [305, 306, 307, 308, 309], thus potentially sharing common biological pathways. Therefore, our working hypothesis is that selective pressures on our genes of interest in Peru are linked to the physiological response to high-altitude, which might be the environmental driver of coevolution.

The significant interaction effects on *CETP* expression vary between sexes in amplitude and direction, with most signals driven by male samples, but significant interaction effects observed in females only, despite sample sizes being consistently lower than for males. Notably, in the tibial artery and heart atrial appendage, two tissues directly relevant to the cardiovascular system, the female-specific interaction effect on *CETP* expression is reversed between rs1967309 genotypes AA and GG, compared to the effects seen in males in skin and brain tissues. Given our *ADCY9*-KD were done in liver cell lines from male donors, future work to fully understand how rs1967309 and rs158477 interact will focus on additional experiments in cells from both male and female donors in these relevant tissues. In a previous study, we showed that inhibition of both Adcy9 and CETP impacted many phenotypes

linked to the ANS in male mice [89], but in the light of our results, these experiments should be repeated in female mice. The function of ANS is important in a number of pathophysiological states involving the cardiovascular system, like myocardial ischemia and cardiac arrhythmias, with significant sex differences reported [310, 311, 312].

The interaction effect between the *ADCY9* and *CETP* SNPs on both respiratory and cardiovascular phenotypes differs between the sexes, with effects on respiratory phenotypes limited to females (Figure 2.14a) and cardiovascular disease phenotype associations showing significant three-way sex-by-SNPs effects (Figure 2.14). Furthermore, the LRLD signal is present mainly in males (Figure 2.6), although the genotype association is also seen in female for a different genotype combination, suggesting the presence of sex-specific selection. This type of selection is very difficult to detect, especially on autosomes, with very few empirical examples found to date in the human genome despite strong theoretical support of their occurrence [313]. However, sexual dimorphism in gene expression between males and females on autosomal genes has been linked to evolutionary pressures [314, 315, 316], possibly with a contribution of epistasis. As the source of selection, we favor the hypothesis of differential survival over differential ability to reproduce, because the genetic combination between *ADCY9* and *CETP* has high chances to be broken up by recombination at each generation. Even in the case where recombination is suppressed in males between these loci, they would still have equal chances to pass the favored combination to both male and female offspring, which would not explain the sex-specific LRLD signal. We see an enrichment for the rs1967309-AA + rs158477-GG in males and rs1967309-AA + rs158477-AA in females, which are the beneficial combination for CAD in the corresponding sex, possibly pointing to a sexually antagonistic selection pressure, where the fittest genotype combination depends on the sex.

Such two-gene selection signature, where only males show strong LRLD, can happen if a specific genotype combination is beneficial in creating males (through differential gamete fitness or in utero survival, for example) or if survival during adulthood is favored with a specific genotype combination compared to other genotypes. In the case of age-dependent differential survival, the genotypic association is expected to be weaker at younger ages, however the LRLD signal between rs1967309 and rs158477 in the LIMAA cohort did not depend on age neither in males nor in females (Supplementary text 9). Since very few

individuals were younger than 20 years old, it is likely that the age range in this cohort is not appropriate to distinguish between the two possibilities. This age-dependent survival therefore remains to be tested in comparison with pediatric cohorts of Peruvians: if the LRLD signal is absent in newborns for example, it will suggest a strong selective pressure acts early in life on boys. To specifically test the in-utero hypothesis, a cohort of stillborn babies with genetic information could allow to evaluate if the genotype combination is more frequent in these. Lastly, it may be that the evolutionary pressure is linked to the sex chromosomes [317, 318], and a three-way interaction between *ADCY9*, *CETP* and Y chromosome haplotypes or mitochondrial haplogroups remains to be explored.

Even though we observed the LRLD signal between rs1967309 and rs158477 in two independent Peruvian cohorts, reducing the likelihood that our result is a false positive, one limitation is that the individuals were recruited in the same city (Lima) in both cohorts. However, we show that both populations are heterogeneous with respect to ancestry (Supplementary figure 2.17), suggesting that they likely represent accurately the Peruvian population. As recent admixture and population structure can strongly influence LRLD, we performed several analyses to consider these confounders, in the full cohorts and in the sex-stratified analyses. All analyses were robust to genome-wide and local ancestry patterns, such that our results are unlikely to be explained by these effects alone (Supplementary text 9). Unfortunately, we could not use expression and phenotypic data from Peruvian individuals, which makes all the links between the selection pressures and the phenotype associations somewhat indirect. Future studies should focus on evaluating the phenotypic impact of the interaction specifically in Peruvians individuals, in cohorts such as the Population Architecture using Genomics and Epidemiology (PAGE) [319], in order to confirm the marginally significant associations found in European cohorts. Indeed, the Peruvian/Andean genomic background could be of importance for the interaction effect observed in this population, which reduces the power of discovery in individuals of unmatched ancestry. Furthermore, not much is known about the strength of this type of selection, and simulations would help evaluate how strong selection would need to be in a single generation to produce this level of LRLD. Another limitation is the low number of samples per tissue in GTEx and the cell composition heterogeneity per tissue and per sample [320], which can be partially captured

by PEER factors and can modulate the eQTL effects. Therefore, our power to detect tissue-specific interaction effects is reduced in this dataset, making it quite remarkable that we were able to observe multiple nominally significant interaction effects between the loci.

Despite these limitations, our results support a functional role for the *ADCY9* intronic SNP rs1967309, likely involved in a molecular mechanism related to *CETP* expression, but this mechanism seems to implicate sex as a modulator in a tissue-specific way, which complicates greatly its understanding. In the dal-OUTCOMES clinical trial, the partial inhibitor of CETP, dalcetrapib, did not decrease the risk of cardiovascular outcomes in the overall population, but rs1967309 in the *ADCY9* gene was associated to the response to the drug, which benefitted AA individuals [24]. Interestingly, rs1967309 AA is found in both the male and female beneficial combinations of genotypes for CAD, the same that are enriched in Peruvians, but without taking rs158477 and sex into account, this association was masked. The modulation of *CETP* expression by rs1967309 could impact CETP's functions that are essential for successfully reducing cardiovascular events. The rs158477 locus could be a key player for these functions, and dalcetrapib may be mimicking its impact, hence explaining the pharmacogenomics association. Furthermore, in the light of our results, some of these effects could differ between men and women [321], which may need to be taken into consideration in the future precision medicine interventions potentially implemented for dalcetrapib.

In conclusion, we discovered a putative epistatic interaction between the pharmacogene *ADCY9* and the drug target gene *CETP*, that appears to be under selection in the Peruvian population. Our approach exemplifies the potential of using evolutionary analyses to help find relationships between pharmacogenes and their drug targets. We characterized the impact of the *ADCY9*/*CETP* interaction on a range of phenotypes and tissues. Our gene expression results in brain tissues suggest that the interaction could play a role in protection against challenges to the nervous system caused by stress such as hypoxia. The female-specific eQTL interaction results in arteries and heart tissues further suggest a link with the cardiovascular system, and the phenotype association results support further this hypothesis. In light of the associations between high altitude-induced hypoxia and cardiovascular system changes, the interaction identified in this study could be involved in both systems: for example, ADCY9 and CETP could act in pathways involved in adaptation to high altitude, which could influence cardiovascular risk via their interaction in a sex-specific manner. Finally, our

findings of an evolutionary relationship between *ADCY9* and *CETP* during recent human evolution points towards a biological link between dalcetrapib's pharmacogene *ADCY9* and its therapeutic target CETP.

6. Acknowledgments

We thank all members of the Hussin lab for their constructive comments and feedback throughout this project, as well as the insightful input from three reviewers and reviewing and senior editors at eLife. This work was completed thanks to computational resources provided by Compute Canada clusters Graham and Beluga. This work was funded by the Institut de Valorisation des Données (IVADO), Health Collaboration Acceleration Fund from the Ministère de l'Économie et de l'Innovation of the Government of Quebec and the Montreal Heart Institute (MHI) Foundation. JGH is a Fonds de la Recherche en Santé (FRQS) Junior one fellow. IG receives a PhD scholarship from the MHI Foundation and MAL holds a PhD scholarship from Canadian Institutes of Health Research. MPD holds the Canada Research Chair in Precision Medicine Data Analysis. JCT holds the Canada Research Chair in Personalized Medicine and the Université de Montréal endowed research chair in atherosclerosis.

7. Funding Sources

Institut de Cardiologie de Montréal

- Isabel Gamache
- Marc-André Legault
- Jean-Christophe Grenier
- Rocio Sanchez
- Eric Rhéaume
- Holly Trochet
- Jean-Claude Tardif
- Marie-Pierre Dubé
- Julie Hussin

Université de Montréal

- Isabel Gamache

Canadian Institutes of Health Research

- Marc-André Legault

Canada Research Chairs

- Jean-Claude Tardif
- Marie-Pierre Dubé

Fonds de Recherche du Québec - Santé

- Julie Hussin

Institut de Valorisation des Données IVADO

- Isabel Gamache
- Jean-Christophe Grenier
- Julie Hussin

Ministère de l'Économie, de la Science et de l'Innovation - Québec (Health collaboration acceleration fund)

- Jean-Claude Tardif
- Marie-Pierre Dubé

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

8. Disclosures

JCT reports grants from Government of Quebec, National Heart, Lung, and Blood Institute of the U.S. National Institutes of Health (NIH), the MHI Foundation, from Bill and Melinda Gates Foundation, Amarin, Esperion, Ionis, Servier, RegenXBio; personal fees from Astra Zeneca, Sanofi, Servier; and personal fees and minor equity interest from Dalcor. Has a patent (US20190070178A1) Methods for Treating or Preventing Cardiovascular Disorders and Lowering Risk of Cardiovascular Events issued to Dalcor, no royalties received, a patent (US20170233812A1) Genetic Markers for Predicting Responsiveness to Therapy with HDL-Raising or HDL Mimicking Agent issued to Dalcor, no royalties received, and a patent (US Provisional Applications No. 62/935,751 and 62/935,865) Methods for using low dose colchicine after myocardial infarction with royalties paid to Invention assigned to the Montreal Heart Institute.

MPD has a patent (US20190070178A1) Methods for Treating or Preventing Cardiovascular Disorders and Lowering Risk of Cardiovascular Events issued to Dalcor, no royalties received, a patent (US20170233812A1) Genetic Markers for Predicting Responsiveness to Therapy with HDL-Raising or HDL Mimicking Agent issued to Dalcor, no royalties received, and a patent (US Provisional Applications No. 62/935,751 and 62/935,865) Methods for using low dose colchicine after myocardial infarction with royalties paid to Invention assigned to the Montreal Heart Institute. M.P.D. reports personal fees and other from Dalcor and personal fees from GlaxoSmithKline, other from AstraZeneca, Pfizer, Servier, Sanofi.

JH has received speaker honoraria from Dalcor and District 3 Innovation Centre

9. Supplementary text

9.1. Data pre-processing

9.1.1. Pre-processing of Native American

The genetic data was obtained following correspondence with Reich et al. 2012 co-authors. The Native American Genetic Dataset (NAGD) dataset being quite sparse and samples coming from many different populations, no missing data threshold nor minor allele frequency or Hardy-Weinberg equilibrium filters were applied prior to the imputation. Harmonization to the hg19 reference genome has been done using GenotypeHarmonizer v.1.4.20 [322] and bcftools v.1.9 [279] with the fixref plugin (-m flip option). Imputation was done using the Sanger Imputation Server [323] using Haplotype Reference Consortium (r1.1) reference panel, with a pre-phasing using SHAPEIT2 r.837 [324] and imputation using PBWT [325]. Post-imputation quality control was done by keeping sites with an INFO score over 0.8 and keeping genotypes having a posterior probability over 0.9. SHAPEIT2 was run to get phased genotypes (parameters: effective size of 10,000, burn of 10, prune of 10, main of 25, states of 400). The obtained VCF was used in the RFMix analysis (see below). SNPs with missing genotypes higher than 90% after imputation were removed for LRLD analysis.

9.1.2. Pre-processing of the LIMAA cohort

A pre-imputation step was conducted keeping only positions passing minor allele frequency (MAF) of 1%, 1% of missing data per site and HWE p-value $>$ 1e-5 using PLINK

v.1.9 [326]. Harmonization to the hg19 reference genome has then been done using GenotypeHarmonizer and bcftools with the fixref plugin (-m flip option). Imputation was done using the Sanger Imputation Server, using Haplotype Reference Consortium (r1.1) reference panel, with a pre-phasing using SHAPEIT2 and imputation using PBWT. Post-imputation quality control was done by keeping sites with an INFO score over 0.8 and keeping genotypes having a posterior probability over 0.9. Furthermore, positions having less than 5% missing rate after the genotyping recoding step were kept and duplicated positions were removed. SHAPEIT2 was run to get phased genotypes (parameters: effective size of 10,000, burn of 10, prune of 10, main of 25, states of 400). Another dataset was built to recover one of our SNPs of interest (rs1967309), which was excluded from our previous pipeline because of their INFO score (0.79). In this new dataset, the INFO score threshold was put to 0.7 and the post-imputation position missing data threshold was set to 35%, being less stringent, but recovering our positions. To make sure imputation quality did not impact our results because of incorrectly imputed genotypes, we redid the imputation of LIMAA with the TOPMED reference panel (https://imputation.biodatacatalyst.nhlbi.nih.gov/#!). The imputation r^2 score with TOPMED is higher than 0.9 for both, and only very limited differences in imputed genotypes are seen (only 5% and 2% of individual allele mismatches in LIMAA for rs1967309 and rs158477, respectively for the 3,243 individuals).

9.1.3. Pre-processing of GTEx genetic data

Starting from the imputed genotyping dataset, we kept bi-allelic SNPs and removed positions with more than 5% missing genotype, remaining 100,986 SNPs to calculate PCA using flashPCA2. To remove the Hispanic group, we reduced the dimensionality of the top 10 Principal Components (PCs) using the R package `UMAP` [327] (default parameters) to obtain a two dimensional representation of the genetic information contained within those PCs. We identified the largest homogeneous group (self-reported 'white') and excluded outlier groups (Supplementary figure 2.25a), used only these individuals for the rest of the analyses. We did our all subsequent analyses with 699 individuals.

9.1.4. Pre-processing of CARTaGENE

CARTaGENE biobank [254] includes 40K individuals from Quebec (Canada) having between 36 and 72 years old. 12,056 individuals were genotyped and among these 911 had

RNAseq performed on whole blood [328, 329]. The genotypes are coming from five different genotyping arrays on which imputation was processed independently. A pre-imputation step was conducted keeping only genotypes passing maf of 1%, 1% of missing data per site and HWE p-value > 1e-5 using PLINK. Harmonization to the Hg19 reference genome has then been done using GenotypeHarmonizer and bcftools with the fixref plugin (-m flip option). Imputation was done using the Sanger Imputation Server, using Haplotype Reference Consortium (r1.1) reference panel, with a pre-phasing using SHAPEIT2 and imputation using PBWT. Post-imputation quality control was done by keeping sites with an INFO score over 0.8 and keeping individual genotypes having a posterior probability over 0.9.

To extract only white European, we used the same filter as for GTEx, except that we removed SNPs having any missing genotypes which could create bias by different chips, then followed the recommendation from flashPCA2, remaining 8,869 SNPs to calculate PCA. We reduced the dimensionality of the top 10 PCs using the R package `UMAP` (default parameters) to obtain a two dimensional representation of the genetic information contained within those PCs. We identified the largest homogeneous group (Supplementary figure 2.25b), which contains a majority of individuals from European descent (self-reported 'white'), and used only those individuals for the rest of the analysis. We kept 11,362 individuals at the end and among these, 911 individuals for which we had RNAseq. For these individuals, we merged samples from different batches, we removed samples who had less than 10 millions of reads, remaining 790 individuals with expression. After filtering out individuals missing the genotype of either rs1967309 or rs158477 SNPs, we did our interaction analysis on 728 individuals.

9.2. Population genetics

9.2.1. iHS analyses

We computed the integrated haplotype score (iHS) [153] for each subpopulation in the 1000 Genomes project (Methods), a statistics that allows us to detect evidence for recent strong positive selection on derived alleles. The SNP rs1967309 is located in a region of high linkage disequilibrium (LD), delimited by recombination hotspots present in all populations. Several SNPs in this LD block exhibit absolute iHS values higher than 2 in non-African populations (Figure 2.2b, Supplementary figure 2.16), specifically in CEU and GBR (highest

signal is a 15 Kb away from rs1967309), CHB, CHS, CDX, KHV, and in all SAS sub-populations, all of which showing signals in several SNPs in less than 200 base pairs from rs1967309. Of note, however, rs1967309 itself does not show value over 2 in any population. In African populations, no signal is seen in this LD block (Supplementary figure 2.16). Other SNPs in *ADCY9* are found to have absolute iHS values higher than 2, especially in the long intron 1 and around the last exon, but characterizing these signals is beyond the scope of this study.

9.2.2. Sex-specific differentiation at rs1967309 in *ADCY9*

We first used F_{ST} to evaluate differences in genotype frequencies between males and females. In the PEL from 1000G, we saw suggestive differences between males and females around rs1967309, but did not replicate in the LIMAA cohort, which suggests it was due to small sample size [330]. Another approach we took was to investigate the impact of sex on our PBS results, by splitting the sample between males and females, and recomputing all PBS values using PEL, MXL and CHB for SNPs on chromosome 16 in each subsample. We report result on chromosome 16 that account for chromosome specific population history, as in our analyses of the full cohort, tests on chromosome 16 were more conservative than on the whole genome (ie. p-values were slightly larger with chromosome 16 alone). Although over the full chromosome, the distribution was not statistically different between males and females ($PBS_{95th-PEL,male} = 0.043$; $PBS_{95th-PEL,female} = 0.040$) as expected, curiously the PEL branch length for all SNPs around rs1967309 increases for males compared to the full-sample results : at rs1967309, the PBS value became 0.096 in males (chromosome 16 empirical p-value = 0.004). On the other hand, for females the value dropped to 0.017 (chromosome 16 empirical p-value = 0.20). No such male-female difference is seen in *CETP*, with the PEL PBS value for rs158477 remaining significantly elevated in both sexes (chromosome 16 empirical $p - value_{rs158477,male} = 0.04$, chromosome 16 empirical $p - value_{rs158477,female} = 0.01$, Supplementary figure 2.18b,c). This suggests that the LD block around rs1967309 is differentiated between males and females in the Peruvians from 1000G. However, we note that the null model for the FST statistic underlying PBS assumes no difference in genotype frequencies between sex (ie. may not be the appropriate tool to address this specific question), and we cannot exclude the possibility of random sampling noise.

9.2.3. Admixture analyses

Recent admixture and migration events can influence LRLD. If segments of the genome are particularly enriched for a specific ancestry, this could lead to inflated LRLD between these segments. Given that the Peruvian is an admixed population between individuals of Native American ancestry (mainly Andean) as well as of European ancestry (Supplementary figure 2.17), we ran several analyses to establish whether our results at *ADCY9*/*CETP* can be explained by admixture patterns.

9.2.4. Local ancestry inference pre-processing

The reference populations used to run RFMix were YRI for the African ancestry, CEU for the European, CHB for the Asian from 1000G, subpopulations in NAGD (Northern American, Central American and Andean) for the Native American ancestry. We estimated local ancestry with RFMix on PEL from 1000G and LIMAA individuals.

For all 1000G populations (YRI, CEU, CHB, PEL), NAGD (Northern American, Central American and Andean) and LIMAA cohort, from the pre-processed datasets (see above) we kept only biallelic SNPs positions, removed SNPs with a MAF under 1% for each subpopulation, with more than 1% of missing individuals, with Hardy-Weinberg equilibrium p-value $< 10^{-4}$ with mid-adjustment using PLINK. We kept overlapping positions between all datasets and extracted the minor allele frequencies for each reference group. To avoid overlapping positions on the genetic maps, when SNPs had the exact same genetic position, we selected the SNP with the higher variance in allele frequencies (using var in R) between the reference groups (all subpopulations except PEL and LIMAA), keeping between 6,742 and 57,238 SNPs per chromosome for RFMix analysis.

9.2.5. Assessing proportions of global Andean ancestry

To see if there could be a potential enrichment or depletion of Andean ancestry at *CETP* and *ADCY9* loci compare to the rest of the genome, we looked at the proportion of attribution of Andean at those loci compared to the overall distribution of all chromosomes. From the 584,797 positions used for RFMix on all chromosomes, 4,476 position intervals were given, and we calculated the proportion of Andean attribution for each interval, then calculated the 95% confidence interval (CI) for all chromosomes which is [0.43 - 0.75]. The proportion at *ADCY9* and *CETP* loci were 0.58 and 0.66 respectively, which suggests that

the correlation between *ADCY9* and *CETP* loci is unlikely to be due to an enrichment or depletion of Andean ancestry at both loci. Results are similar when only considering chromosome 16 to calculate the 95% CI.

9.2.6. LRLD in the Andean population from NAGD

Another question is to assess if the association was already present in the non-admixed ancestral Andean population. If this is the case, the association cannot be explained by the random distribution of Andean segments across the Peruvian genome. We computed LRLD as described in Peruvians in the Andean population from NAGD and we found that the association between rs1967309 and rs158477 is also significant (*ADCY9*/*CETP* empirical p-value=0.04, Figure 2.4a,b, Supplementary table 2.3). We note that, in this population, strong association signals with rs158477 are also seen at other SNPs in the *ADCY9* LD block region. This result provides convincing evidence that the results in PEL and LIMAA are not due to random distribution of admixed segments but rather might have been inherited from the Andean population, where it was already present, and is maintained since then by selection.

In the Andean population, the association between rs1967309 and rs158477 is not significant when we stratified by sex (Supplementary table 2.3), but we still see significant association signals with rs158477 at other SNPs in *ADCY9* LD block in both sexes (Figure 2.9)

9.2.7. Comparison between Peruvian cohorts

To evaluate the genetic difference between Peruvian from 1000G and LIMAA, we performed a PCA starting from the phased data files. We kept only biallelic SNPs with a MAF higher than 5% in each cohort and kept only positions with no missing genotype. We followed the suggestion given by flashPCA2 (https://github.com/gabraham/flashpca) [331], remaining 18,345 SNPs for the PCA. We then did a UMAP on 50 PCs given by flashPCA2 using the `UMAP` package on R (default parameters) (Supplementary figure 2.20). As seen in the UMAP analysis, population structure exists in LIMAA, and PEL samples are mainly part of the largest subgroup observed in Supplementary figure 2.17e, which was kept for LIMAA analyses to remove any confounders linked to population subdivision (see below).

Also, the LIMAA cohort was initially recruited as part of a tuberculosis study [271], but our PCA and UMAP analysis showed no separation according to disease state.

9.2.8. Null distributions of LRLD

To evaluate how likely it is to observe, specifically in the admixed Peruvian population, a genotype correlation of $r^2 = 0.08$ between SNPs that are approximately 53 Mb apart on the same chromosome like between rs1967309 and rs158477, we have used two approaches. The first one was specific to the two genes under study, *ADCY9* and *CETP*, and therefore controls for all genomic factors specific to these regions. We selected all SNPs with MAF>0.05 in the two genes, and computed r^2 values for all 37,802 pairs (461 SNPs in *ADCY9* and 82 SNPs in *CETP*), yielding a null distribution for the expected genetic correlation between these genes. We then compared our r^2 value for rs1967309 and rs158477 to this distribution, with its rank being reported as an empirical p-value. This is referred to in the Results section as "*ADCY9*/*CETP* empirical p-value".

This approach is appropriate to correct for the genomic context specific to our genes of interest, but does not account neither for allele frequencies (most SNPs in the null will be at lower frequencies than our two SNPs) nor for overall admixture levels in the genome of this sample, thus we used a second empirical approach to account for these important confounders. For this genome-wide null distribution of the LRLD matching our SNPs, we generated one set of pairs of SNPs and evaluated LRLD between these random pairs in both LIMAA cohort and PEL from 1000G. Since frequencies in the LIMAA cohort are likely better estimates of allele frequencies in the Peruvian population because of the size of the sample, we started our selection based on SNPs' characteristic in this cohort: we extracted pairs of biallelic SNPs from chromosome 1 to 18, (the other chromosomes being too small) with a MAF between 15% and 30%, separated by between 50-60 Mb and 61±10 cM based on the PEL genetic map from 1000G. If SNPs in a pair shared coordinates on the genetic map (in cM) with another SNP from another pair, we kept only one of these pairs. We ended up with 3,576 non-overlapping SNP pairs for calculating the LRLD null distribution matching our rs1967309-rs158477 pair obtained from the LIMAA cohort. For analysis in PEL from 1000G, we added an extra frequency filtering step to remove pairs for which one or both SNPs had a MAF below 5% in PEL, leaving 3,513 pairs for analysis for PEL of 1000G. To calculate an empirical p-value in PEL, we evaluated the number of pairs which had a LRLD

value larger to the observed value for rs1967309-rs158477 and divided this number by the total number pairs (n=3,513). This is referred to in the Results section as "genome-wide empirical p-value".

From the 3,513 pairs of SNPs sampled to create the genome-wide null distribution in both sexes in PEL, we stratified by sex and recomputed null distributions for males and females in the same way as for the full cohorts, also with a MAF filter at 5%, leaving 3,505 pairs in males and 3,512 in females in PEL. In males, the r^2 value between rs1967309 and rs158477 was the highest of the distribution (genome-wide empirical p-value < 2.85 x 10^{-4}), but for females, it was in the 20^{th} percentile (genome-wide empirical p-value = 0.80).

9.2.9. Permutation analysis of sex-specific LRLD at the positions rs1967309 and rs158477

A second null distribution was derived for evaluating if the LRLD difference between sex for the rs1967309-rs158477 pair was significant, given the significant LRLD observed at these loci. We permuted the sex labels within the cohort and split them into two random groups of 42 pseudo-males and pseudo-females, while making sure an equal number of real males and females (21 of each) are found in each random group, yielding a total of 919 unique random splits that respected these conditions for the 85 PEL individuals. For each iteration, we calculated LRLD between rs1967309-rs158477 for each group and computed the absolute difference between them. To calculate a p-value, we evaluated the number of iterations that had a LRLD difference of more than or equal to the observed difference for the rs1967309-rs158477 pair between true males and females. The true absolute difference in r^2 values between rs1967309 and rs158477 (0.346) is the third highest value in this null distribution (p-value=0.002) (Supplementary figure 2.19c).

9.2.10. Genotype association between rs1967309 and rs158477 in LIMAA

In the LIMAA cohort, we performed a genotype association test using a χ^2 test with four degrees of freedom (χ^2_4) with a permutation scheme to obtain the p-values, as reported in [159], to control for the marginal one-locus genotype counts. To avoid the potential effects of population subdivision on LRLD [155, 285], we only kept individuals in the largest, likely more homogeneous, group seen in the UMAP performed on the first 50 PCs with PEL from 1000G (Supplementary figure 2.17e). Two smaller distinct groups were identified in

the UMAP analysis and these individuals were excluded from our analysis (cross shaped individuals in Supplementary figure 2.17e), leaving 3,243 individuals for analysis. The permutation scheme consists in permuting the rs1967309 values 1,000 times and computing the number χ^2_4 values obtained by permutation that are higher than the observed value for the rs1967309/rs158477 pair. For LIMAA, the χ^2_4 value is 82.0 (permutation p-value < 0.001). We then performed the same analysis by stratifying by sex, and obtained a χ^2_4 value of 56.6 (permutation p-value = 0.001) in males and a χ^2_4 value of 37.0 (permutation p-value = 0.017) in females. We note that performing this analysis in the full cohort of 3,509 individuals (without excluding individuals from subpopulations shown in Supplementary figure 2.17e) yield very similar results (full cohort χ^2_4=77.6, male χ^2_4=56.5, female χ^2_4=34.5). To assess which combination is driving the effect, we used an empirical combination-specific test: the p-value is obtained by breaking the real rs1967309-rs158477 genotype combinations by permuting rs1967309 genotypes and evaluating how many permutated samples show an enrichment of a specific genotype combination as large as in the real data. Interestingly, the combination driving the highly significant male effect is an excess of rs1967309-AA + rs158477-GG (combination-specific permutation p-value < 0.001), whereas in female, the result seems to be driven by rs1967309-AA + rs158477-AA (combination-specific permutation p-value = 0.014). These sex-specific genotypic effects could not be captured by a linear model and can explain why the r^2 value in LIMAA is smaller than in PEL. Additionally, we note that in both sexes (but mainly in males), the low-frequency rs1967309-GG + rs158477-GA combination is enriched in LIMAA (observed counts is 112 whereas expected according to allele frequencies at both loci is 59.7).

Finally, to evaluate the effect genome-wide, we calculated the χ^2_4 for all 3,576 pairs from the above described genome-wide null distribution for LRLD, then compared these with the value obtained for the rs1967309/rs158477 pair, in all individuals, males and females of LIMAA. In all groups, the rs1967309/rs158477 χ^2_4 values were in the top values (genome-wide empirical p-value$_{all}$ = 0.0003; genome-wide empirical p-value$_{males}$ = 0.002; genome-wide empirical p-value$_{females}$ = 0.001), meaning that the association is significant genome-wide, as found in PEL using r^2 (Figure 2.3d).

Despite lower power in 1000G PEL sample, we replicated the rs1967309-AA + rs158477-GG enrichment in males in PEL using a 2x2 χ^2 test, comparing specifically the rs1967309-AA + rs158477-GG to the three other combinations (rs1967309-nonAA+rs158477-GG; rs1967309-AA+rs158477-nonGG; rs1967309-nonAA+rs158477-nonGG, permutation p-value = 0.018). The rs1967309-AA + rs158477-AA association seen in females does not replicate (permutation p-value = 0.51) possibly due to low sample size (observed counts is 1, expected counts is 0.66).

Age was available in the LIMAA cohort, enabling us to test whether the LRLD pattern is associated with age, which could suggest a survival benefit if the association is not seen at younger ages. No correlation was seen between genotype and age for rs1967309 and rs158477, and age distributions between males rs1967309-AA+rs158477-GG and females rs1967309-AA+rs158477-GG were not significantly different. Because sample size was large enough in this cohort to perform a stratified analysis, we further split the cohort into nearly balanced age categories in males (0-19 years old: 435; 20-25: 464, 26-35: 523; over 35: 519) to establish if the LRLD is present in a specific sub-group. To test if the enrichment of rs1967309-AA + rs158477-GG in males varies between age group, we calculated the expected frequencies using the frequencies in all age combined in males only (using the whole sample allele frequencies did not change the results). First, we generated a 2x2 contingency table comparing rs1967309-AA + rs158477-GG versus the three others (see above), then calculated a χ^2, then we used a permutation test permuting rs1967309 genotypes 1,000 times to assess statistical significance. The empirical p-values suggest differences between age groups (permutation $p-value_{0-19} = 0.12$; permutation $p-value_{20-25} = 0.21$; permutation $p-value_{26-35} = 0.01$; permutation $p-value_{>35} = 0.04$), with significant p-values in older groups only. We thus more formally tested if the association differs between age groups between 0 and 25 years old (n=899) and above 25 (n=1042): we performed a χ^2 test based on a 2x2 contingency table using the chisq.test() function in R, comparing the rs1967309-AA + rs158477-GG versus all other genotypes for the 2 age groups, permutating age values 1,000 times to estimate an empirical p-value. There was no significant difference between age group ($\chi^2 = 0.02$, permutation p-value=0.88). Similar results were obtained when the number of individuals were balanced across the two groups (cut off of 26 years old) or when the four initial age groups were used.

Finally, we also considered how the imputation quality in LIMAA could affect the main result of LRLD, because of imputation from non-representative reference panel populations is known to be problematic. We recomputed the genotype correlation (r^2) in LIMAA with our two SNPs imputed with the TOPMED panel, a more representative panel than the Haplotype Reference Consortium initially used. The value obtained is 0.0047 compared to 0.0046 before, showing that imputation quality is unlikely to have affected our results.

9.3. Expression data

9.3.1. *ADCY9* and *CETP* expression quantification from RNAseq data

By analysing more in depth the *ADCY9* gene and its isoforms, we noticed that a considerable proportion of reads were assigned to a specific isoform, *ADCY9*-205 (ENST00000574721.1), a 2.4 Kb long retained intron that does not have a validated status. It was removed from the gene definition file (GTF) to remove any noise from spurious transcription. All GTEx data was therefore reprocessed for *ADCY9* and *CETP* by the same pipeline (see below) to obtain transcription levels per sample at the gene level, consistent for the two genes across cohorts.

For each eQTL dataset (GEUVADIS, GTEx, CARTaGENE), we recalculated the top 5 PCs using genotype data with flashPCA2. For duplicated samples, we kept the sample with the highest read count in the library and removed samples which had a total of less than 10 million of reads. We trimmed the sequencing reads from Illumina adaptors and bad quality ends (BQ>20) using TrimGalore!. We mapped the alignment files on Hg38 human genome reference using STAR v2.6.1a [215] with the Ensembl 87 genome annotation, then estimated count for each gene using RSEM v1.3.1 [218]. For GTEx, we separated each tissue at this step, then removed tissues with less than 50 samples, leaving samples from 49 different tissues to avoid over-interpretation due to low sample size while maximizing the number of tissues to be tested. We kept the genes which had more than 6 reads in at least 20% of the sample. We then normalized expression data using limma (TMM normalization) [332] and voom [223]. We calculated PEER factors [233] on the normalized expressions. For all sex-stratified analysis, we kept sex-stratified tissues that had at least 50 samples, and recomputed PEER factors with samples from only one sex (which we term sPEER factors).

To test if *ADCY9* and *CETP* expression is correlated across tissues in humans, we used data GTEx and performed a linear regression correcting for the first 5 PCs, age, sex, the collection site (SMCENTER), the sequencing platform (SMGEBTCHT) and total ischemic time (TRISCHD). We find that *ADCY9* and *CETP* gene expression levels are negatively correlated (and significantly($p<0.05$) so for Adipose-Subcutaneous, Adrenal Gland, Artery Coronary, Artery Tibial, Brain-Cerebellar Hemisphere/Cerebellum/Cortex/Frontal Cortex/Putamen (basal ganglia), Breast-Mammary Tissue, Esophagus-Gastro esophageal Junction/Muscularis, Heart-Left Ventricle, Lung, Prostate, Small Intestine-Terminal, Spleen, Stomach, Uterus, Whole blood), except in skin tissues and cells cultured fibroblast, for which a significant positive correlation is found (Supplementary figure 2.21).

9.3.2. Expression Quantitative Trait Loci (eQTL) analysis for rs1967309 and rs158477

We first looked at the effects of the SNPs independently on their respective genes. The covariates include the first 5 PCs, age (except for GEUVADIS, information not available), sex, as well as PEER factors, calculated to take into account hidden factors. In GTEx, we added additional covariates: the collection site (SMCENTER), the sequencing platform (SMGEBTCHT) and total ischemic time (TRISCHD). One limitation is there is no standardized way of deciding how many PEER factors to include. We tested the robustness of results to the inclusion of different numbers of PEER factors in the models and we report them all for GEUVADIS, CARTaGENE (CaG) and GTEx for transparency (Supplementary figures 2.22-2.24). The maximum number of PEER factors considered follows recommendation by GTEx based on sample size for each tissue.

SNP rs1967309 is a cis eQTL of *ADCY9* in whole blood in CARTaGENE (p-value=4.46 x 10-13, β= -0.10, N=728, 10 PEER factors) with AA individuals having increased *ADCY9* expression compared to GG individuals. This effect is replicated in whole blood samples from GTEx (N=559), and several other tissues in GTEx (with esophagus being the most significant), but some tissues show an inversion of the direction of the effect, such as lung and thyroid. These results may differ from GTEx reported eQTL results, because of the removal of *ADCY9*-205 isoform (see above), the different expression normalization method and filters by ethnicity applied here. SNP rs158477 is found as a cis eQTL of *CETP* in GEUVADIS (p-value =1.69 x 10^{-4}, β=0.26) lymphoblastoid cell lines, replicated in GTEx

(EBV transformed lymphocytes) as well as in many other GTEx tissues. Tissues with p-value below 0.05 across most PEER factors values (if not all) are: adipose tissues, hippocampus, liver, lung, small intestine, muscle, stomach, thyroid, with GG genotype having consistently less *CETP* expression than AA.

We next tested whether the SNPs are trans eQTL for the genes. We found nominally significant associations between rs1967309 and *CETP* expression in the ovary (p-value=0.0017, N=138, max PEER factors = 15) and hippocampus (p-value=0.049, N=150, max PEER factors = 30), results that are stable across PEER factors values, two tissues for which rs1967309 is not significantly associated with *ADCY9* expression (p-value>0.05). We found nominally significant associations between rs158477 and *ADCY9* expression in the brain-cerebellar hemisphere and liver.

We next evaluated the interaction effect between rs1967309 and rs158477 on gene expression levels, despite somewhat low statistical power, especially given the small number of samples by tissue. Because the appropriate value of the number of PEER factors to be included on the model is not obvious, we report all values of PEER until the maximum suggested by GTEx for the GTEx tissues (15 for N<150, 30 for 150 $\leq$ N<250, 45 for 250 $\leq$ N<350, and 60 for N$\geq$350). We also required that the interaction term rs1967309*rs158477 for a tissue had p-values under 0.1 for a majority of values of the number of PEER factors included, to qualify a result to be suggestive of an interaction effect. In the GEUVADIS dataset, which has 287 samples, the interaction was significant on *CETP* expression and stable across PEER factors (Supplementary figure 2.22a), which could mean that the effect of a SNP could be modulated by the other SNP. To evaluate this effect further and make sure this is not due to outlier effects or other statistical flukes, we stratified by genotypes of each SNP and investigated the effect of the other SNP on *CETP* expression. We used a linear regression with the same covariates mentioned above. We first stratified by the genotype of rs1967309, then evaluated the effect of rs158477 on *CETP* expression. SNP rs158477 is significant in the AA of rs1967309 (p-value=0.03, β=0.45, OR=[1.05-2.36], n=46) and AG (p-value=0.009, β=0.24, OR=[1.07-1.53], n=143), but not for GG (p-value=0.58, β =0.07, OR=[0.83-1.40], n=96) (Figure 2.10b), potentially showing a mitigation of the eQTL effect of rs158477 for each alternative allele of rs1967309 on *CETP* expression. We also evaluated the effect of rs1967309 when we stratified by rs158477 on *CETP* expression. SNP rs1967309

is significant only for GG of rs158477 (p-value=0.05, β=0.3, OR=[1.00-1.83], n=72), and not for GA (p-value=0.51, β=-0.06, OR=[0.77-1.14], n=139) nor AA (p-value=0.68, β=-0.07, OR=[0.66-1.30], n=74). The second dataset that we used was the GTEx, in which we evaluated 49 tissues. Since the effects across tissues are likely not independent, we did not correct for multiple testing, keeping a suggestive threshold at 0.10 and a significant threshold at 0.05, but those values need to be reached for a majority values of the number of PEER factors included to be convincing. Among the 49 tissues (Supplementary figure 2.23), those with p-values under 0.10 for several numbers of PEER factors are hippocampus (N=150), hypothalamus (N=156), brain spinal cord (cervical c-1) (N=114), substantia nigra (N=100) and skin sun-exposed (N=507). Among those tissues, rs1967309 is only a cis-eQTL of *ADCY9* in the substantia nigra.

Since the selective pressure differ between sexes, we stratified our expression analysis by sex. For *CETP* expression analysis, there are no consistent signals for GEUVADIS, possibly reflecting lack of power or that there is no sex-specific effect in lymphoblastoid cell lines (Supplementary figure 2.22c). However, in CaG, the significant interaction found is present only in male (again for the genotypic coding, Supplementary figure 2.22d). In GTEx, we see significant interaction effects in males in tissues that had signals with sex-combined, such as brain hippocampus (N_{male}=105), hypothalamus (N_{male}=112) and spinal cord cervical (N_{male}=72), skin sun-exposed for *CETP* (N_{male}=330) (Figure 2.10d, Supplementary figure 2.24). We note that for most brain tissues, the low sample size in females does not allow to conclude on the presence of an interaction effect in that sex. In these tissues, the direction of the effect in males is reversed compared to what is observed in GEUVADIS with sexes combined (Figure 2.10b), whereas the highly significant result in skin shows an effect consistent with the sex-combined GEUVADIS result (p-value=0.0017, β=-0.32). More specifically, in the sun-exposed skin samples, in rs1967309 AA males, copies of the rs158477 A allele increase *CETP* expression by 0.49 (95% CI 0.12-0.87) on average. In rs1967309 AG males, the effect of rs158477 is null ($p-value_{AG}$=0.33) and the effect of the rs158477 A allele is suggestive in rs1967309 GG individual ($p-value_{GG}$=0.10) with a decrease of the *CETP* expression. Conversely, if we look at the effect of rs1967309 on *CETP* expression in skin sun-exposed when we stratified by rs158477 in males, SNP rs1967309 is neither a significant

eQTL of *CETP* in GG of rs158477 in males (p=0.11, n=89) nor for GA (p=0.65, n=164), but is a significant eQTL for AA males (p=0.026, β=-0.46, OR=[-0.87, -0.06], n=76).

We also identified new tissues where the interaction is either suggestive or significant in females, in artery tibial (N_{female}=156), heart atrial appendage (N_{female}=97), spleen (N_{female}=69) and stomach (N_{female}=105) (Supplementary figure 2.24), with an effect reversed compared to the initial GEUVADIS result (Supplementary figure 2.22a). For the pituitary tissue, it is significant in both sex (additive coding in female and genotypic coding in male, N_{male}=156, N_{female}=63), but the direction of the additive coding is reversed between sexes, possibly explaining why the sex-combined analysis did not show any signal. We note that the newly discovered signals are mainly for females, indicating that the signal was hidden by the male effects (or absence of effects), likely because of higher sample sizes.

9.4. Experiments

9.4.1. Real-time PCR quantification

Reverse transcription was performed from 500 ng total RNA in a 20 ml reaction using High-Capacity cDNA Reverse Transcription Kit (Applied Biosystems cat #4368814). RNA quantification was assessed using Agilent RNA 6000 Nano Kit for Bioanalyzer 2100 System (Agilent Technologies). Primers were designed using the Beacon designer software v.8 (Premier Biosoft) (Supplementary table 2.5). The real-time PCR was carried out with SYBR-Green reaction mix (BioRad cat #1725274). The thermal cycling program was 3 min at 95°C for initial denaturation followed by 40 cycles of denaturation for 10 sec at 95°C, 30 sec annealing at 60°C and 30 sec extension at 72°C. qPCR assay was normalized with PGK1 and HBS1L genes.

9.4.2. Western Blot analysis.

200 ml of cell media from HepG2 transfected cells were concentrated using Amicon Ultra 0.5 ml 10 kDa cutoff units (cat #UFC501096) to 25 ml. Proteins were separated on 10% TGX-acrylamide gel. After O/N electrotransfer at 10 volts to PVDF membranes, CETP protein was determined using a primary anti-CETP rabbit monoclonal antibody (Abcam cat #ab157183) 1:1000 in 3% BSA, TBS, tween 20 0.5%, O/N 4°C, followed by HRP-conjugated

secondary antibody goat anti-rabbit 1:10 000 in 3% BSA 1h at room temperature. Detection was performed using Western Lightning ECL Pro (Perkin Elmer cat #NEL122001EA). Proteins levels were normalized with total proteins loaded.

9.5. Phenotype associations

9.5.1. Two-way and three-way interaction models in UK Biobank

With rs1967309 coded under the genotypic model, allowing to capture non-additive effects, we tested if the effect of the interaction term was significant for a phenotype Y using a likelihood ratio test (LRT) by comparing the following models:

$$Y\ rs158477 + rs1967309 + sex + age + PC(1-5)(m1)$$

$$Y\ rs158477 * rs1967309 + sex + age + PC(1-5)(m2)$$

$$Y\ rs158477 * rs1967309 * sex + age + PC(1-5)(m3)$$

We used the R function glm with family = "binomial" and compared models using the following: anova(a, b, test = "LRT"), with a=m1 and b=m2 to test for two-way interaction effects, and a=m2 and b=m3 to test for three-way interaction effects. Individually, SNP rs1967309 (f_A=39%) is nominally associated with heart rate, and rs158477 (f_G=47%) with the systolic blood pressure (Figure 2.13), both results being mainly driven by association in females. Both SNPs are nominally associated with waist-hip ratio, rs1967309 in females, rs158477 in males. None of these effects are genome-wide significant.

9.5.2. Phenotype associations in GTEx

In GTEx, we had the variable MHHRTATT (phv00169162.v8.p2) for cardiovascular disease. This variable is defined as Heart attack, acute myocardial infarction, acute coronary syndrome. GTEx also has phenotypes, including cardiovascular traits. The variable DTHFUCOD (First Underlying Cause Of Death) was used to identify individuals whose cause of death included Heart Attack/Stroke, Heart Disease, Acute Myocardial Infarction, Possible MI, who were considered as cases. From the 699 samples kept, we excluded 6 for which the phenotype was missing or unknown for MHHRTATT variable and for which the cause

of death was unrelated to heart disease, yielding a total of 130 cases and 563 controls. We added as covariates: sex, age and top 5 PCs. For this phenotype, neither rs1967309 nor rs158477 are associated with those phenotypes when taken alone. However, the interaction and both SNPs in the equation are significant (p-value$\leq$0.05) or close to be significant (p-value$\leq$0.10) for the phenotype (p-value$_{MHHRATT}$=0.01, Estimate$_{MHHRATT}$=-0.54) (Figure 2.13). Like for *CETP* expression, this means that for each G allele for rs1967309, there is a decrease of the effect of rs158477 on cardiovascular outcome. A difference with *CETP*'s expression is that there is an inversion of the direction of effect. In other word, for AA of rs1967309, directions of the effect of rs158477 are positive, with GG having less probability to have an event than AA ($Estimate_{MHHRATT}$=0.47), but for GG of rs1967309, estimates of rs158477 are negative ($Estimate_{EstimateMHHRATT}$=-0.79). Those results are consistent with the direction of dalcetrapib pharmacogenomic analysis. Considering that the GG genotype of rs158477, with less *CETP*'s expression, is a proxy for dalcetrapib, which is an inhibition of *CETP*, the same gradient is present for rs1967309. In AA of rs1967309, there is less heart disease with dalcetrapib [24]. In GG of rs1967309, there is more heart disease with dalcetrapib. More study of this interaction is needed to understand the mechanism. However, this could lead to new insights into the potential biological mechanism behind the pharmacogenomic association involving the gene *ADCY9* with cardiovascular outcome of dalcetrapib.

10. Supplementary figures

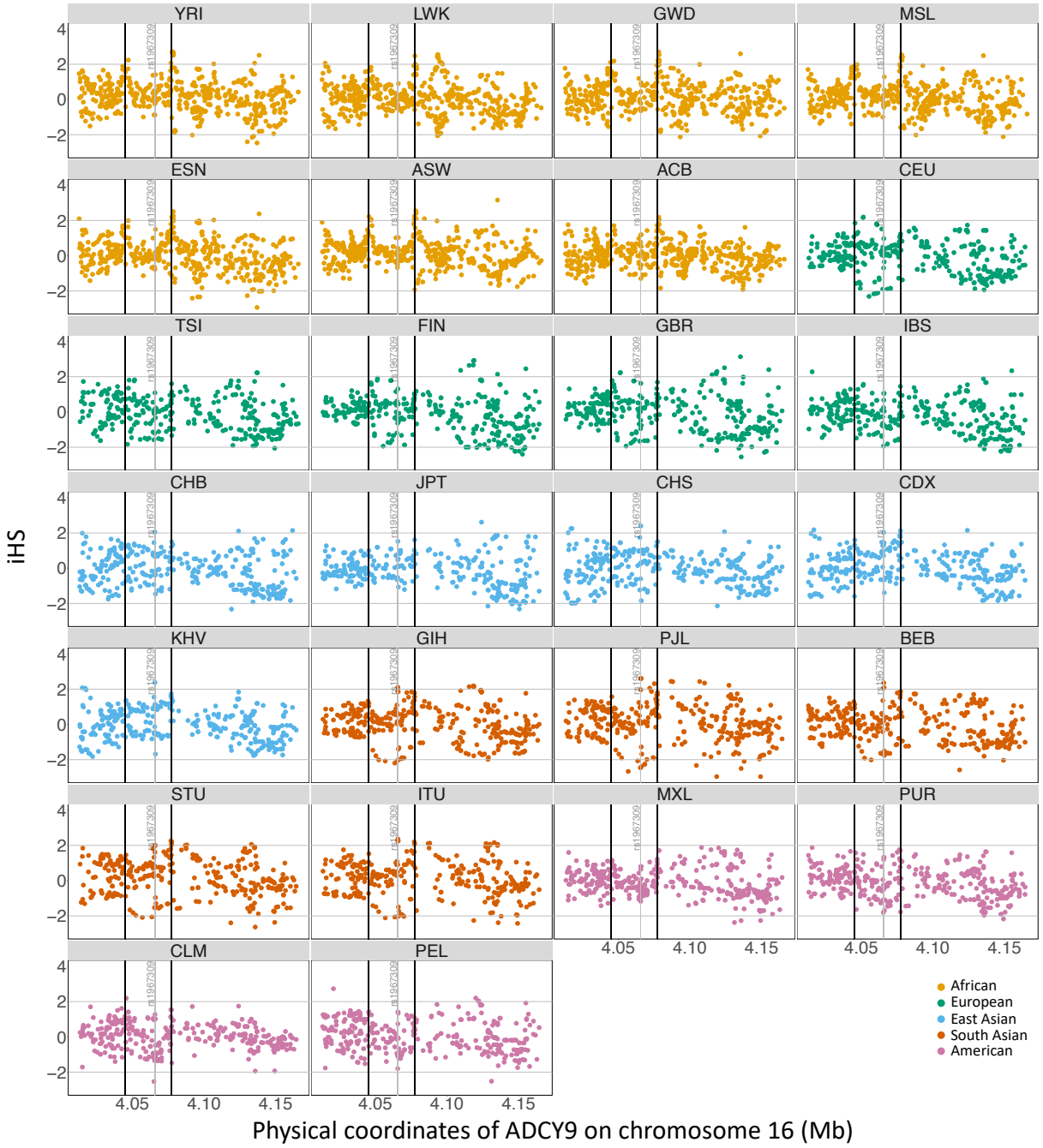

Fig. 2.16. Selection signature in *ADCY9*.
iHS values and recombination for all populations in the *ADCY9* gene. Vertical black lines represent the highest recombination rates around rs1967309 from 1000G population-specific genetic maps. Horizontal line represents the value at 2 and -2. Different colors represent one super population. In order of color: African, European, East Asia, South Asia and America. Abbreviations for the subpopulation of 1000G can be found here https://www.internationalgenome.org/category/population/

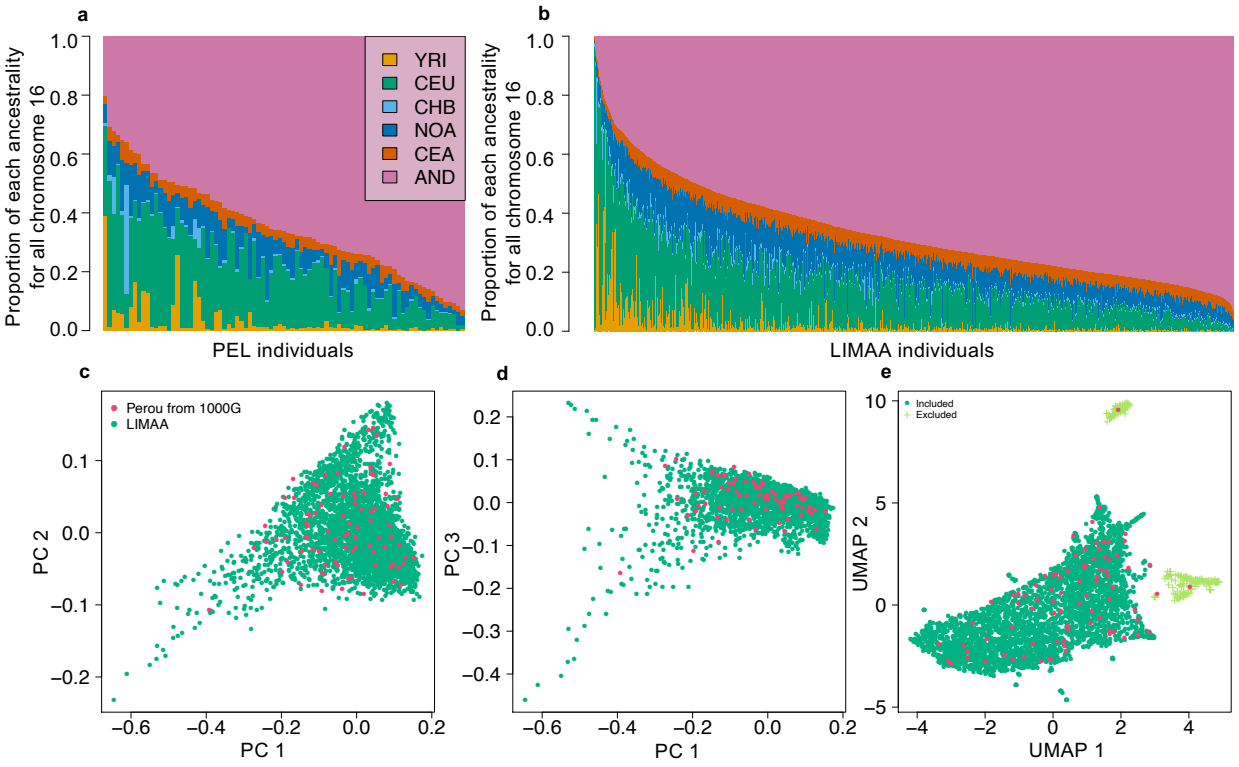

Fig. 2.17. Population structure of Peruvian from LIMAA and Peruvian from 1000G. Ancestry distribution on all chromosomes in the Peruvian from 1000G (a) and LIMAA cohort (b). Overall weighted proportion given by RFMix using reference populations from 1000G and Native American Genetic Dataset (NAGD) for the Peruvian population from 1000G (a) and from LIMAA cohort (b). 1000G populations YRI, CEU and CHB were chosen to represent African, European and Asian ancestry, respectively. (c,d) Principal Component Analysis using flashPCA on Peruvian from 1000G and LIMAA cohort. The top three PCs is shown. (e) UMAP analysis on the top 50 PCs. To limit confounders due to population structure, we excluded individuals in LIMAA coming from the two small groups identified by the UMAP (cross shaped light green symbols in (e)). Abbreviations for 1000G can be found here: https://www.internationalgenome.org/category/population/. Abbreviations for the Native American (NAGD): NOA: northern American; CEA: central American; AND: Andean.

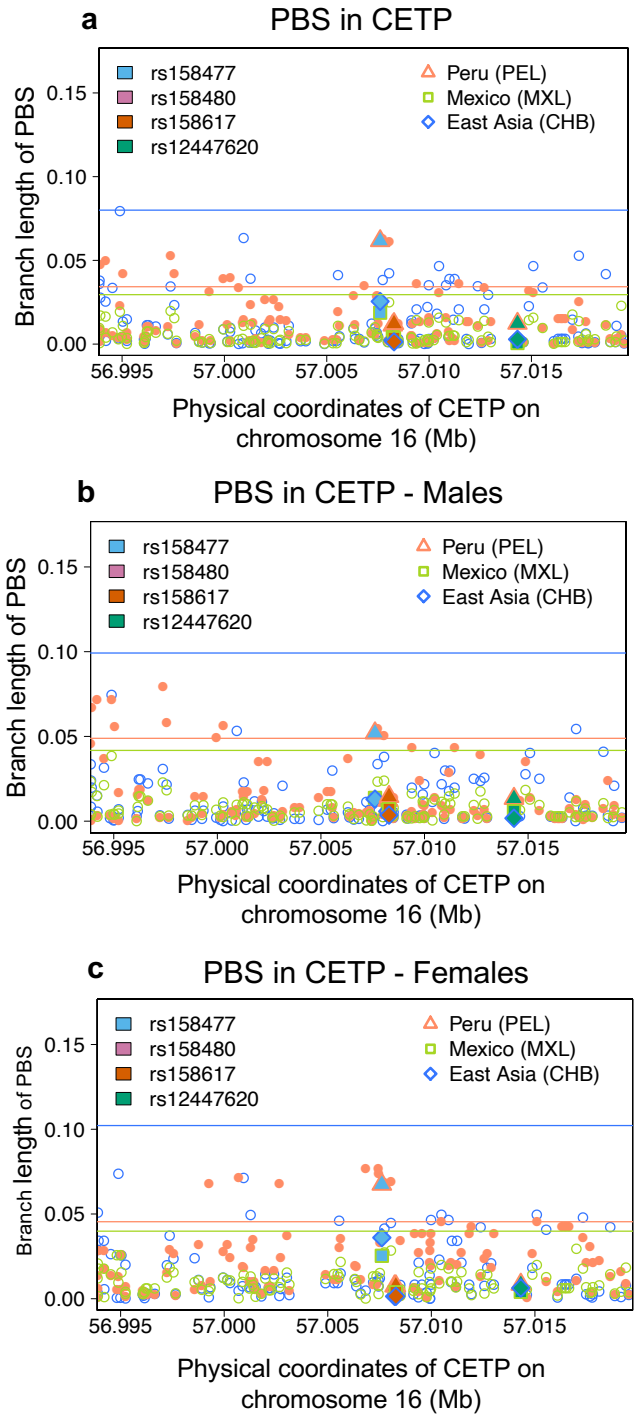

Fig. 2.18. Populational differentiation of *CETP* gene using PBS statistic. PBS values in the *CETP* gene, comparing the CHB (outgroup), MXL and PEL identified by different colors, overall (a), in males (b) and in females (c). Horizontal lines represent the 95th percentile PBS value genome-wide (a) or the chromosome 16 (b,c) for each population. Position with r^2 higher than the 99th percentile in the Peruvian population from the 1000G are represented by colored shape.

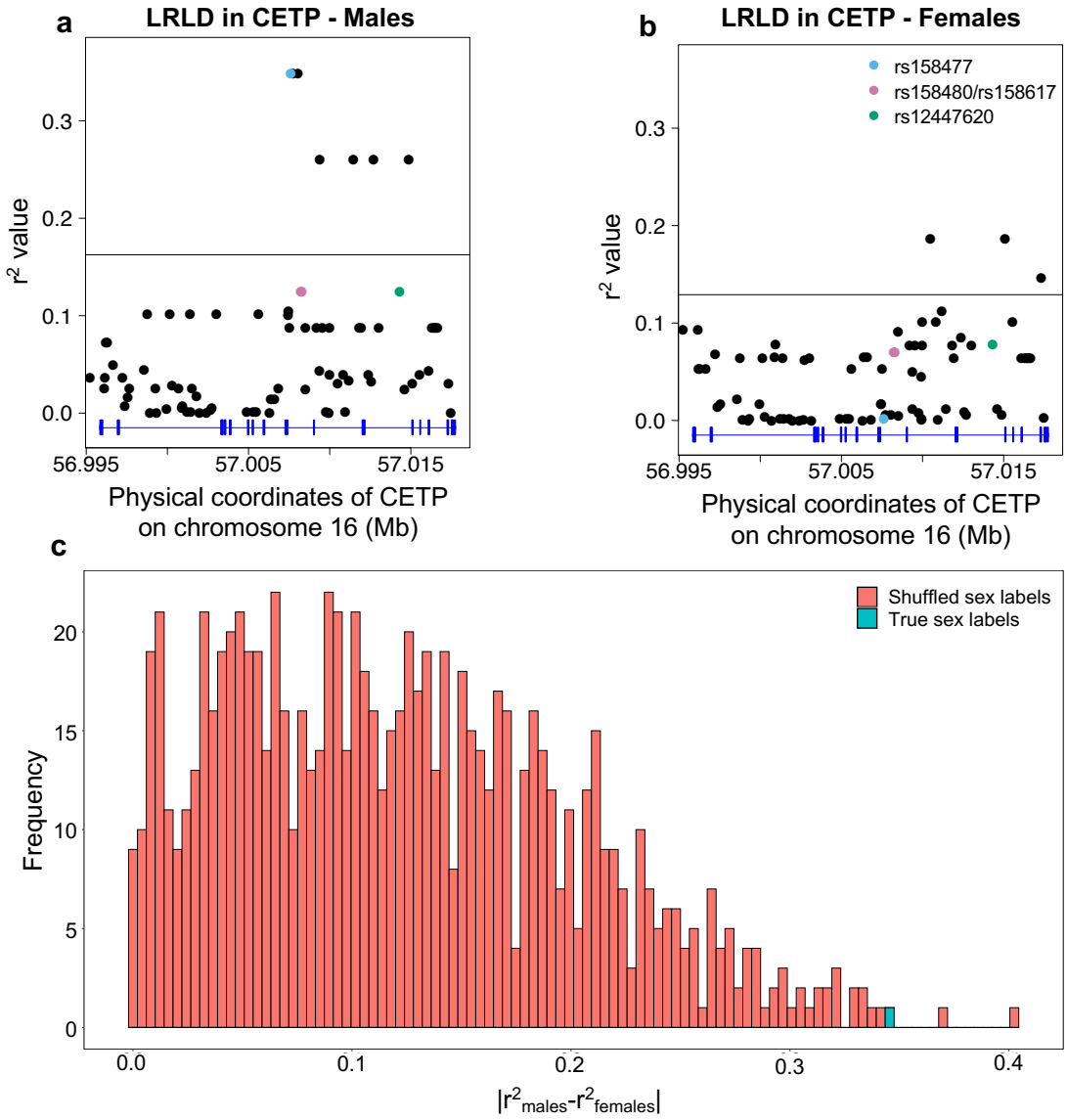

Fig. 2.19. Long-range linkage disequilibrium shown in *CETP* for the PEL population from 1000G, stratified by sex.
Genotype correlation (r^2) between the 3 loci identified in *CETP* (see Figure 2.2a) to be higher than the 99th percentile and all SNPs with MAF>5% in *ADCY9*, in males (a) and females (b). The horizontal black line is the 99th of all those comparisons between *ADCY9* and *CETP* by sex. (c) Distribution of absolute difference of genotype correlation values obtained during the permutation analysis that shuffled the sex label for rs1967309 and rs158477 (red), compared to the value obtain with the real sex labels (blue).

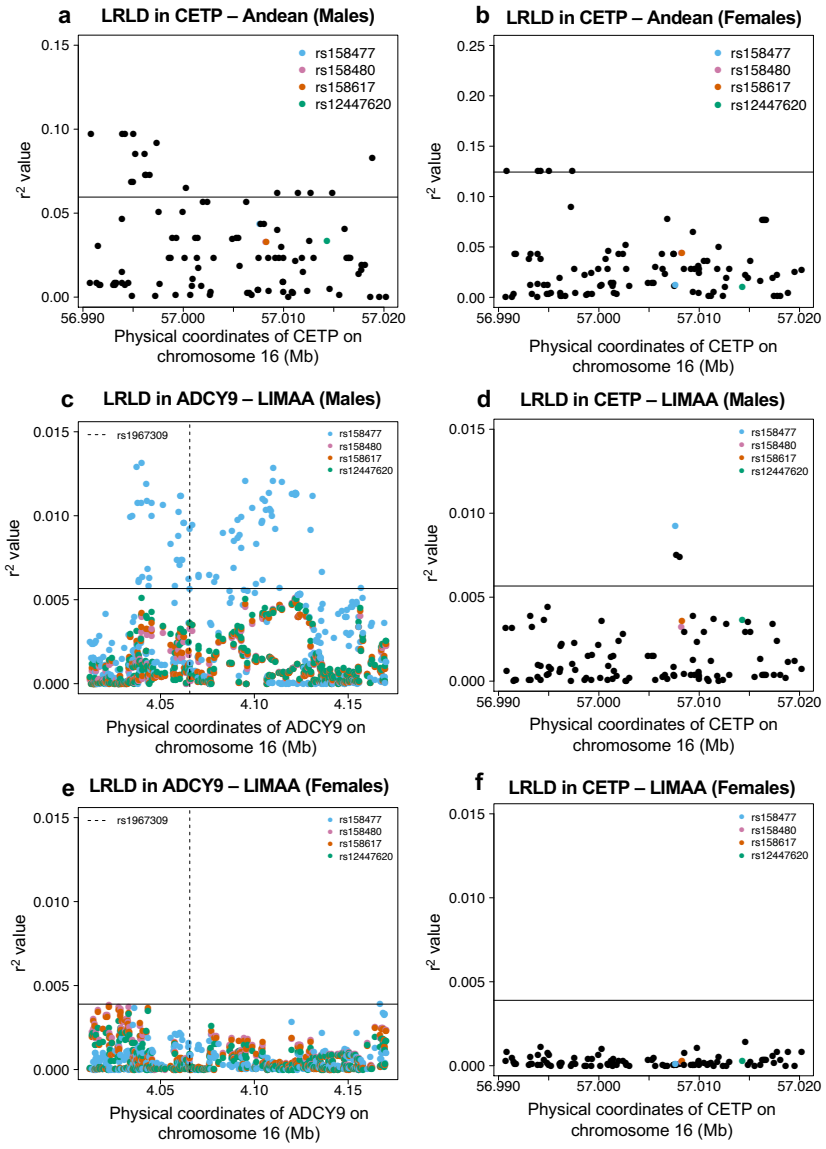

Fig. 2.20. Long-range linkage disequilibrium in the Andean population from NAGD (a,b) and LIMAA cohort (c-f).
(a,b,d,f) Genotype correlation (r^2) between rs1967309 and all SNPs with MAF>5% in *CETP*, for the Andean population from NAGD (a,b) and the LIMAA cohort (d,f). (c,e) Genotype correlation between the 3 loci identified in Figure 2.3a to be higher than the 99th percentile and all SNPs with MAF>5% in *ADCY9* in LIMAA. Males (N_{Andean}=54, N_{LIMAA}=1941) (a,c,d) and females (N_{Andean}=34, N_{LIMAA}=1302) (b,e,f) are shown separately. The horizontal line is the 95th (a,b) and 99th (c-f) percentile of all comparisons between *ADCY9* and *CETP* genes.

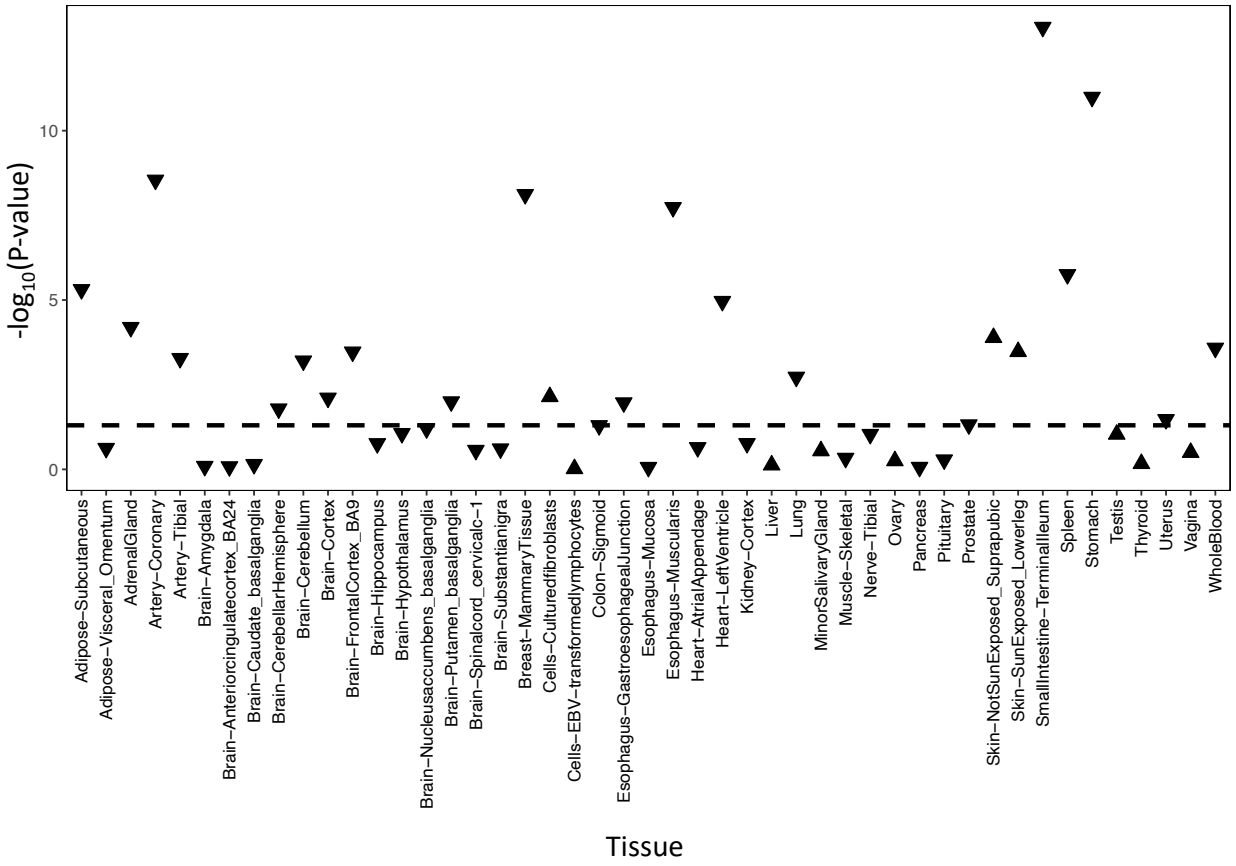

Fig. 2.21. Significance of the correlation between *ADCY9* and *CETP* expression across GTEx tissues.

P-values are presented on a -$\log_{10}$ scale and are obtained from a linear regression on normalized expression with correction for age, sex, top 5 PCs, ischemic time death, sequencing platform, and sequencing center. Regular triangles mean that both gene expression levels are positively correlated, inverted triangles mean that both gene expression levels are inversely correlated. The dashed line represents the p-value at 0.05.

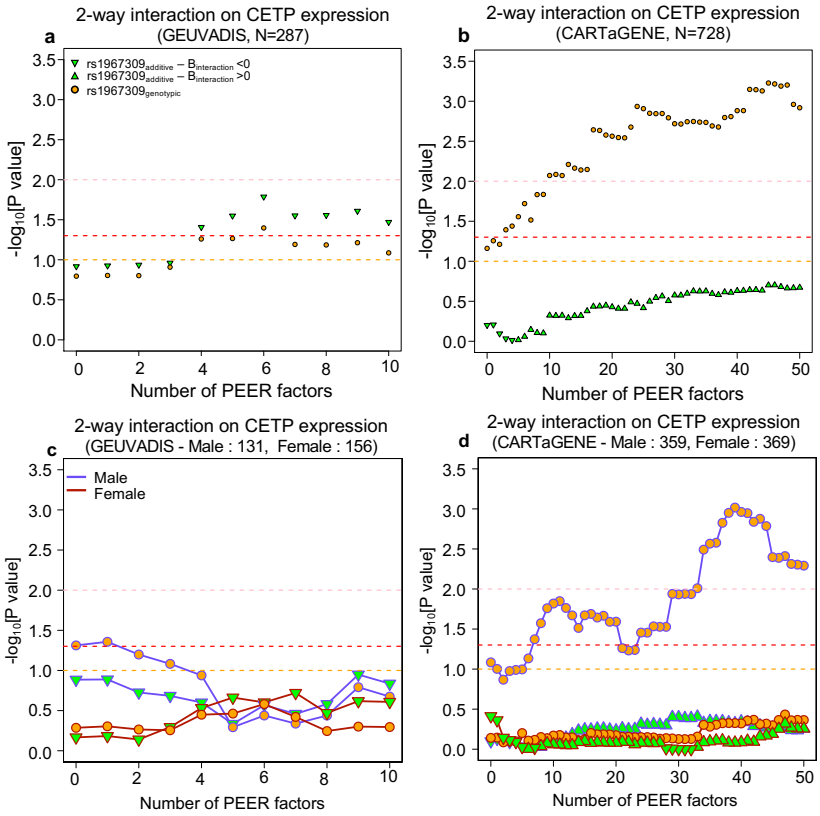

Fig. 2.22. Epistatic effects between rs1967309 and rs158477 on *CETP* expression in GEUVADIS (LCL, N=287) and CARTaGENE (Whole blood samples, N=728).
P-values are presented on a -$\log_{10}$ scale and are reported in function of the number of PEER/sPEER factors in GEUVADIS (LCL) (a,c) and CARTaGENE (b,d) in sex-combined (a,b) and sex-stratified (c,d) analyses. For all models, rs158477 is coded as additive (GG=0, GA=1, AA=2). In the additive model (green triangle), rs1967309 is coded as additive (AA=0, AG=1, GG=2), p-values are obtained using a linear regression in R. In the genotypic model (orange circle), rs1967309 is coded as a genotypic variable and p-values are obtained from a likelihood ratio test comparing models with and without the interaction term between the SNPs. The orange, red and pink lines represent p-values of 0.1, 0.05 and 0.01 respectively. The sample sizes reported are the number of individuals left after removing participants with missing genotypes for rs1967309 and/or rs158477. In (c,d), the color of the lines represents the sex label.

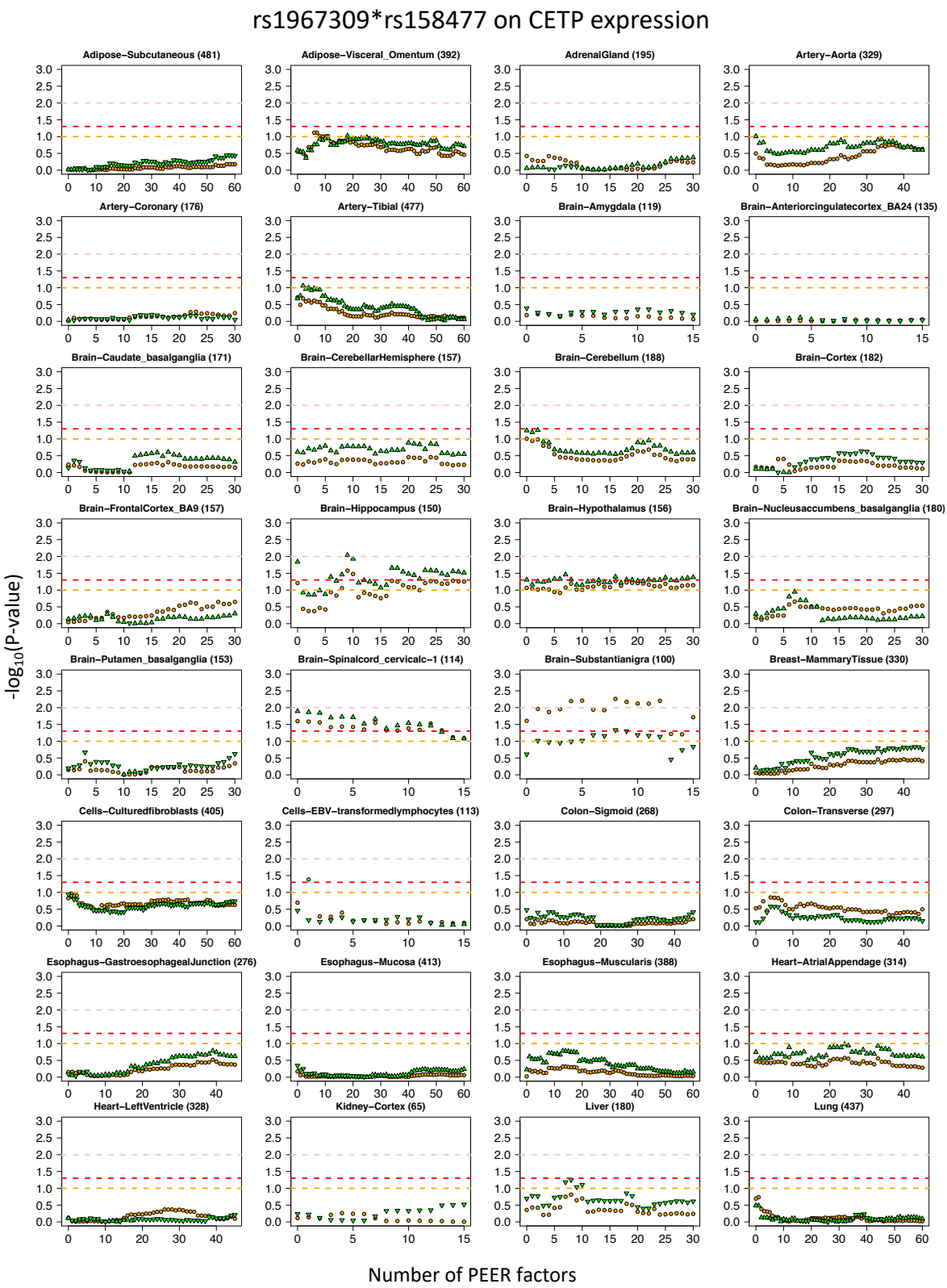

Fig. 2.23. Sex-combined epistatic effect p-values for the interaction between rs1967309 and rs158477 on *CETP* expression depending on the number of PEER factors in GTEx by tissue.

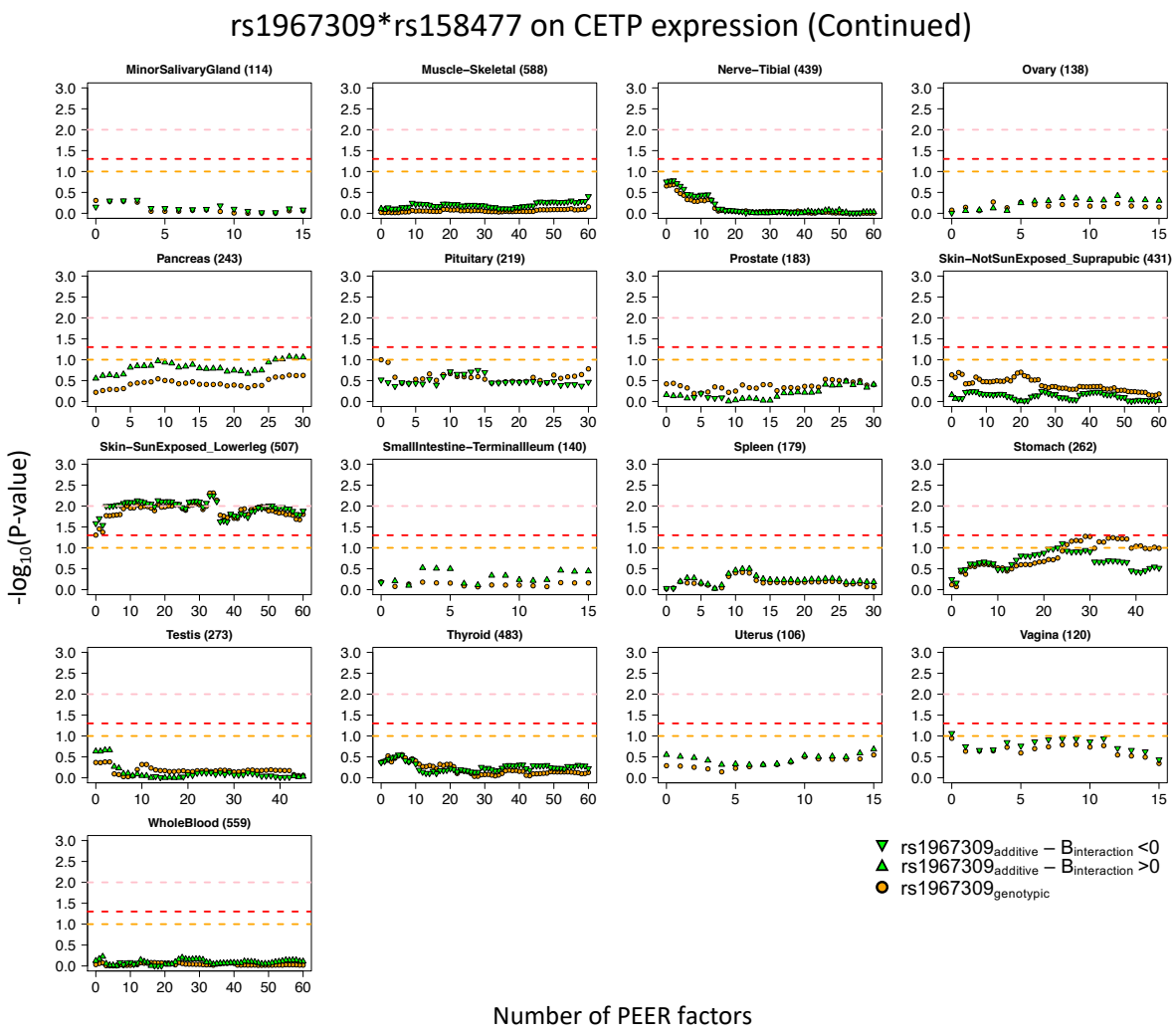

Fig. 2.23. Sex-combined epistatic effect p-values for the interaction between rs1967309 and rs158477 on *CETP* expression depending on the number of PEER factors in GTEx by tissue. P-values are presented on a -$\log_{10}$ scale. For all models, rs158477 is coded as additive (GG=0, GA=1, AA=2). In the additive model (green triangle), rs1967309 is coded as additive (AA=0, AG=1, GG=2), p-values are obtained using a linear regression in R. In the genotypic model (orange circle), rs1967309 is coded as a genotypic variable and p-values are obtained from a likelihood ratio test comparing models with and without the interaction term between the SNPs. The orange, red and pink lines represent p-values of 0.1, 0.05 and 0.01 respectively. The tissue type and the number of samples for each, used in the analysis, are reported in the titles of the subgraphs.

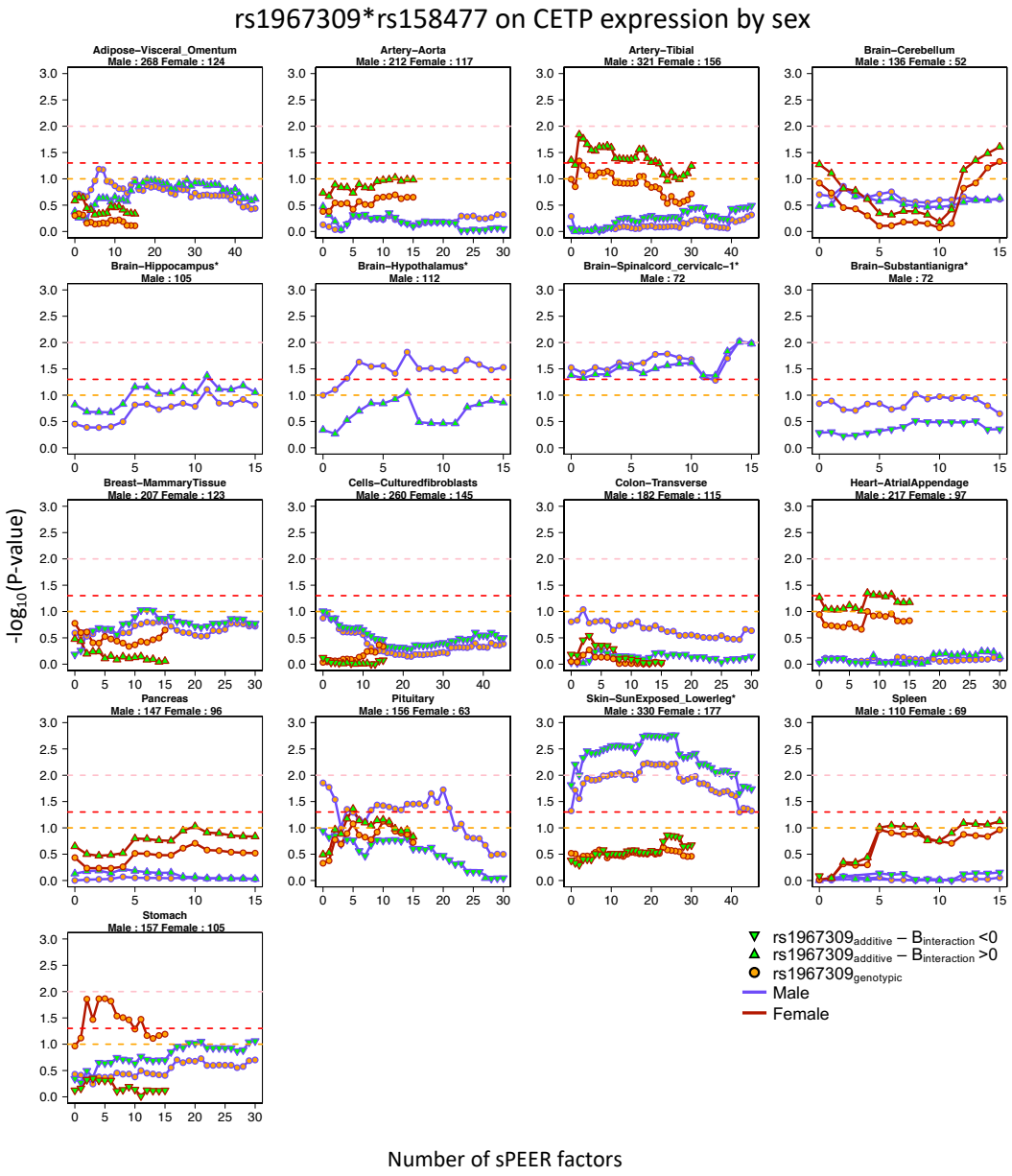

Fig. 2.24. Sex-specific epistatic effects between rs1967309 and rs158477 on *CETP* expression depending on the number of sPEER factors in GTEx by tissue.
P-values are presented on a -$\log_{10}$ scale. For all models, rs158477 is coded as additive (GG=0, GA=1, AA=2). In the additive model (green triangle), rs1967309 is coded as additive (AA=0, AG=1, GG=2), p-values are obtained using a linear regression in R. In the genotypic model (orange circle), rs1967309 is coded as a genotypic variable and p-values are obtained from a likelihood ratio test comparing models with and without the interaction term between the SNPs. The orange, red and pink lines represent p-values of 0.1, 0.05 and 0.01 respectively. The tissue type and the number of samples for each, used in the analysis, are reported in the titles of the subgraphs. The color of lines represents the sex label. Only tissues with at least one value under 0.10 are showed. Tissues with an asterisk (*) next to their title are tissues showing a the suggestive/significant effect in the sex-combined analysis.

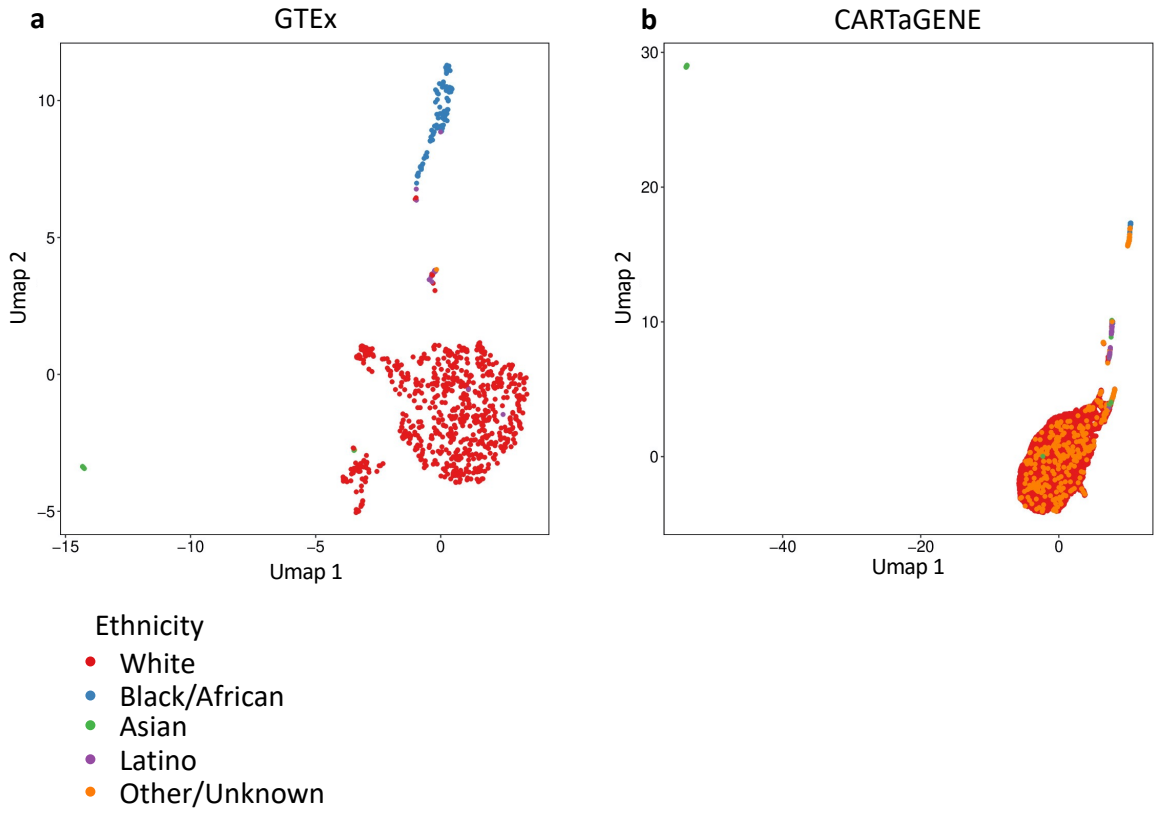

Fig. 2.25. Population structure in datasets analysed.
We estimate population structure using UMAP on the top 10 PCs generated with flashPCA2 on (a) GTEx (N=699) and (b) CARTaGENE (N=12,056) biobanks. The self-reported white non-Latino individuals were selected for further analyses.

11. Supplementary tables

Cohort	Population	Sex	Number	r^2	p-value *ADCY9-CETP*
1000G	YRI	All	108	0.0236	0.11
	CEU	All	99	0.0003	0.86
	GBR	All	91	0.0117	0.28
	CHB	All	103	0.004	0.53
	MXL	All	64	0.0007	0.83
	PEL*	All	85	0.0796	5.42×10^{-3}
		Male	41	0.3483	8.23×10^{-5}
		Female	44	0.0016	0.78
LIMAA	LIMAA	All	3243	0.0046	3.24×10^{-3}
		Male	1941	0.0097	3.71×10^{-3}
		Female	1302	0.0003	0.52
NAGD	Northern Amerind (NOA)	All	81	0.0084	0.44
		Male	27	0.0634	0.16
		Female	54	0.0699	0.07
	Central Amerind (CEA)	All	81	0.0281	0.12
		Male	34	0.0316	0.28
		Female	47	0.0257	0.24
	Andean (AND)	All	88	0.0293	0.04
		Male	54	0.0436	0.09
		Female	34	0.0125	0.55

Table 2.3. Long-range linkage disequilibrium analysis in three datasets, and in subsets of the cohorts.

Number of individuals (N) in each subset is reported. P-values correspond to the *ADCY9/CETP* empirical p-values computed as described in Section Long-range linkage disequilibrium in Methods. r^2 were obtained from the geno-r^2 option of vcftools software. For 1000G populations, abbreviations can be found here https://www.internationalgenome.org/category/population/. * Discovery cohort.

Variable ID	**UK Biobank variable location**	**Number of samples used for interaction**
Category 100011 - Blood pressure - Physical measures - UK Biobank Assessment Centre		
Pulse rate at baseline (Pulse rate) Units: bpm	Data-Field 102 (automatic entry) or Data-Field 95 (manual entry), to be derived as follows: · Pulse rate, automated reading (Data-Field 102) used mean of available measures for instance 0 (baseline) only. If a manual measure is available for an individual (Data Field 95 below) then do not use this automated reading (assumed to be abnormal). · Pulse rate (during blood-pressure measurement) (Data-Field 95), use Instance 0 (baseline). Use mean when there are multiple measures for a same individual.	All=395,319 Male=182,279 Female=213,040
Diastolic blood pressure at baseline (Diastolic BP) Units: mmHg	Data-Field 4079 (automatic entry) or Data-Field 94 (manual entry), as follow: · Diastolic blood pressure, automated reading: Data-Field 4079, use mean of available measures for instance 0 (baseline) only. If a manual measure is available for an individual (Data Field 94) then do not use this automated reading (assumed to be abnormal). · Diastolic blood pressure, manual reading: Data-Field 94, use mean of available measures for instance 0 (baseline) only.	All=395,384 Male=182,326 Female=213,058
Systolic blood pressure at baseline (Systolic BP) Units: mmHg	Data-Field 4080 (automatic entry) or Data-Field 93 (manual entry), as follow: 1) Systolic blood pressure, automated reading: Data-Field 4080, use mean of available measures for instance 0 (baseline) only. If a manual measure is available for an individual (Data Field 93) then do not use this automated reading (assumed to be abnormal). 2) Systolic blood pressure, manual reading: Data-Field 93, use mean of available measures for instance 0 (baseline) only.	All=395,353 Male=182,316 Female=213,037

Variable ID	**UK Biobank variable location**	**Number of samples used for interaction**
Category 100010 - Body size measures - Anthropometry - Physical measures - UK Biobank Assessment Centre		
Waist circumference at baseline (Waist circumference) Units: cm	Data field 48, use mean of available measures for instance 0 (baseline) only.	All=395,006 Male=182,089 Female=212,917
Hip circumference at baseline (Hip circumference) Units: cm	Data field 49, use mean of available measures for instance 0 (baseline) only.	All=394,651 Male=181,988 Female=212,663
Waist-hip ratio	Compute waist/hip	All=394,944 Male=182,056 Female=212,888
Weight Units: Kg	Data-Field 21002 (automatic entry) or Data-Field 3160 (manual entry), as follow: 3) Weight: Data-Field 21002, use mean of available measures for instance 0 (baseline) only. Only if unavailable, then use: 4) Weight, manual reading: Data-Field 3160, use mean of available measures for instance 0 (baseline) only.	All=394,377 Male=181,732 Female=212,645
Height Units: cm	Data-Field 50 or 12144. 5) Standing height: Data Field 50, used mean of available measures for instance 0 (baseline) only. Only if unavailable, then use: 6) Height: Data-Field 12144, used mean of available measures, as this is a singular instance field	All=394,871 Male=181,969 Female=212,902
UK Biobank BMI (BMI) Units: Kg/m2	Data field 21001, used mean of available measures for instance 0 (baseline) only.	All=394,173 Male=181,705 Female=212,468

Variable ID	**UK Biobank variable location**	**Number of samples used for interaction**
Category 100009 - Impedance measures - Anthropometry - Physical measures - UK Biobank Assessment Centre		
Trunk fat percentage (% Trunk fat) Units: %	Data field 23127, use mean of available measures for instance 0 (baseline) only.	All=388,569 Male=178,837 Female=209,732
Body fat percentage (% Body fat) Units: %	Data field 23099, use mean of available measures for instance 0 (baseline) only.	All=388,600 Male=178,752 Female=209,848
Basal metabolic rate Units: KJ	Data field 23105, use mean of available measures for instance 0 (baseline) only.	All=388,585 Male=178,758 Female=209,827
Whole body water mass Unites: Kg	Data field 23102, use mean of available measures for instance 0 (baseline) only.	All=388,719 Male=178,881 Female=209.838
Category 100020 - Spirometry - Physical measures - UK Biobank Assessment Centre		
Forced vital capacity (FVC) Units: L	Data field 20151, use mean if more than one measure.	All=297,461 Male=138,909 Female=158,552
Forced expiratory volume in 1-second (FEV1) Units: L	Data field 20150, use mean if more than one measure.	All=297,499 Male=138,937 Female=158,562
Category 100057 - Sleep - Lifestyle and environment - Touchscreen - UK Biobank Assessment Centre		
Sleep duration Units: hours/day	Data field 1160, use mean of available measures for instance 0 (baseline) only.	All=393,133 Male=181,452 Female=211,681
Category 100072 - Early life factors - Verbal interview - UK Biobank Assessment Centre		
Birth weight Units: Kg	Data field 20022, use mean if more than one measure.	All=227,244 Male=89,715 Female=137,529

Variable ID	**UK Biobank variable location**	**Number of samples used for interaction**
Category 717 - Biomarkers		
Apolipoprotein A1 (ApoA) Units : g/L	Data field 30630, use mean of available measures for instance 0 (baseline) only. Standardized using the mean : (x-mean)/sd	All=413,138 Male=190,454 Female=222,684
High Density Lipoprotein (HDL-c) Units : mmol/L	Data field 30760, use mean of available measures for instance 0 (baseline) only. Standardized using the mean : (x-mean)/sd	
Lipoprotein (a) (Lp(a)) Units : nmol/L	Data field 30780, use mean of available measures for instance 0 (baseline) only. Standardized using the mean : (x-mean)/sd	
C-Reactive Protein (CRP) Units : mmol/L	Data field 30710, use mean of available measures for instance 0 (baseline) only. Ln transformation, then standardized using the mean: (x-mean)/sd	
Low Density Lipoprotein (LDL-c) Units : mmol/L	Data field 30790, use mean of available measures for instance 0 (baseline) only. Standardized using the mean : (x-mean)/sd	
Apolipoprotein B (ApoB) Units : g/L	Data field 30640, use mean of available measures for instance 0 (baseline) only. Standardized using the mean : (x-mean)/sd	
Category of operation procedure codes (OPCS) and hospitalization or death record codes(ICD9/ICD10)		
Coronary artery disease (CAD)	Prevalent or incident	(cases/controls) All=413,138 (44,713/368,425) Male=190,454 (29,910/160,544) Female=222,684 (14,803/207,881)

Variable ID	**UK Biobank variable location**	**Number of samples used for interaction**
Myocardial Infarction (MI)	Prevalent or incident	(cases/controls) All=413,138 (18,559/394,579) Male=190,454 (13,812/176,642) Female=222,684 (4,747/217,937)

Table 2.4. Details on metabolic and clinical variables extracted from the UK Biobank

Species	Gene	Strain	Sequence
Human	*ADCY9*	Forward	5' CTGAGGTTCAAGAACATCC 3'
		Reverse	5' TGATTAATGGGCGGCTTA 3'
	CETP	Forward	5' CTACCTGTCTTTCCATAA 3'
		Reverse	5' CATGATGTTAGAGATGAC 3'
	HBS1L	Forward	5' ACAAGAATGAGGCAACAG 3'
		Reverse	5' AGATACTCCAGGCACTTC 3'
	PGK1	Forward	5' GTGGAGGAAGAAGGGAAG 3'
		Reverse	5' AAGCATCATTGACATAGACAT 3'

Table 2.5. Primers sequence for real-time PCR quantification in HepG2 cells for the KD-*ADCY9* and KD-*CETP* experimentations

Chapitre 3

CETP alternative splicing variation impacts human traits

Contributions à ce chapitre

Mes contributions à l'article inclut dans ce chapitre sont les suivantes en tant que première auteure :

- Formulation des hypothèses de recherche
- Planification et réalisation des analyses en transcriptomique
  - Traitements des données d'expression (filtrage, normalisation, calcul des facteurs cachés PEER)
  - Analyses des tests statistiques (eQTL, sQTL, régressions logistiques, épistasie)
  - Quantification des jonctions d'épissage alternatif (PSI) avec les logiciels MAJIQ et ASpli
- Réalisation des analyses de randomisation mendélienne
  - Calcul des variables instrumentales pour l'exposition
  - Performance des analyses statistiques
- Rédaction intégrale de l'article
- Création de toutes les figures

Ce travail n'aurait pas été possible sans l'aide des personnes suivantes :

- Julie Hussin
  - Supervision du projet
  - Aide à la conceptualisation et la rédaction de l'article
- Marc-André Legault

  - Enseignement des notions de randomisation mendélienne (MR)
  - Aide à la conceptualisation des analyses de MR et à l'évaluation des résultats
- Jean-Christophe Grenier
  - Aide à l'obtention de la base de données GTEx
  - Aide à la relecture de l'article
- Marie-Pierre Dubé
  - Co-supervision du projet
  - Aide à sa conceptualisation
- Jean-Claude Tardif et Éric Rhéaume
  - Aident à la discussion de résultats

CETP alternative splicing variation impacts human traits

by

Isabel Gamache[1,2], Marc-André Legault[3,4], Jean-Christophe Grenier[2], Éric Rhéaume[2], Jean-Claude Tardif[1,2], Marie-Pierre Dubé[1,2,5], and Julie G Hussin[1,2]

([1]) Faculty of Medicine, Université de Montréal, Montréal, Québec, Canada

([2]) Montreal Heart Institute, Canada

([3]) McGill University, Faculty of Science, Department of Computer Science, Montréal, Canada

([4]) Mila, Montréal, Canada

([5]) Université de Montréal Beaulieu-Saucier Pharmacogenomics Centre, Canada

The article will soon be submitted to a scientific journal.

1. Abstract

The cholesteryl ester transfer protein (CETP) is an important protein in reverse cholesterol transport and has been identified as a significant factor associated with cardiovascular disease (CVD), making it a widely studied pharmaceutical target. Three protein-coding isoforms of *CETP* exist, distinguished by the alternative splicing of one exon each. The isoform primarily responsible for cholesterol-related functions in the plasma is well studied, but specific functions of each isoform remain poorly understood. In this study, we demonstrate the significance of considering *CETP*'s isoforms in analyses of human traits. Using bulk RNA-seq data from multiple tissues, we characterized the expression patterns and genetic regulation determinants of *CETP* transcripts. Leveraging publicly available GWAS summary statistics, we conducted multivariable Mendelian Randomisation (MVMR) to estimate the impact of variation in isoform proportions on phenotypes, highlighting the importance of *CETP*'s isoforms in pituitary and thyroid glands. Furthermore, we uncovered tissue-specific

associations between *CETP*'s isoforms and CVD-associated phenotypes. Additionally, we observed that the epistatic interaction previously reported between *CETP* and *ADCY9*, a gene implicated in modulating a *CETP* modulator's response, may be mediated through the regulation of alternative splicing of exon 9. Our results underscore the importance of a comprehensive understanding of *CETP*'s isoforms, which can significantly impact both fundamental and clinical research efforts.

Keywords: CETP, Alternative Splicing, Transcriptomics, Phenotypic association analysis, Mendelian Randomisation

2. Introduction

The cholesteryl ester transfer protein (CETP) mediates the exchange of cholesterol esters (CE) and triglycerides (TG) between high-density lipoproteins (HDL) and lower density lipoproteins (LDL) [36, 44]. It is expressed in a wide variety of tissues [40, 41, 42, 333], and plasmatic CETP is mainly produced by liver and adipocytes [40]. Many factors can modulate its production, including dietary cholesterol [46, 47, 334], thyroid stimulating hormones (TSH) [335], hyperinsulinemia [336], obesity [337] and acute infection with Epstein-Barr virus (EBV) [338]. CETP's plasmatic function is mostly linked to the modulation of lipoprotein metabolism, but intracellular functions have also been reported, notably linked with lipid homeostasis, such as lipid transport from the endoplasmic reticulum (ER) to lipid droplets, TG biosynthesis, lipid storage, or its cholesterol content in the membrane [36, 44, 49, 51, 52, 339, 340, 341, 342]. CETP's impact on lipid concentration in different lipoprotein particles has led to the hypothesis that its pharmacological inhibition leading to increased HDL-cholesterol (HDL-C) may be beneficial in treating cardiovascular diseases (CVD) [53, 343, 344, 345], resulting in the development of several CETP inhibitors [76, 77, 79, 81, 321]. These inhibitors act on lipoprotein-associated lipid profile at the plasmatic levels and can also act within cells [346, 347, 348, 349, 350]. Other phenotypes have also been associated with this gene, such as response to sepsis [54, 55, 56] and age-related macular degeneration (AMD) [59, 60, 351].

While previous expression quantitative trait loci (eQTL) studies, examined the impact of genetic variants on global *CETP* expression levels, there is a lack of research exploring *CETP*'s expression regulation at the mRNA isoform-level. *CETP* has three reported

protein-coding transcripts: the full-length transcript (Ensembl CETP-201 transcript), the exon 9-spliced out transcript (CETP-202), and a transcript featuring an alternative first exon (CETP-203). A majority of studies have focused on CETP-201, as it is the predominant form that is secreted from most of cells and exerts plasmatic activities [40]. Meanwhile, the protein derived from CETP-202, that is not secreted, functions as an inhibitor, binding to CETP-201 protein and impeding its secretion [42, 66, 67, 352, 353]. Notably, to our knowledge, there is currently no clear knowledge about CETP-203's function, which is nevertheless reported as a coding protein in Ensembl [64].

Alternative splicing refers to the process by which exons from the same gene are joined in different combinations during mRNA maturation, leading to different isoforms which can have different functions and effect on phenotypes [43, 202, 354, 355, 356]. Genomic regions surrounding and within an alternative exon are important for the regulation of the splicing events since they can impact the recognition of splicing binding factors. In *CETP*, exon 9 is spliced in the *CETP-202* isoform and the exonic mutation rs5883 within this exon influences the recognition of a splicing factor, thereby modulating the occurrence of this splicing event [357, 358, 359]. This mutation was also found to be linked to HDL-C level and coronary artery disease (CAD) in a sex-specific manner [357]. Interestingly, an intronic mutation located near exon 9, rs158477, in strong linkage disequilibrium (LD) with rs5883 (D'>0.99 in the CEU population from 1000 Genomes project [252]), is involved in an epistatic interaction with *ADCY9* impacting *CETP* expression levels [360]. This epistatic interaction appears to have been under sex-specific selection in the Peruvian population, with an over-representation of some chromosome 16 haplotypes involving rs158477 in *CETP* and rs1967309 in *ADCY9* that differ between males and females. Rs1967309 genotype located at approximatively 50 MB from the *CETP* gene determines the CVD benefits in response to the CETP inhibitor dalcetrapib [24].

In this study, we characterized the genetic regulation of expression and splicing for each *CETP* isoform using data from the Genotype-Tissue Expression (GTEx) project. Furthermore, we demonstrate that changes in the proportion of *CETP*'s isoforms have a significant impact on various phenotypes using Mendelian Randomisation (MR) analysis based on publicly available summary statistics. Moreover, our findings indicate that changes in the

proportion of *CETP* isoforms are associated with CAD in a tissue-specific manner. Additionally, in line with previous results on the evolutionary link between *CETP* and *ADCY9* genes [360], our study shows that the interaction between these genes has a significant impact on alternative splicing of exon 9. Lastly, variations in the proportion of *CETP* isoforms may indicate important functions of *CETP* within the pituitary or thyroid glands, highlighting the need for further investigation into the role of these isoforms in physiological processes. Generally, our study highlights the crucial importance of studying the individual isoforms of *CETP*.

3. Methods

3.1. *CETP* transcript definitions

Gene-level *CETP* (ENSG00000087237) refers to the total gene expression, containing all isoforms count during its quantification. *CETP-201* (ENST00000200676.8) contains all 16 exons of *CETP*. *CETP-202* (ENST00000379780.6) has the same exons as *CETP-201*, except for alternative splicing of the 9th exon. *CETP-203* (ENST00000566128.1) differs from *CETP-201* only by an alternative exon 1.

3.2. Datasets

We used two datasets for which we had both RNA-seq data and genotyping. The first dataset is obtained from the Genotype-Tissue Expression v8 (GTEx) [116], accessed through dbGaP (phs000424.v8.p2, dbgap project #19088), which includes RNA sequencing across 54 tissues and 948 donors with genetic information available. The cohort contains mainly of European descent (84.6%), ranging in age from 20 to 79 years old. Due to the small sample sizes of GTEx, we applied PCA and ethnicity-based filtering methods to keep the largest homogeneous group, meaning that we only kept self-reported white non-Latino individuals, as described previously [360], resulting in 699 individuals for our analyses, comprising 66 % males and 34 % females. The second dataset used is the GEUVADIS dataset [253] from 1000 Genomes project, which is accessible at https://www.internationalgenome.org/data-portal/data-collection/geuvadis. We kept a total of 287 non-duplicated European samples (CEU, GBR, FIN, TSI).

3.3. BAM processing

For the GTEx (GRCh38) and GEUVADIS (GRCh37) [253] datasets, we extracted the region of *CETP* from the bam files. In GRCh37 (chromosome 16), the region spanned from 56,985,762 to 57,027,757, while in GRCh38 (chromosome 16), it ranged from 56,951,923 to 56,993,845. We performed this extraction using samtools [279], then we used Picard tools (Broad Institute, 2019) to get Fastq files including the unpaired reads. We trimmed the Illumina adaptors from sequencing reads and removed bad quality ends (BQ>20) using TrimGalore! [276]. The read files were then mapped to the GRCh38 human genome reference using STAR v2.6.1a [215]. During the mapping process, we utilized specific parameters such as outSAMstrandField, outFilterIntronMotifs (RemoveNoncanonicalUnannotated) and outSAMattributes (NH, nM, MD).

3.4. Expression analyses

For the GTEx dataset, pre-processing of expression data was done as described previously [360] : briefly, we performed PCA on genetic data, we computed PEER factors [273], transcripts were quantified using RSEM [218] and their expression were normalized using limma [332] and voom [223]. Tissues with less than 50 samples were excluded, resulting in samples from 49 different tissues being retained for analysis. In the GTEx dataset, a gene-wide expression quantification quantitative trait locus (eQTL) analysis was performed to assess the association between mutations and gene expression. The genomic region from 56,941,980 to 57,003,666 (20,000bp before and after *CETP* locus) on chromosome 16 (GRCh38) was selected. Positions with MAF below 5% were removed, and only biallelic positions were retained. Indels and SNPs with a Hardy-Weinberg equilibrium p-value below 0.01 were also excluded, resulting in 174 positions for analysis. During the eQTL analyses, two-sided linear regressions on the *CETP* gene expression and its three protein-coding transcripts were done using R (v.3.6.0) [361]. Each SNP was coded based on the number of non-reference alleles. The covariates included the first five Principal Components (PCs) computed using FlashPCA2 [362], as well as age, sex, collection site (SMCENTER), sequencing platform (SMGEBTCHT), total ischemic time (TRISCHD), and PEER factors. For the PC analysis, we utilized the imputed genotyping dataset of GTEx v8 using the same filters as mentioned

above, and for the filtering based on linkage disequilibrium, we followed the recommendations provided by flashPCA2 (https://github.com/gabraham/flashpca). After the filtering process, we retained 100,986 SNPs for performing PCA. The maximum number of PEER factors considered followed the GTEx consortium recommendation based on sample size for each tissue. To assess potential differences in genetic regulation between expression of *CETP* at gene-level and its isoforms, the estimates of their eQTLs were compared using a t-test for each pair (gene-level *CETP* vs isoform, isoform vs isoform). Only SNPs that passed the threshold of p-value$<$0.05/[49 tissues * 7 LD blocks] = 0.0001 for at least one of the transcripts in the comparison were considered (Supplementary figure 3.8).

3.5. Alternative splicing analysis

We estimated Percent Spliced-In (PSI) values using ASpli [231] from the processes BAM files from above. In this study, PSI values estimated by ASpli represent the proportion of reads covering the splicing event, such as alternative exon 1, to the total number of reads covering the junction, including reads associated with alternative exon 1 and regular exon 1 (Figure 3.2a). The minimum read length was set to 80, and the maximum intron size considered was 10,000 bp. As an alternative approach, we also replicated all results using MAJIQ software [229] (Supplementary text 6, Supplementary Figures 3.15 and 3.16). We note that when using ASpli, the alternative exon 1 was not detected as a splicing event. To address this, we modified the coordinate of the start of the exon 1 of *CETP-203* in the GTF file to ensure overlap with exon 1 of *CETP-201* and *CETP-202*. After estimating PSI values, we filtered out samples with less than 10 reads covering the analyzed splicing junctions. We further restricted the analysis to junctions in tissues with at least 50 samples passing this filter, remaining five tissues for the alternative exon 1 (AS1) and eleven for the alternative splicing exon 9 (AS9) (Supplementary Figure 3.16). To estimate splicing quantitative trait loci (sQTL), we performed linear regressions using the same positions and covariates as used in the eQTL analyses. These covariates included the first five PCs, age, sex, collection site (SMCENTER), the sequencing platform (SMGEBTCHT), total ischemic time (TRISCHD), and PEER factors. We restricted the number of PEER factors to 15, as was done by GTEx. To assess tissue-specific genetic regulation of alternative splicing, we selected the top sQTL for AS1 and the top sQTL for AS9. We then evaluated if their effects on respective PSI

values exhibited heterogeneity across tissue. We considered values to be significant if their p-value, obtained using the `metagen` function from the meta package in R [363], was below 0.05.

3.6. Epistasis analyses

To examine the interaction between SNPs in *ADCY9* and *CETP* on *CETP* alternative splicing events, we performed a screening of interaction effects between SNP rs158477 in *CETP* and all biallelic SNPs with MAF>0.05 in the *ADCY9* locus. For the *ADCY9* locus, we kept position from 3,946,204 to 4,135,397 (Chromosome 16, GRCh38, 20 kb around *ADCY9* locus, Supplementary Figure 3.17). We did the same analysis with SNP rs1967309 in *ADCY9* and all biallelic SNPs with MAF>0.05 in the *CETP* locus (Supplementary text 6). SNPs with a Hardy-Weinberg equilibrium p-value below 0.01 were excluded. Interactions between SNPs were evaluated using the following model in R: lm(p $\sim$ rs1967309*rs158477+Covariates), where the covariates are the same as those used in the splicing QTL analysis without interaction.

3.7. Isoform quantification

Transcript quantifications were estimated using Transcript per million (TPM). The abundance values (TPM) for each *CETP* isoform were obtained from the GTEx V8 online server. From the remaining 699 individuals, samples with a sum of TPM greater than 0.5 for all three isoforms were selected for the quantification of isoform frequencies (Supplementary Figure 3.5). The transcript proportions were estimated by combining PSI values from both alternative splicing junctions, meaning AS1 and AS9 (Supplementary Figure 3.5). The proportions of each transcript were computed using the following formulas :

CETP-201 = 1-PSI$_{AS1}$ -PSI$_{AS9}$

CETP-202 = PSI$_{AS9}$

CETP-203 = PSI$_{AS1}$

3.8. LD block inference

To infer LD blocks for both *CETP* and *ADCY9* genes, we retained only SNPs in the GTEx dataset with a MAF>0.05 for 20 KB regions surrounding the genes. Using the `BigLD` function from the gpart R package [364], we inferred the LD block with a CLQcut of 0.3 and we generated plots for using the `LDblockHeatmap` function from the same package (Figure 3.1a, Supplementary Figure 3.17).

3.9. Logistic regression on CAD in GTEx

To evaluate the potential association between alternative splicing and cardiovascular diseases, we conducted a logistic regression analysis in the GTEx dataset. The subjects with CAD were identified based on the variable MHHRTATT (phv00169162.v8.p2) (140 cases) and MHHRTDIS (phv00169163.v8.p2) (126 cases), along with their respective cause of death (Supplementary file 1), using the variable DTHFUCOD (First Underlying Cause Of Death) from GTEx. Individuals were classified as control if they were not a case for both MHHRTATT and MHHRTDIS. In addition, we excluded individuals from the control group if they had the phenotype of "heart disease" (MHHRTDISB (phv00169164.v8.p2)), if their cause of death was associated to the heart but not linked to the two phenotypes mentioned, or if their cause of death was unknown (Supplementary file 2). This resulted in a total of 197 cases and 371 controls included in our analysis. To mitigate potential bias caused by PSI values of 0, we stratified the PSI values in four categories: samples with PSI values of 0 were grouped together, and the remaining samples were stratified into terciles, ensuring similar sample sizes in each group. The logistic regression was performed using this categorial variable instead of the individual PSI values, then compared the model with and without the PSI categories using anova() in R. Since PEER factors may correct for the effect of cardiovascular disease on the transcriptomic profile, we instead calculated 5 SVA [232] on the gene expression protecting the cardiovascular disease status. The other covariables included in the regression analysis were the first five PCs, age, sex, collection site (SMCENTER), the sequencing platform (SMGEBTCHT), total ischemic time (TRISCHD), and we also added *CETP* expression.

3.10. Mendelian Randomization

To assess the causal relationship between a change in the occurrence of alternative splicing events, we performed univariable and multivariable two-sample mendelian randomisation (MR) analyses.

3.10.1. Exposure Data

For the instrumental variables (IV) of *CETP* expression, we utilized the results from our gene-level *CETP* eQTL analysis. For the instruments for alternative exon 1 (AS1) and alternative splicing of exon 9 (AS9), we utilized the results from our sQTL analyses. All participants in the GTEx dataset were of European descent ancestry.

3.10.2. Outcome Data

We acquired summary statistics from online PheWAS results on European descent individuals, see supplementary tables 3.1 and 3.2 for a list of phenotypes and where they were obtained from. There should be no participant overlap between exposure and outcome database. Phenotypes were selected based on literature, observed in online PheWAS results accessed at December 2022 (https://atlas.ctglab.nl/, https://pheweb.sph.umich.edu/), or for other reasons mentioned in results.

3.10.3. SNP Exclusion Criteria

To determine the IVs for the three exposures, we applied several exclusion criteria. These criteria involved the following thresholds : a derived F-statistic below 10, a p-value above 0.001, and a MAF below 0.01. The derived F-statistic was estimated using an ANOVA model (aov() in R), comparing the model with covariates to the model with covariates and the SNP. We harmonized the exposure and outcome data using the `harmonise_data` function from the TwoSampleMR package in R [365], and subsequently removed any ambiguous and palindromic SNPs to ensure accurate and reliable results. To assess the correlation structure among the SNPs, we calculated the correlation matrix using the `ld_matrix` function from the R package ieugwasr [366], considering the 699 individuals of European descent in the GTEx dataset. We pruned high LD (r^2>0.9) pairs of variants using two strategies. In the univariable model, we removed the SNP with the smallest derived F-statistic. In the multivariable model, we avoided removing the SNP with the largest derived F-statistic for

each exposure and instead randomly removed one SNP from each pair of SNPs with a correlation exceeding 0.90. Following these filters, only the spleen and thyroid tissues had at least 3 SNPs remaining for each exposure, indicating that these tissues were suitable for further analysis. We performed our analysis in the thyroid tissue since it showed the most significant associations for all three exposures. Detailed summary statistics of the SNPs retained for the thyroid tissue analysis are provided in supplementary files 3 and 4.

3.10.4. Mendelian Randomisation Analysis

We conducted two sample MR analyses using two methods : the Inverse variance weighted (IVW) method and Mendelian randomization Egger regression (MR-Egger) method. There are three core assumptions in MR : i) The variant causes the exposure, ii) there are no confounders of the variant-outcome relationship, iii) the variant does not affect the outcome, except by its effect on the exposure. If the instrument satisfy those assumptions, the IVW methods provides an unbiased estimator of the causal effect. MR-Egger relaxes the conventional instrumental variable assumptions by allowing for directional pleiotropy and is used here as a sensitivity analysis. If the MR-Egger regression intercept coefficient is significantly different from 0 (p-value$<$0.05), it indicates the presence of directional pleiotropy. In univariable models, the correlation between *CETP* expression, AS1 and AS9 may induce bias due to violations of the 3^{rd} instrumental assumption. To address this, we employed MultiVariable Mendelian Randomization (MVMR) to investigate the causal effect of one exposure on the outcome conditional on the other two exposures. For example, we examined the causal relationship of AS9 by controlling for the effects of *CETP* expression and AS1. During the MVMR-Egger, we performed this test three times, using each exposure as the reference once to evaluate their intercept. If the intercept was significant in all three tests, it suggested the presence of confounding effects, and the results from MVMR-Egger were considered. If the intercept was not significant for all three tests, we relied on the results from MV-IVW. However, we compared the estimates from MV-IVW with those from MVMR-Egger to ensure consistency. Consistency between IVW and MR-Egger was assessed by examining if the beta coefficients were in the same direction and of comparable magnitude. All MR analyses were conducted using TwoSampleMR and MendelianRandomization [367] R packages. Bonferroni correction was applied to account for multiple testing with a significance threshold of p-value=0.003 (0.05/17 phenotypes).

4. Results

4.1. Tissue-Specific Genetic Regulation of *CETP* Isoforms Reveals Distinct Regulatory Patterns

CETP regulatory region has been previously described in the literature, with tissue-specific eQTLs located in the upstream region of the gene [47]. Its expression has been detected in many tissues, but is typically mostly found in adipose, breast, spleen and liver [359]. Since most of the studies that examine *CETP* expression only evaluate overall gene-level expression, hence ignoring isoform-level patterns, our goal is to study differences in genetic regulation in an isoform-specific way to determine their specificities. Using GTEx RNA-seq data, we confirmed that the most expressed isoform is the full-length transcript, *CETP-201*, whereas the alternative exon 9 transcript, *CETP-202*, and the isoform with an alternative starting exon 1, *CETP-203*, are less expressed in all tissues except in Lymphoblastoid Cell Line (LCL) (Supplementary text 6, Supplementary Figures 3.5 and 3.6). To study the genetic regulation of each of the *CETP* isoforms and compare them to gene-level *CETP* regulation, we identified *CETP* eQTLs, within a 20 kb region surrounding the locus (Methods). As previously reported in the literature [47], significant gene-level eQTLs are found upstream of the gene, in the LD block B-3 (Figure 3.1), and show heterogeneity across tissues, confirming a tissue-specific regulation (Supplementary text 6, Supplementary Figure 3.7). The impact of the statistically most significant eQTL, rs56156922 (Figure 3.1), strongly varies across tissues (Supplementary Figure 3.7, with the small intestine showing the strongest effect, potentially linked to its role in cholesterol absorption [47]. Notably, in the testis and ovary, rs56156922 exhibits opposite effects on *CETP* expression compared to other tissues (Supplementary text 6), suggesting a potential involvement of these organs in sex-specific traits associated with CETP activity, since eQTLs exhibiting opposite effects between tissues have been suggested to play a role in the development of complex trait [368].

To investigate isoform-specific regulation, we compared the estimated effects of significant eQTLs of each isoform to those of gene-level *CETP* (Methods). The genetic regulation of gene-level *CETP* is not significantly different from that of *CETP-201* and *CETP-202* (Supplementary Figure 3.8), with *CETP-201* showing significant association within the same LD block and displaying tissue-specificity (Supplementary text 6, Supplementary Figure 3.7).

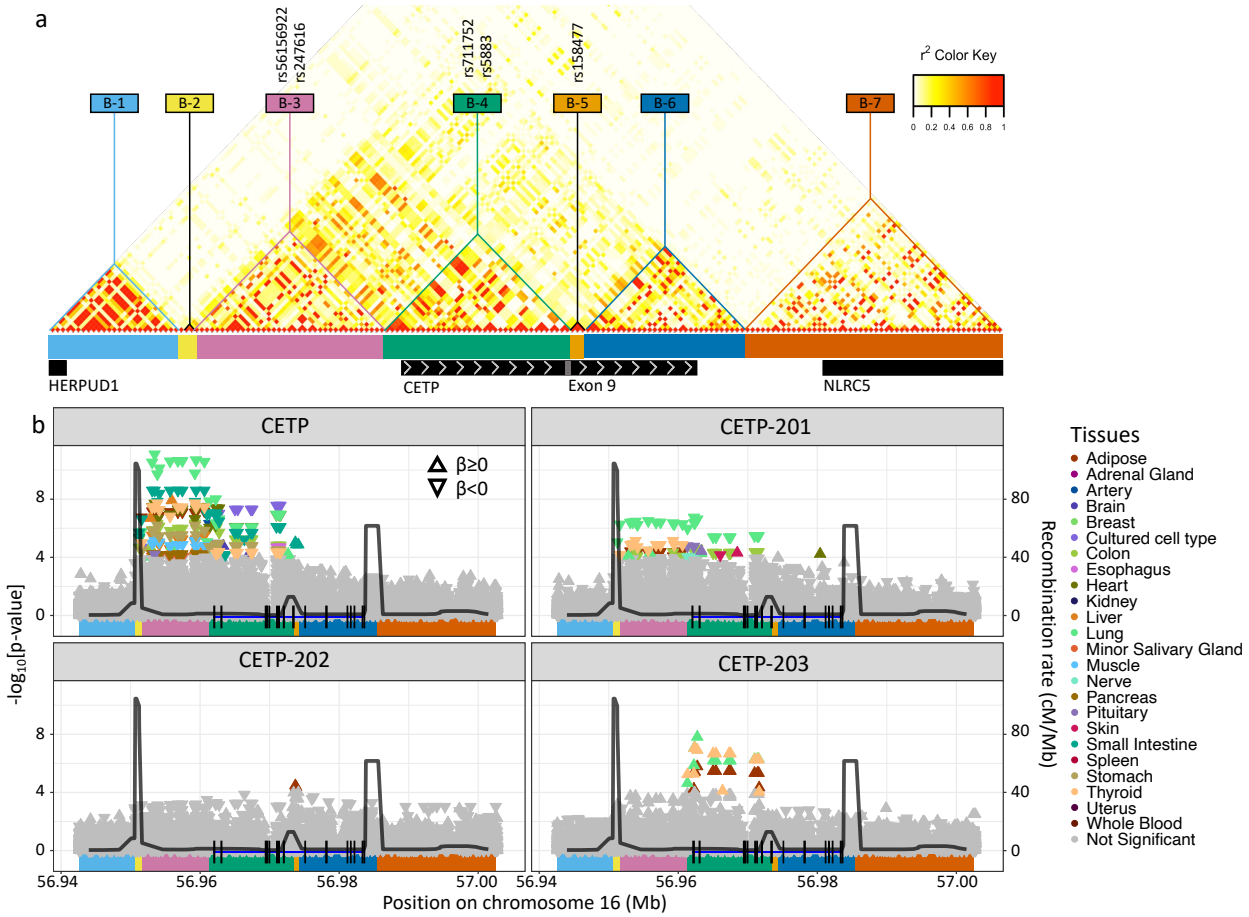

Fig. 3.1. Gene Structure and Genetic Regulation of *CETP* Transcript Expression. (a) Linkage Disequilibrium (LD) blocks within the *CETP* locus, with surrounding genes identified below and grey arrows indicating 5' to 3'. Blocks were estimated in GTEx participants from European descent using `BigLD` of gpart package with CLQ cut at 0.3 for SNPs having a MAF above 5%. IDs of SNPs of interest in this study are localized within their respective LD blocks. Position of the exon 9 within *CETP* is represented by a grey box. (b) cis-eQTLs of gene-level *CETP* expression (*CETP*) and its transcripts (*CETP-201, CETP-202, CETP-203*) for 49 tissues of GTEx, grouped by 24 tissue labels. LD blocks are represented under each plot using the color code from (a). SNPs below the threshold (p-value$<$0.05/(49*7)) are colored. The *CETP* exons are represented in each plot, specific to each isoform. Black lines represent the recombination rate in CEU of 1000G.

CETP-202 does not exhibit distinct significant eQTLs in the *CETP* locus, suggesting that both isoforms are regulated by the same genetic region. In contrast, distinct significant eQTLs are found for *CETP-203*, located in LD block B-4 (Figure 3.1), which showed significant differences in effect sizes compared to gene-level, *CETP-201* and *CETP-202* eQTLs (Supplementary Figure 3.8) in tissues where *CETP* is the most expressed, although there was no clear evidence of tissue-specific effects for *CETP-203* (Supplementary text 6). These

results indicate that *CETP-203* is differently regulated from the other two isoforms and points towards a novel putative regulatory region not captured by gene-level eQTL analysis.

4.2. Genetic Regulation of *CETP* Isoforms Revealed through Alternative Splicing Analysis

Analyzing *CETP* eQTLs at the isoform level did not result in distinguishing the specific regulatory mechanisms of the main isoforms, *CETP-201* and *CETP-202*, hindering their comprehensive characterization. However, an approach based on alternative splicing (AS) can provide a better understanding of their distinct genetic regulations. *CETP* isoforms *CETP-202* and *CETP-203* arise from distinct splicing events, specifically, the alternative splicing of exon 9 (AS9) (Figure 3.2a, Junction 2) and alternative exon 1 (AS1) (Figure 3.2a, Junction 1), respectively. To further assess isoform-specific genetic regulation, we first computed the Proportion-Spliced-In (PSI) values for each AS event using ASpli [231] in GTEx data, and we identified splicing quantification trait loci (sQTLs) within a 20 KB region surrounding the *CETP* locus (Methods). Notably, SNPs within the LD block B-4, which encompassed significant *CETP-203* eQTLs (Figure 3.1b) exhibit strong associations with AS1 (Figure 3.2b), primarily in visceral adipocytes, breast, spleen and thyroid tissues, but no tissue-specific effects were detected, since we did not detect overall heterogeneity of its estimate across all tissues (Supplementary text 6). The overlap between these sQTLs and *CETP-203* eQTLs indicates that alternative exon 1, which capture a change in the proportion of alternative exon 1, quantification effectively captures *CETP-203* expression levels. Considering that the expression of *CETP-202* may also be captured through the measurement of the alternative splicing event AS9, we detected sQTL for AS9 in multiple tissues (Figure 3.2b), including adipose tissues, pituitary gland, spleen and thyroid. As for AS1, no tissue-specific effects are observed (Supplementary text 6). The signals are primarily located at the end of the LD block B-4 (Figure 3.2b), with the previously reported rs5883 SNP displaying the most significant association [357, 358, 359], which is however not identified as an eQTL of *CETP*.

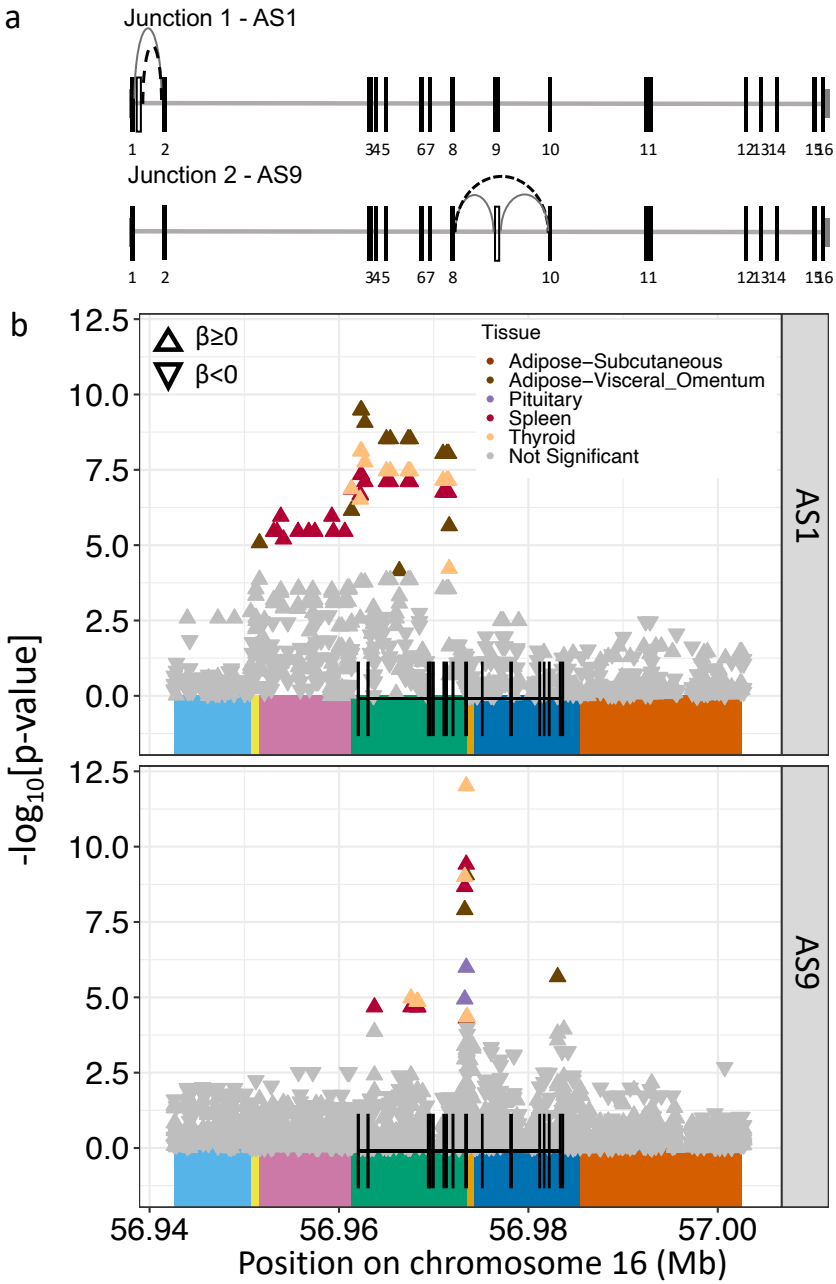

Fig. 3.2. Genetic control of alternative splicing at the *CETP* locus.
(a) Diagram of the two splicing junctions in *CETP* gene, where empty boxes represent spliced exons, and the dashed black lines represent the splicing junction variation characterized in ASpli analyses. PSI values represent the proportion of reads associated with the dashed black lines to the total number of reads for both the dashed and full lines. (b) sQTL of the alternative exon 1 (AS1) and alternative splicing of exon 9 (AS9) in the gene *CETP* in different tissues from GTEx data. LD blocks from Figure 1a are represented below the graph. Only tissues with SNPs passing the significance threshold (p-value<0.05/[49 tissues x 7 LD blocks]) are colored. *CETP* gene and its 16 exons are represented at the bottom of each plot.

4.3. Causal Relationships and Tissue-Specific Effects of *CETP* Isoforms on Cardiovascular Disease Phenotypes through Alternative Splicing Analysis

The CETP protein has been extensively implicated in cardiovascular disease (CVD) [53, 343, 344, 345], probably due to its impact on the concentration of cholesterol in LDL particles (LDL-C). To investigate the causal relationship between each isoform and CVD-associated phenotypes, we performed Mendelian Randomisation (MR) analyses utilizing eQTLs of gene-level *CETP* expression and sQTLs of AS1 and AS9 as instruments, and coronary artery disease (CAD), as well as HDL-C, LDL-C and triglyceride (TG) levels as outcomes) (Methods). We employed multivariable models (Methods, Supplementary Figure 3.9) allowing for isoform-specific contributions while controlling for gene-level expression. When the causal effects of both AS events on a phenotype are consistently in the same direction, *CETP-201* expression variation is likely the influential factor for the phenotype, since it lacks AS1 and AS9. However, if the causal effects of AS1 and AS9 are reversed, either *CETP-202* or *CETP-203* variation is considered the most impactful, as they display only one of the splicing events.

We used the significant eQTLs and sQTLs found in the thyroid tissue, an organ that displays high expression of all three isoforms (Supplementary Figure 3.5) and strong associations for all three exposures: gene-level *CETP* expression, AS1 and AS9 variations (Figures 3.1 and 3.2). *CETP* expression causally modulates all outcomes, as expected [36, 44, 53, 343, 344, 345, 351]. We also found that an increase in the proportion of both AS1 and AS9 increase the HDL-C concentration, as well as a decrease in LDL-C, TG, and CAD occurrence (Figure 3.3a), robust to the adjustment for gene-level *CETP* expression. The results remained significant in a sensitivity analysis accounting for directional pleiotropy using the MR-Egger models (Supplementary text 6, Supplementary Figure 3.11) [122]. Since the effects of both AS events align, our findings indicate that variations in *CETP-201* proportions impacts these phenotypes. Specifically, an increase in *CETP-201* leads to increased LDL-C, TG and CAD, but decreased HDL-C, which is consistent with existing literature, as it is the only isoform known to be secreted in plasma [40].

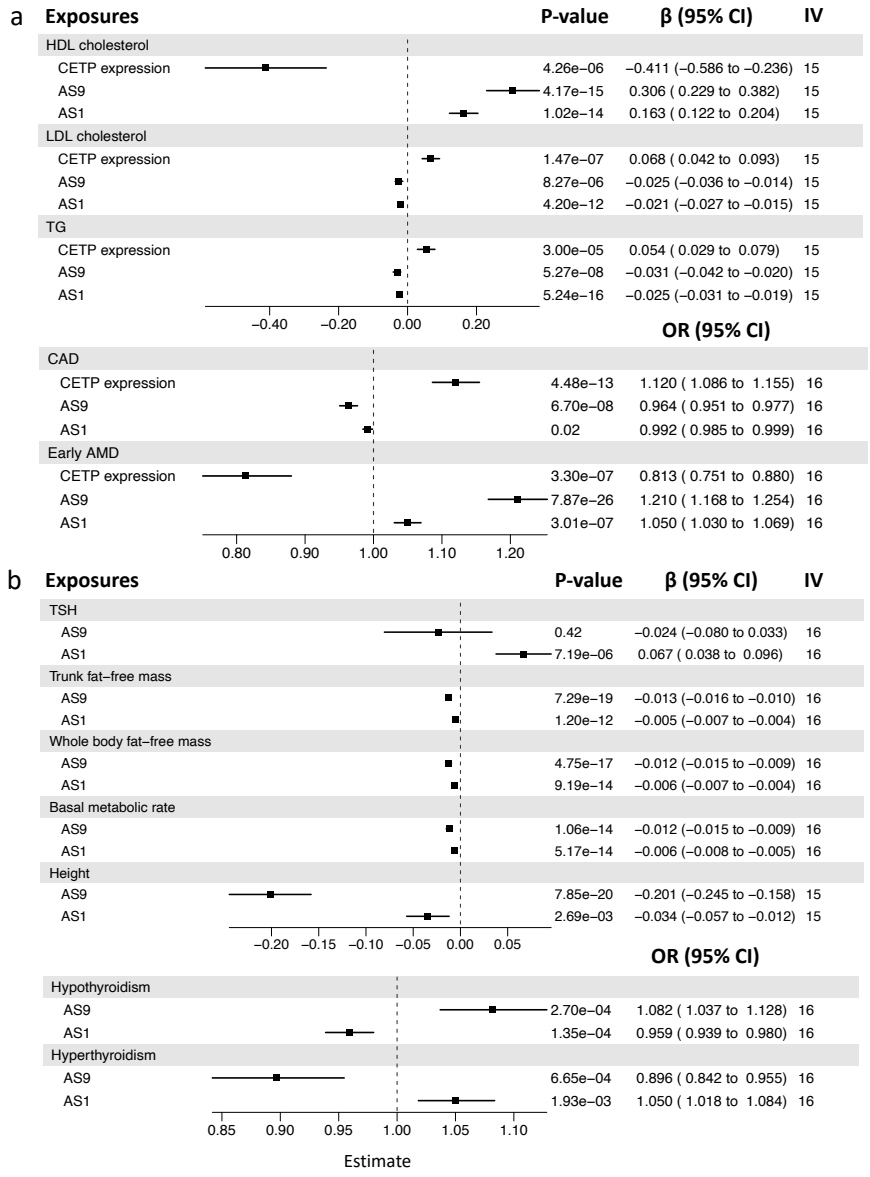

Fig. 3.3. Multivariable Mendelian Randomisation on *CETP* expression and alternative splicing events as exposures and CETP-relevant traits as outcomes.
Effects of change in the proportion of *CETP* isoforms using multivariable Mendelian Randomisation (MR) on phenotypes (a) previously associated with gene-level *CETP* expression and (b) associated with thyroid/pituitary glands. Results are from the IVW test, which takes into account gene-level *CETP* expression, alternative exon 1 (AS1) and alternative splicing of exon 9 (AS9). Estimates (β) represent the effect of a change of 1 standard deviation on the outcomes. IV represents the number of instrumental variables (IV) used in the analysis. Results for univariable MR are show in Supplementary figures 3.10 and 3.13 and for MR-Egger in Supplementary figures 3.11 and 3.14.

Given that the transcriptome profile can differ across cell types, potentially affecting isoform dynamics, we sought to investigate whether variation in *CETP* isoform proportion among samples is differentially associated with CAD events across the various tissues available in GTEx (Methods). We report significant associations between AS9 proportion and CAD in adipocyte tissues (P-value$_{Subcutaneous}$=0.01, N=180; P-value$_{Visceral}$=0.04, N=234) and in thyroid (P-value=0.005, N=277) (Supplementary figure 3.12). The MR results above are consistent with the observed trend in subcutaneous adipocyte tissue, wherein individuals with higher levels of AS9 have a lower proportion of CAD events. Additionally, it is worth reporting that the pattern in liver, which is the main contributor to plasmatic CETP, shows the same trend, although the association is not statistically significant (P-value=0.29) potentially due to a smaller sample size (N=56). This trend, seen in two tissues that contribute significantly to CETP secretion in plasma [40], could be explained by the fact that an increase of CETP-202 (increase of AS9) decreases the secretion of CETP-201 in the plasma [40, 42, 66, 67, 352]. However, we identified an effect in the opposite direction for visceral adipocytes, another tissue involved in plasma CETP secretion, and for the thyroid, where *CETP* expression is the second highest (Supplementary Figure 3.5) but lacks a known internal function associated with CETP. Altogether, these findings indicate tissue-specific effects of the main isoform on CAD traits, highlighting the complex interplay between isoform-specific regulation and phenotypic outcomes.

4.4. Change in isoform proportion shows causal relationships with phenotypes distinct to *CETP* expression

In addition to CAD, other phenotypes have previously been associated with CETP protein levels or genetic variants within the *CETP* locus in the literature and online databases. The list of these phenotypes is reported in Supplementary Tables 3.1 and 3.2. For each phenotype, we performed MR to identify the isoform impacting a given phenotype.

CETP expression has been found to be associated with early AMD [59, 60], where an increase of *CETP* expression is associated with a decrease of the prevalence, which is in the opposite direction to the association with CAD. This opposite effect could thus be due to isoform specific effects. Multivariable MR showed a positive correlation between early

AMD and both AS events, robust to controlling for *CETP* expression and pleiotropy (Supplementary text 6, Supplementary Figures 3.10 and 3.11). We observed that variation in AS9 is more significantly associated with early AMD (p-value=7.87 x 10^{-26}) than a change in *CETP* expression levels (p-value=3.30 x 10^{-7}), and with similar effect sizes in opposite direction, suggesting that variation in the isoform proportion may also be important for this disease.

While the modulation of plasmatic CETP activity by thyroid hormones and thyroid dysfunction is known [369, 370], the specific role of CETP inside this tissue, where it is highly expressed, remains unknown. We thus used thyroid dysfunction, specifically results from GWAS on hyperthyroidism and hypothyroidism phenotypes (Supplementary Tables 3.1 and 3.2), to evaluate the effect of a change in AS on thyroid function. In multivariable MR models, changes in both AS events were significantly associated with hypothyroidism (p-value$_{AS9}$=2.70 x 10^{-4}; p-value$_{AS1}$=1.35 x 10^{-4}) and hyperthyroidism (p-value$_{AS9}$=6.65 x 10^{-4}; p-value$_{AS1}$=1.93 x 10^{-3}) (Figure 3.3b), but not with *CETP* expression (Supplementary Figure 3.13). The results for hypothyroidism remained significant after controlling for pleiotropy (Supplementary text 6, Supplementary Figure 3.14). Thyroid function is also associated with the thyroid stimulating hormone (TSH), produced by the pituitary gland. We evaluated the causal relationship of AS events on TSH production and found an increase of AS1 to be associated with an increase of TSH levels (p-value=7.19 x 10^{-6}) (Figure 3.3b), robust to controlling for *CETP* expression and pleiotropy (Supplementary Figures 3.13 and 3.14). This result suggests that variation in AS1 may impact the production of TSH. Pituitary and thyroid hormones are known to modulate fat-free mass, basal metabolic rate and height [371, 372, 373] which are traits that have all been associated with genetic variants in *CETP* in online GWASatlas [374]. Based on publicly available GWAS statistics for these phenotypes, we found significant causal relationships only with AS events (Figure 3.3b) but not with *CETP* expression (Supplementary figures 3.13 and 3.14). A change in the proportion of *CETP* isoforms thus have significant effects on anthropometric traits independently of *CETP* expression, potentially through the regulation of pituitary or thyroid hormones. Altogether, these findings indicate that *CETP* isoforms *CETP-202* and/or *CETP-203* may significantly impact pituitary and/or thyroid glands functions.

4.5. Variation in alternative splicing of exon 9 impacts pulmonary and pregnancy phenotypes

CETP has been involved in one of the first sex-specific genetic interactions reported to date in humans, whereby different combinations of genotypes at rs1967309 in *ADCY9* and at rs158477 in *CETP* influenced *CETP* expression in multiple tissues [360]. The SNP rs158477, located 188 bp away from the end of exon 9, may impact splicing of this exon, as adjacent regions to spliced exons are known to be involved in splicing regulation [190]. To assess whether the epistatic interaction could impact alternative splicing of exon 9, we examined the evidence for interaction between the SNP rs158477 in *CETP* and polymorphisms in *ADCY9* on AS9 variation across tissues in GTEx (Methods).

We found significant interaction effects between rs158477 in *CETP* and several SNPs in *ADCY9* on AS9 (Figure 3.4a), with the *ADCY9* SNP rs4786452 showing the strongest association (p-value$_{Interaction}$=5.78 x 10^{-7}, breast mammary tissue). SNP rs4786452 is in strong LD with rs1967309 (D'>0.99 in the CEU population from 1000 Genomes project [252]) but has a lower minor allele frequency (MAF=0.15 in GTEx, Supplementary text 6). We note that, because of low MAF at this SNP, one combination of genotypes was missing (CC-rs4786452/GG-rs158477) which can lead to false positive results in interaction analyses. However, the interaction remained significant when merging CC with CG of rs4786452 (p-value$_{Interaction}$=1.70 x 10^{-7}) (Figure 3.4b), which indicate that the presence of a C allele at rs4786452 interacts with genotypes at rs158477 to modulate AS9 variation. These results support the hypothesis that the *ADCY9* locus regulates *CETP* expression and suggest that it may be involved in alternative splicing of exon 9 through an interaction with rs158477.

The genetic interaction between *CETP* and *ADCY9* was shown to be under selection in the Peruvian population, but the selective pressure remains unknown. However, two pulmonary phenotypes, known to be altered in high-altitude populations [375, 376], were found to be significantly impacted by the rs1967309 x rs158477 interaction using the UK biobank data, driven by females [360], namely forced expiratory volume in 1-second (FEV1) and forced vital capacity (FVC). FVC is a measure of lung capacity and indicates restrictive lung disorders, while FEV1 is more commonly used as a measure for obstructive airway disorders. To investigate whether alternative splicing of *CETP* may causally impact these

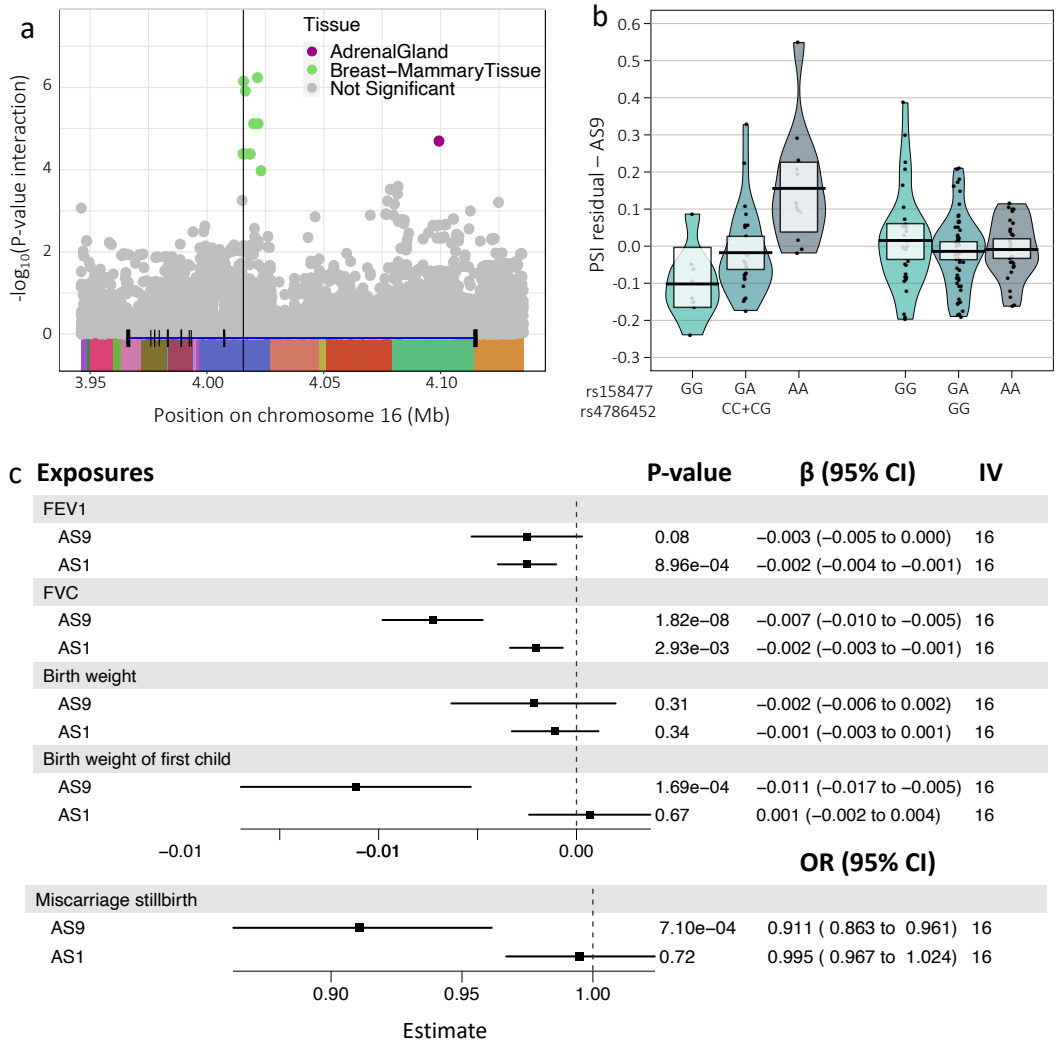

Fig. 3.4. Role of alternative splicing in the epistatic interaction between *ADCY9* and *CETP*.

(a) Epistatic interaction between all SNPs in *ADCY9* with MAF>0.05 and the SNP rs158477 in *CETP* on Proportion-Spliced-In (PSI) values for alternative exon 1 (AS1) and alternative splicing of exon 9 (AS9) across tissues in GTEx. Only tissues with SNPs passing significance threshold (p-value<0.05/49) are colored. *ADCY9* and its exons are represented at the bottom and colored boxes represent LD blocks estimated with `BigLD` of gpart package, with a CLQ cut-off at 0.3 for SNPs with MAF>0.05 in the European descent population of GTEx. The vertical black line represents the position of SNP rs1967309. (b) PSI values of AS9 (corrected for covariates used in the regression model) by genotype combination for rs4786452 (ADCY9) and rs158477 (CETP) in breast mammary tissue. Genotypes CC and CG for rs4786452 were combined for robustness, due to the absence of the CC-rs4786452/GG-rs158477 combination. (c) Multivariable Mendelian Randomization (MR) for AS1 and AS9 exposures on outcomes potentially linked to the previously reported selective pressure in Peruvians. Estimates (Beta) represent the effect of a change of 1 standard deviation on the outcomes. IV represents the number of instrumental variables (IV) used in the analysis. Results for *CETP* expression and univariable MR are show in Supplementary figure 3.13 and for MR-Egger in Supplementary figure 3.14.

phenotypes, we performed MR analysis on these outcomes (Methods, Supplementary Tables 3.1 and 3.2). Our analysis revealed an association between AS9 and FVC (p-value$_{AS9}$=1.82 x 10^{-8}), but not with FEV1 (p-value$_{AS9}$=0.08) (Figure 3.4b, Supplementary figures 3.13 and 3.14).

Another hypothesis put forward to explain the strong selective pattern observed in Peruvians was the impact of *CETP* modulation in pregnancy or in early life [360]. During the early stages of pregnancy, maternal plasmatic CETP activity and *CETP* expression in the placenta are increased [377, 378, 379], and can modulate newborn's body weight at birth [380]. We thus evaluated the impact of AS variation on weight (first child's weight at birth and the individual's weight at birth), as well as on stillbirth and miscarriage, through MR (Methods, Supplementary Tables 3.1 and 3.2). The weight of the first child was negatively associated with AS9 (p-value=1.69 x 10^{-4}) (Figure 3.4, Supplementary figures 3.13 and 3.14), but not birth weight of the individual (p-value$_{AS9}$=0.31). Furthermore, an increase in the AS9 was associated with a decreased probability of pregnancy complications (miscarriage, stillbirth) (p-value=7.10 x 10^{-4}). These results suggest that variation in *CETP* isoform proportions in women is causally associated with pulmonary capacity and pregnancy complications, providing further evidence into the source of the selective pressure acting on *ADCY9* and *CETP*'s combinations of genotypes, potentially modulated through the coregulation of AS9.

5. Discussion

In this study, we comprehensively examined the regulation of *CETP* isoforms and identified potential novel functions of CETP. We demonstrated the utility of alternative splicing approaches in estimating isoform levels : although alternative splicing methods typically require high coverage of splicing junctions, resulting in fewer suitable samples for analysis, they offer greater precision in quantification. Furthermore, as each splicing junction only occur in one isoform in *CETP*, combining splicing measurements helped to identify which isoform contributes the most to the effect on a phenotype and allowed us to reveal new associations.

Using multivariable mendelian randomisation, we established significant causal relationships between alternative splicing events of *CETP* and phenotypes associated with *CETP* expression, such as lipid profile, CAD and early AMD. Notably, *CETP-201*, which lacks

both AS1 and AS9 emerged as the most important isoform for these phenotypes, in line with its presence in plasma and role in lipid metabolism. These findings support the notion that studies assessing plasmatic CETP or *CETP* expression using gene-level *CETP* are effectively assessing CETP-201 functions. Furthermore, we showed that changes in the proportion of *CETP* isoforms affected CAD differently across tissues, suggesting tissue-specific and isoform-specific mechanisms of CETP action. Specifically, data from liver and subcutaneous adipose tissue, which are two major producers of plasmatic CETP [40], showed an increase in AS9 that is associated with a decrease in CAD risk, consistent with our MR results. In contrast, the visceral adipose tissue, another major CETP producer, and the thyroid, which has the second highest *CETP* expression (despite not being known as a major contributor of plasma CETP possibly due at least in part to its much smaller size) showed an association in the opposite direction. It raises the question of whether CETP inhibitors, which can enter cells due to their lipophilic nature [348, 349], will differentially target each protein isoform. Whether the tissue-specific and isoform-specific effects uncovered here could modulate their potential protective effect against CVD requires further investigation.

AMD is a multifactorial disease caused by damage to the macula. While it involves the accumulation of extracellular lipids, among other factors, the exact pathophysiological changes leading to this accumulation remain to be elucidated [60, 381]. Cholesterol accumulates in the Subretinal pigmented Epithelial (RPE), which participates in cholesterol exchange with its neighboring tissue, the neural retina, where *CETP* is expressed [382]. Therefore, the process of reverse cholesterol can be crucial to prevent cholesterol accumulation and inflammation. In contrast to the causal relationships observed on CAD, the impact of change in the variation of AS9 on early AMD was similar in effect size to that of *CETP* expression, suggesting significance of isoform proportions in disease development, possibly through lipid homeostasis within the cell [42, 67], but future studies should investigate CETP's isoforms in eye tissues. Likewise, our results on *CETP* isoforms in thyroid and pituitary glands, which are not known to produce plasmatic CETP, suggests an intracellular function for *CETP* in these tissues, potentially with distinct roles played by different isoforms. Specifically, we found that changes in alternative splicing were causally associated with pituitary and thyroid gland function, as well as many phenotypes influenced by the hormones they secrete.

Previous findings identified a sex-specific epistatic interaction between *ADCY9* and *CETP* genes, specifically with SNP rs158477 located near exon 9 [360]. Here, we report an epistatic interaction between *ADCY9* locus and rs158477 on the alternative splicing of exon 9, suggesting that the interaction between both genes may act to modulate *CETP* isoforms proportions. Interestingly, *ADCY9* can activate protein kinase A (PKA), which was found to regulate alternative splicing of other genes [195, 383]. Furthermore, alternative splicing is known to contribute to adaptive evolutionary changes as it is one mechanism leading to fast response to environmental changes [354]. The observed epistatic interaction on AS9 could thus contribute to the coevolution event reported between the two genes. Changes in the proportion of *CETP* isoforms are associated with pregnancy-related phenotypes and pulmonary capacity, both of which are influenced by living at high altitudes [384, 385, 386, 387, 388], in line with the selective pressure being observed in the Peruvian population.

There are several limitations in our study. Our analyses investigating the impact of AS changes had limited statistical power given the available number of samples in GTEx, and were restricted to tissues with the highest *CETP* expression, nevertheless generating interesting hypotheses to follow up on. Indeed, replication studies are needed, particularly for the association between change in *CETP* isoforms and CAD occurrence. Additionally, while our analyses focused on identified cis-QTLs, it is important to consider the potential importance of distal regulation, which warrants analyses in larger datasets. This is especially true in the context of the previously identified epistatic interactions involving *CETP*. It is also noteworthy that, in our MR analyses, the three exposures, i.e. *CETP* expression, alternative splicing of exon 9 and alternative exon 1, are not completely independent, which may lead to overcorrection of the effect in the multivariable analysis. To gain a better understanding of the relationship between the three exposures and the phenotypes, it could be beneficial to perform phenotype analysis considering the interaction between the exposures. However, datasets including information of subject with both phenotype and expression data generally have limited statistical power for such analyses, given the limited sample sizes. Furthermore, another limitation is that we were unable to perform sex-specific sQTL analyses due to the limited statistical power resulting from the available number of samples in GTEx. However, the pregnancy-related results from our MR analyses suggest there may be interesting CETP

functions specific to women phenotypes. Therefore, conducting follow-up analyses of sex-specific sQTL in larger cohorts is necessary to further explore sex differences in *CETP* regulation.

In conclusion, our study clearly shows that each *CETP* transcript has its own specific genetic regulation, either directly involving alternative splicing or in their expression modulation through genetic loci. Using these distinct features, we observed that a change in the proportion of its isoforms may have an important impact on phenotypes already known to be associated with *CETP* expression and activity, but also revealed new associations, which suggests a potentially important effect of *CETP* isoforms in the pituitary and/or thyroid glands. Furthermore, we propose that epistatic interactions involving *CETP* could be mechanistically linked to alternative splicing, which could in turn have importance repercussions on human cardiovascular, pulmonary and reproductive health.

6. Supplementary text

6.1. Truncated isoforms in LCL

6.1.1. Sashimi plot

Sashimi plots were utilized to visualize splice junctions of the *CETP* gene using the IGV software [389]. For the GTEx dataset, all the samples were then merged by tissue, while the European subset were merged in the GEUVADIS dataset, using samtools to merge the samples for plotting in IGV. For the visualization in IGV, we applied a coverage filter of 1000X for a junction in LCL and 2000X for the thyroid gland. This filter allowed us to focus on the junctions associated with the three isoforms of *CETP* (Supplementary Figure 3.6a).

6.1.2. Truncated isoforms

As observed for *CETP* at the gene level, all three isoforms also seem to be expressed across tissues associated to lipid and macrophage with *CETP-201* being the predominant isoform (range of 25.9% to 100%), followed by *CETP-202* (range of 0.0% to 74.1%), then *CETP-203* (range of 0.0% to 35.7%) (Supplementary Figure 3.5b). Interestingly, cells-EBV-transformed lymphocytes (or Lymphoblastoid Cell Line, LCL) exhibited minimal expression of *CETP-201*. In particular, the cell line showed almost no coverage of exon 1 to exon 6, but it was still possible to detect alternative splicing of exon 9 (Supplementary Figure

3.6a), representing two unreported *CETP* transcripts. We confirmed that these transcripts were also observed in the LCL from the GEUVADIS dataset [253]. This finding may be attributed to a mechanism specific to the transformation process by EBV [390, 391]. To gain further insights, we investigated the transcription factors associated to EBV infection, obtained from [391] and visualized using the http://epigenomegateway.wustl.edu/ browser (Supplementary Figure 3.6b). We observed two of its transcription factors, *EBNA2* and *EBNLP*, had enhancer sites in intron 2 and between exons 7 and 10 (Supplementary Figure 3.6b). During an EBV infection, pathways involved in fatty acid synthesis are induced in newly infected B-cells [392], potentially affecting *CETP* expression, since its activity is increased during infection [338]. While many functional domains are retained, the protein structure of these unknown *CETP* transcripts undergoes significant changes, rendering their functions unknown. Further research is needed to investigate these new transcripts, and whether they are found in other tissues, as these isoforms, if present in other tissues, would be masked by the presence of *CETP-201* and *CETP-202*.

6.2. Supplementary results on eQTL and sQTL analyses

6.2.1. Residual LD between LD blocks

As we observed significant associations in two LD blocks for gene-level *CETP* and for *CETP-201*, we aimed to determine whether the signals in the less significant LD blocks (B-4) were generated by a residual effect of LD or by a second regulatory region. To investigate this, we included the most significant SNP of B-3 as a covariate in the linear regression analysis of SNPs in B-4. The significance of SNPs within B-4 disappeared, confirming that it is indeed caused by residual LD rather than a second independent regulatory region.

6.2.2. Tissue-specific regulation of *CETP* expression

eQTLs exhibiting opposite effects between tissues have been suggested to play a role in the development of complex traits [368]. As *CETP* expression has been reported to be tissue-specific [47], confirmed by our results (Supp Figure 1), we sought to assess tissue-specificity in the regulation of *CETP* expression. A single SNP for *CETP-202* barely passed our eQTL threshold in only one tissue (Figure 3.1b), representing little evidence for *CETP-202*-specific eQTLs, thus we did not perform tissue-specific analyses for this isoform. We identified the strongest eQTLs for gene-level *CETP*, *CETP-201* and *CETP-203*: rs56156922 for gene-level

CETP, rs247616 for *CETP-201* (within LD blocks B-3) and rs711752 for *CETP-203* (within LD blocks B-4). The mean and the standard deviation of the effect size for each eQTL was estimated for each tissue (Supplementary Figure 3.7). We evaluated the difference of effect sizes for these SNPs for each pair of tissues with an eQTL p-value under 0.0001 in at least one tissue, and then counted the number of tissues for which the SNP showed significant difference (T-test, p-value$<$0.001 to control for the number of tissue comparisons). For each SNP, we also looked at the overall heterogeneity across all tissues with a Q-cochran test from the package meta [363]. We confirmed that eQTL effect sizes differ between tissues (gene-level CETP: p-value$_{Q-Cochran}<10^{-10}$; *CETP-201*: p-value$_{Q-Cochran}<10^{-4}$; *CETP-203*: p-value$_{Q-Cochran}<10^{-7}$). *CETP-203* exhibited significant differences in its eQTL effects across tissues at rs711752, however, the effects were consistently in the same direction, suggesting that they may simply indicate stronger statistical evidence in tissues where this isoform is most highly expressed. We describe below the tissue-specific findings at the gene-level and for *CETP-201*.

The effects of rs56156922 on gene-level *CETP* expression in the small intestine is significantly more negative compared to most of the other tissues (Supplementary Figure 3.7). This observation could be associated with the modulation of *CETP* expression by a fat-rich diet, as the small intestine is responsible for absorbing cholesterol [47]. Although not reaching significance as an eQTL in the testis and ovary, we noticed that rs56156922 exhibited opposite directions of effects on *CETP* expression in these sex-specific tissues compared to other tissues, and these differences were statistically significant (significantly different estimates for 27 and 28 tissue comparisons, respectively). These organs are involved in the production of sexual hormones, and have been involved in differences between sexes for various phenotypes associated with CETP activity, such as HDL-cholesterol levels and cardiovascular disease [393, 394, 395, 396]. This suggests that the differential regulation of *CETP* in these tissues may play a role in the sex-specific characteristics of these complex traits.

The effects of rs247616 on *CETP-201* expression are significantly different, and in opposite directions, for brain amygdala and whole blood compared to most other tissues (Supplementary Figure 3.7). Interestingly, the pattern in amygdala was not detected when looking at gene-level *CETP* expression eQTLs. The amygdala has been associated with Alzheimer's disease [397], and some *CETP* polymorphisms have been associated with this disease, possibly

modifying brain structure and neurodegenerative disease susceptibility [294, 398]. Moreover, amygdala activity has been associated with bone marrow activity, arterial inflammation, and with risk of cardiovascular disease events [399] and atherosclerotic risk [400]. These findings suggest a potential association between the *CETP-201* isoform and the development of diseases affecting the amygdala. However, further studies are needed to explore this relationship in more detail.

6.2.3. Tissue-specificity in alternative splicing

To assess tissue-specificity regulation of alternative splicing, we identified the strongest splicing quantitative trait loci (sQTL) for alternative exon 1 (AS1) and alternative splicing of exon 9 (AS9).

Similar to tissue-specific analyses of eQTLs above, we evaluated the difference of effect sizes for the top sQTL for each splicing event (rs711752 for AS1, rs5883 for AS9) for each pair of tissues with an sQTL p-value under 0.0001 in at least one tissue, and then counted the number of tissues for which the SNP showed significant difference (T-test, p-value$<$0.001). For each SNP, we also looked at the overall heterogeneity across all tissues with a Q-cochran test from the package meta [363], considered as significant if the p-value were under 0.05. For AS1, the strongest sQTL, rs711752, is also the strongest eQTL for *CETP-203*. This SNP has previously been linked to metabolic syndrome and dyslipidemia [401, 402]. We did not detect tissue-specific effects for sQTLs in tissues expressing high level of *CETP-203* (Breast-mammary tissue, Spleen, Thyroid, Visceral adipocyte, p-value$_{Q-Cochran}$=0.23). Likewise, no tissue-specific effect was observed for rs5883 in AS9 (p-value$_{Q-Cochran}$=0.07). This indicates that the genetic regulation of alternative splicing of the exon 9 is likely to be preserved across different tissues.

6.3. Mendelian Randomization Analyses

To better understand the impact of *CETP* expression at the gene level and the role of isoforms on phenotypes, we employed univariable and multivariable models (Methods, Supplementary Figure 3.9) using three exposures, namely gene-level *CETP* expression, AS1 and AS9. Univariable models represent the conventional approach of assessing causal relationship between the three exposures and phenotypes independently, while the multivariable

model allows to consider the contributions of specific isoforms while controlling for gene-level expression.

Results of multivariable (MV) models are described in the main text. Here, we describe additional results on the associations with gene-level *CETP* expression, as well as comparison of MV models with univariable models and (MV)MR-Egger models for AS1 and AS9 exposures.

6.3.1. Lipid profile

Univariable and multivariable IVW models replicated the well-known association between an increase of *CETP* expression with decrease of HDL-c levels, as well as an increase of LDL-c and TG (Supplementary figure 3.10). AS1 and AS9 are also associated with these outcomes, but with weaker effects and in the opposite direction of *CETP* expression. We performed MR-Egger analysis (Supplementary figure 3.11) to evaluate the impact of pleiotropy on our results. The intercept indicated significant pleiotropy in the univariable models between AS and HDL-c level. However, after accounting for *CETP* expression in the MV model, the intercept was no longer significant, suggesting that *CETP* expression influences the causal relationship between AS and HDL-c. Furthermore, the estimates between MR-Egger and IVW were in the same direction.

6.3.2. Diseases known to be associated with *CETP* expression: CAD and early AMD.

Causal relationships between *CETP* expression and CAD or early AMD were successfully replicated in univariable and MV analyses (Supplementary figures 3.10 and 3.11). In the univariable models, we observed a nominal significant correlation between increased AS1 and increased risk of early AMD in the IVW test (p-value=0.04) and significant association between increased AS9 and increased risk of early AMD (p-value=7.26 x 10^{-5}). Those association replicated in the multivariable model (p-value$_{AS1}$=3.01 x 10^{-7}, p-value$_{AS9}$=7.87 x 10^{-26}), which included additional instrumental variables, thereby increasing statistical power. Similar to the findings for lipid profiles, the associations between AS1 and AS9 with early AMD were in the opposite direction compared to *CETP* expression.

The intercepts of the MR-Egger analysis were significant for these phenotypes (Supplementary figure 3.11), and the effect estimates were consistent with those obtained from the IVW method.

6.3.3. Pituitary and thyroid

CETP was found to be highly expressed in pituitary and thyroid glands. We investigated the impact of changes in *CETP* expression on diseases related to the thyroid (hypo/hyperthyroidism) and the hormone TSH, which is produced by the adrenal gland and can affect thyroid function. Our analysis did not reveal any significant causal relationship between *CETP* expression and these conditions in any of the models (Figure 3.3, Supplementary figures 3.13 and 3.14). Contrary to multivariable models, univariable models also did not show causal relationship between AS1 or AS9 and thyroid phenotypes after Bonferroni correction, except for a slight association with AS1 for TSH. The intercepts of MR-Egger were not significant.

6.3.4. Anthropometric traits

We observed a weakly significant causal relationship between *CETP* expression fat-free mass and basal metabolic rate (BMR) in the univariable models (Supplementary figures 3.13 and 3.14). However, these associations did not persist in multivariable models or MR-Egger analyses. Additionally, body height, which had genome-wide significant associations in the *CETP* locus, did not show any relationship with *CETP* expression. For alternative splicing exposures, in the univariable models, AS9 showed a negative association with fat-free mass, BMR and body height, which were stronger in the multivariable models, and persisted in MR-Egger (Supplementary figures 3.13 and 3.14). The estimates for AS9 were consistent across all models, indicating robust results. AS1 showed similar effects for fat-free mass and BMR, but not for body height.

6.3.5. Pulmonary phenotypes

CETP expression showed a strong association with forced expiratory volume in 1 second (FEV1) in both the univariable and multivariable models for IVW and MV-MR-Egger analyses (Supplementary figures 3.13 and 3.14). The univariable MR-Egger model was nominally significant, but the estimate was coherent and the intercept was not significant, indicating no significant horizontal pleiotropy. Associations with forced vital capacity (FVC) were less

pronounced and did not persist in the MR-Egger analysis, even in the absence of detected horizontal pleiotropy. This suggests that *CETP* expression, may play an important role in lung function, specifically in relation to obstructive lung diseases, which can be detected by FEV1.

On the other hand, AS9 did not pass the Bonferroni corrected threshold with FVC in the univariable IVW model, but became significant in the multivariable IVW and MR-Egger models (Supplementary figures 3.13 and 3.14), potentially due to increased statistical power. No horizontal pleiotropy was detected by MR-Egger. AS1 association did not pass the Bonferroni corrected threshold for FVC. However, it should be noted that there may also be an association with height, as taller individuals tend to have higher FVC [403].

6.3.6. Pregnancy-related phenotypes

On phenotypes related to pregnancy, we found that *CETP* expression is associated with the birth weight of the participant, but in the opposite direction compared to the birth weight of the first child. However, as birth weight is influenced by fetal sex, with male fetuses generally having higher birth weights [404], the opposite direction observed may be due to sex-specific effect. Unfortunately, the available summary data did not allow to stratify by the sex of the participant for birth weight and the sex of the fetuses for the birth weight of the first child.

Alternative splicing (AS1 and AS9) did not show significant associations with the birth weight of the participant: the univariable model showed a suggestive association between AS1 and birth weight, which was lost in the multivariable model. AS9 was not significant in the univariable models (IVW and MR-Egger), but was significant in the multivariable model for birth weight of first child. Once again, the influence of fetal sex could be relevant, but this hypothesis could not be tested with the data at hands.

Regarding the stillbirth/miscarriage phenotypes, neither *CETP* expression nor AS1 showed significant associations. AS9, on the other hand, showed suggestive associations in the univariable models, which became significant in the multivariable IVW model (Supplementary figures 3.13 and 3.14). The intercept of MR-Egger was not significant and the estimates were consistent across the models.

6.4. Epistasis interaction with rs1967309 in ADCY9

Since we observed the effect of a change in alternative splicing of the exon 9 could impact phenotypes, we next evaluated the effect of the interaction between *ADCY9* and *CETP* genes. We detected significant associations between the SNP rs158477 in *CETP* and SNPs within the LD block containing rs1967309 in *ADCY9* on alternative splicing of exon 9 (Figure 3.4a), with the strongest association with the *ADCY9* SNP rs4786452 in breast mammary tissue (p-value$_{Interaction}$=5.78 x 10^{-7}), whereas the interaction between rs158477 (CETP) and rs1967309 (ADCY9) was not significant for neither alternative splicing events in any of the studied tissues, and only nominally significant for interaction with sex in breast mammary tissue with alternative splicing of exon 9 (p-value=0.02). The mutation rs4786452, however, is in high LD with the mutation rs1967309 in almost all populations from 1000 Genomes project (D'>0.99), but with a lower minor allele frequency in breast mammary tissue (MAF$_{rs4786452}$=15% vs MAF$_{rs1967309}$=38%). The C allele of rs4786452 was always on the same haplotype as the A allele of rs1967309, for which we observed an enrichment in the Peruvian population while considering an interaction with rs158477. This suggests an epistasis interaction between rs158477 and *ADCY9* locus on AS9.

6.5. Analyses with MAJIQ

MAJIQ is a commonly used software for estimating Percent Spliced-In (PSI) values, but it can introduce complexities in analyzing simple alternative splicing events like AS9. In order to validate the performance of a newer and less well-known software called ASpli, we conducted a comparison of PSI values between the two software tools.

We estimated the PSI values from the processes bam file from above with MAJIQ v2 [229]. The configuration files contain the information of the length of RNA-seq reads of 100 and Hg38 reference panel. We did not allow de-novo junctions nor intron retention during the build for MAJIQ. Each tissue had at most three junctions detected. The first one (Junction 1) is for an alternative exon 1, which differentiate isoform *CETP-203* to isoforms *CETP-201/202* and is the same as AS1 from ASpli. The second one (Junction 2) starts at the end of exon 8 and goes to either the beginning of exon 9 or 10. The third (Junction 3) starts at either the end of exon 8 or 9, then finishes at the start of exon 10. The combination of both junctions is similar to AS9 from ASpli.

PSI values were estimated while running MAJIQ with default parameters on all samples separately. 5, 5 and 8 tissues had more than 50 samples with PSI values for junction 1, 2 and 3 respectively, which is less than what was obtained with ASpli. Less samples got a PSI value with MAJIQ, with default parameter, than with ASpli, for which we put a filter to at least 10 reads for a junction instead of the default 5 reads (Supplementary Figure 3.16).

6.5.1. PSI values comparison

ASpli and MAJIQ have two distinct approaches to quantified PSI values. While comparing PSI values for both methods, we observed strong correlation for alternative exon 1 (AS1) in ASpli with junction 1 in MAJIQ, and skipping exon 9 (AS9) in ASpli with junction 2 and 3 in MAJIQ (Supplementary Figure 3.15). Only LCL showed lesser correlation for junction 2 and exon skipping 9, potentially due to the truncated transcript in this cell line (Supplementary Figure 3.6).

We performed sQTL as we did for ASpli in the main method. Hit regions for sQTL were highly similar between both methods. However, observing the strength of the association with a derived F-statistic, in some tissues, we observed an increase in the strength of the association for AS1 with ASpli, but in other tissues, such as spleen, we observed a slight decrease in the strength of the association for AS9.

In MAJIQ for the junction 1, which was estimated with adipocyte, thyroid, breast and spleen tissues, we did not observed an heterogeneity of the effect of the most significant sQTLAS1 rs711752 for this junction (p-value$_{Q-Cochran}$=0.17). For junction 2, which was estimated with the same tissues as junction 1, we also did not observe an heterogeneity of the effect of the most significant sQTL$_{AS9}$ rs5883 for this junction (p-value$_{Q-Cochran}$=0.27). However, for junction 3, which was estimated with the same tissues than the other two junctions, plus pituitary, lung and LCL, we observed significant heterogeneity of the effect of rs5883 on PSI values (p-value$_{Q-Cochran}$=0.005), which was mostly caused by the LCL (p-value$_{Q-CochranwithoutLCL}$=0.08). We did not observe heterogeneity with ASpli caused by LCL, potentially since AS9 is approximately the average from junctions 2 and 3, which could increase the variance of the effect of rs5883 on this junction.

7. Supplementary tables

Outcome	Source	Consortium	Sample Size	Origin
HDL cholesterol	ieu open gwas project : ieu-b-109	UK Biobank	403943	PMID : 32203549
LDL cholesterol	ieu open gwas project : ieu-b-110	UK Biobank	440546	PMID : 32203549
TG	ieu open gwas project : ieu-b-111	UK Biobank	441016	PMID : 32203549
Trunk fat-free mass	ieu open gwas project : ukb-b-17409	MRC-IEU	454508	Ben Elsworth
Whole body fat-free mass	ieu open gwas project : ukb-b-13354	MRC-IEU	454850	Ben Elsworth
Height	GWASatlas id : 3187	UK Biobank	385748	PMID: 31427789
Basal metabolic rate	ieu open gwas project : ukb-b-16446	MRC-IEU	454874	PMID : NA
Thyroid Stimulating Hormone	ieu open gwas project : prot-a-530	NA	3301	PMID : 29875488
Birth weight	ieu open gwas project : ukb-b-13378	MRC-IEU	261932	Ben Elsworth
Birth weight of first child	ieu open gwas project : ukb-b-3357	MRC-IEU	200272	Ben Elsworth
FEV1	ieu open gwas project : ukb-b-19657	MRC-IEU	421986	Ben Elsworth
FVC	ieu open gwas project : ukb-b-7953	MRC-IEU	421986	Ben Elsworth

Table 3.1. Source and information about the continuous phenotypes used in the paper.

The information included the source and the ID to retrieve the data, the consortium from which the data originated, the sample size, and whether the data is derived from a published paper or a laboratory source.

Outcome	Source	Consortium	Number of Cases	Number of Controls	PMID
CVD-related	van der Harst, Pim (2017), "CAD meta-analysis", Mendeley Data, V1, doi: 10.17632/gbbsrpx6bs.1	UK Biobank, CARDIoGRAMplusC4D, Myocardial Infarction Genetics, CARDIoGRAM Exome	122733	424528	PMID : 29212778
Early age-related macular degeneration (AMD)	ieu open gwas project : ebi-a-GCST010723	International AMD genomics consortium (IAMDGC)	14034	91214	PMID : 32843070
Hypothyroidism	GWASatlas id : 4376	NA	3440	49983	PMID : 30367059
Hyperthyroidism	GWASatlas id : 4375	NA	1840	49983	PMID : 30367059
Miscarriage stillbirth	pheweb.org/UKB-TOPMed : 634	UK Biobank	5214	212254	PMID : NA

Table 3.2. Source and information about the discrete phenotypes used in the paper.

The information included the source and the ID required to retrieve the data, the consortium from which the data originated, the number of cases and controls included in the study, and the reference to the paper where the data is reported.

8. Supplementary files

Supplementary file 1. List of the first underlying cause of death (DTHFUCOD) from GTEx dataset, specifically belonging to the variables MHHRTATT and MHHRTDIS. These variables provide information about the CAD-related causes of death for the individuals included in the GTEx dataset.

Supplementary file 2. List of the first underlying cause of death (DTHFUCOD) from GTEx dataset that were removed from the control group in our analysis. This list includes the cause of death associated to the variable MHHRTDISB, the cause of death associated to heart but not linked to MHHRTATT and MHHRTDIS, as well as cases where the cause of death is unknown.

Supplementary file 3. Statistical summary of all SNPs that were used in at least one Mendelian Randomisation analysis conducted on thyroid tissue for the three exposures: eQTL of gene-level *CETP* expression, sQTL of Alternative splicing of exon 9 (AS9) and sQTL of alternative exon 1 (AS1). The information presented includes the beta coefficient, the effect allele, the p-value (pval) of the association, the standard error (se), the derived F-statistic (Fstat) of each SNP for the three exposures. We also identified the outcomes for which the SNP was used and in which analysis. The SNPs included in this list are those that passed the filtering criteria, which required a derived F-statistic of at least 10, a p-value below 0.001, and a minor allele frequency above 0.01 for at least one of the three exposures. Additionally, the SNPs were filtered based on correlation (r^2>0.90) before being included in the analysis.

Supplementary file 4. Genetic correlation matrix for all SNPs that were used in at least one Mendelian Randomisation analysis. The correlation matrix was calculated using the `ld_matrix` function from the R package ieugwasr and was performed specifically on the 699 individuals of European descent in the GTEx dataset. The SNPs included in this matric are those that passed the filtering criteria, which required a derived F-statistic of at least 10, a p-value below 0.001, and a minor allele frequency above 0.01 for at least one of the three exposures. Additionally, the SNPs were filtered based on correlation (r^2>0.90) before being included in the analysis.

9. Supplementary figures

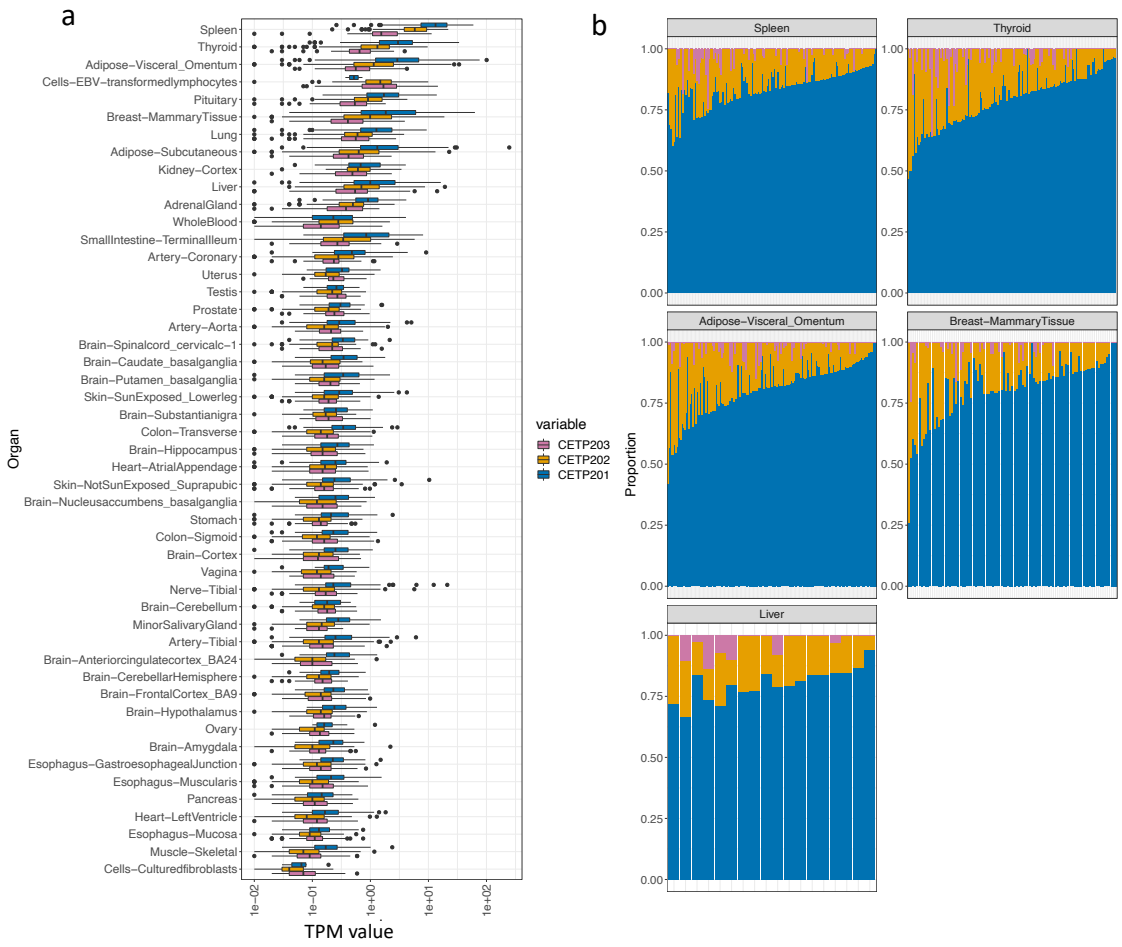

Fig. 3.5. Expression of *CETP* transcripts by tissue in GTEx dataset.
(a) Transcript per million (TPM) for the three protein coding *CETP* isoforms generated by RSEM in GTEx. Values of 0 were removed from this graph and x axis was log-transformed. (b) Proportion of each transcript for each sample for 5 tissues, reported using Proportion-Spliced-In (PSI) values estimated by ASpli. Samples are included only if they had non-zero values for alternative exon 1 and alternative splicing of exon 9.

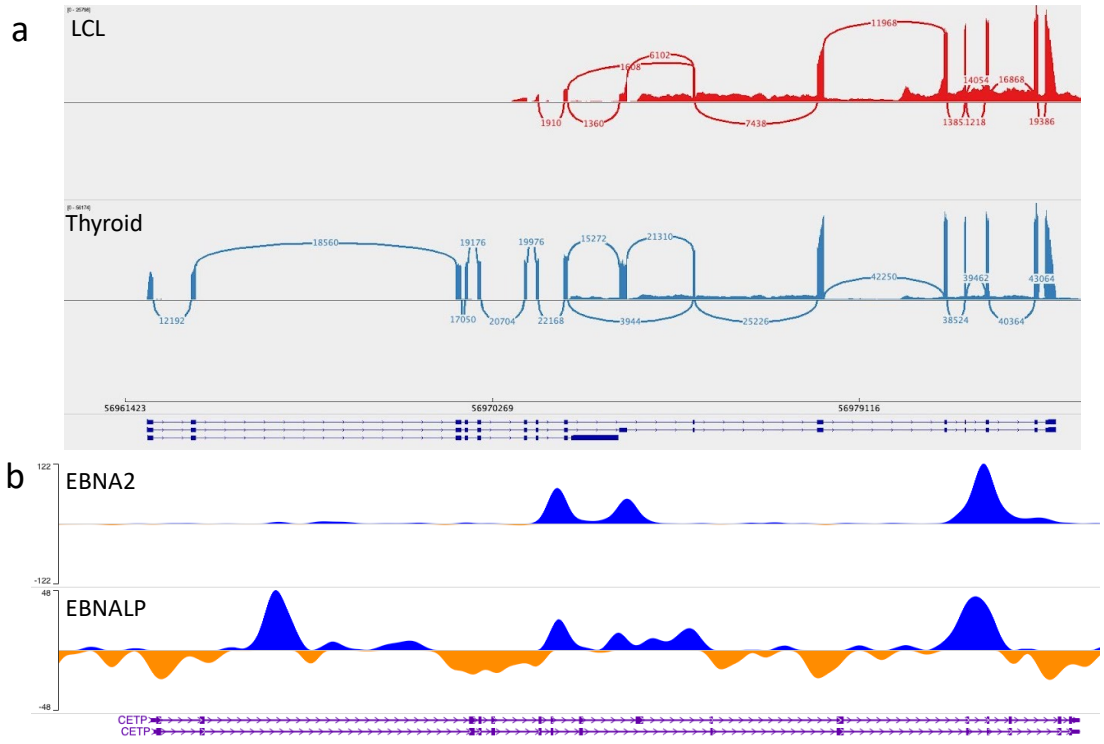

Fig. 3.6. *CETP* isoform in cells-EBV-transformed lymphocytes (LCL)
(a) Sashimi plot of the *CETP* gene in the *CETP* isoform in LCL and thyroid tissue visualized in IGV, made from merging all samples from the LCL (n=174) and thyroid (n=653). Junctions are shown when there was a minimum of 1000 reads junctions for LCL and 2000 for Thyroid. (b) Promoter region of Epstein-Barr virus (EBV) genes from http://epigenomegateway.wustl.edu/browser/ database compared to the coverage of *CETP* in LCL from GEUVADIS.

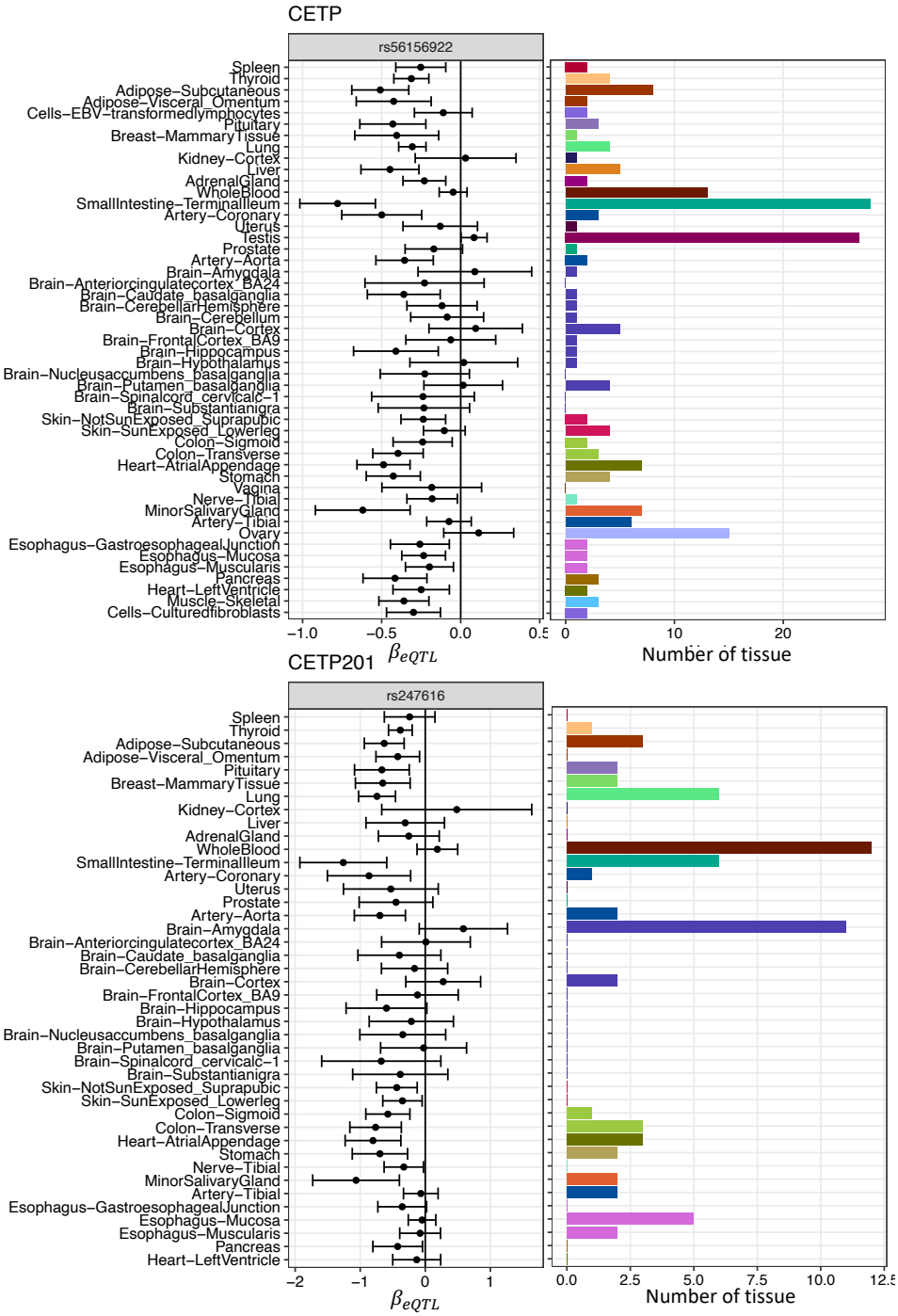

Fig. 3.7. Tissue-specificity of eQTLs for gene-level *CETP* (top) and *CETP-201* (bottom) across tissues

The SNPs chosen are the strongest eQTL in the LD block B-3. The barplot represent the number of tissues the effect size estimates significantly differ from the effect size estimate from the labeled tissue. Tissues are sorted according to the gene-level expression of *CETP*.

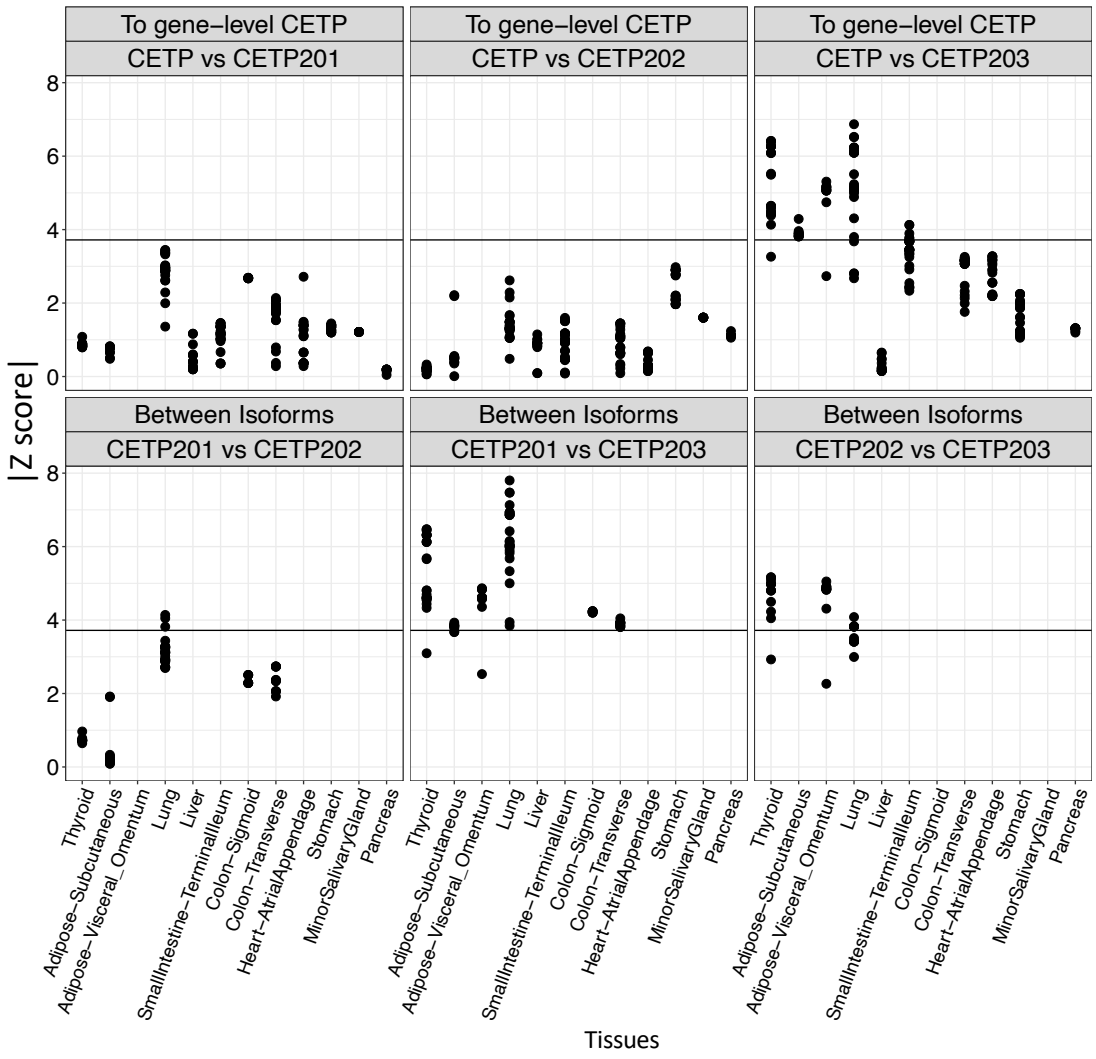

Fig. 3.8. Comparison of the effect size estimate of isoform-level with gene-level *CETP* expression (top) and between isoforms (bottom)
Absolute Z-score of a T-test comparing the effects of all SNPs between two analyses with a p-value < 0.0001 (CETP-wide significance) for at least one analysis in the comparison. Tissues are sorted according to the gene-level expression of *CETP*. Black horizontal lines represent the Z-score corresponding to a p-value of 0.001.

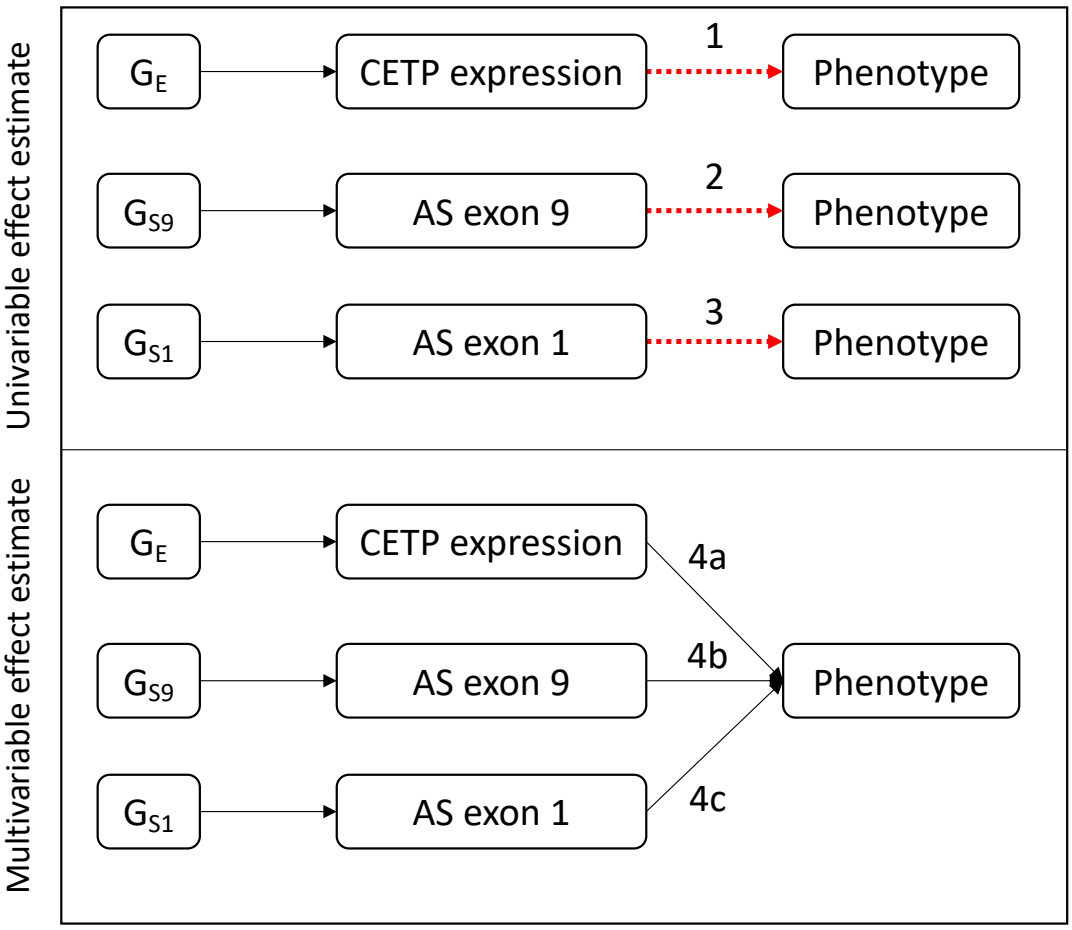

Fig. 3.9. Representation of the effect estimated in the univariable (Top) and multivariable (Bottom) mendelian randomisation (MR) analyses

Arrows indicate causal effect studied in each test. Effect estimate 1 is an univariable MR on *CETP* expression. Effect estimate 2 is an univariable MR on alternative splicing of exon 9. Effect estimate 3 is an univariable MR for alternative exon 1. Effect estimates 4 is a multivariable MR model for which we considered *CETP* expression, alternative splicing of exon 9 and alternative exon 1. Effect estimate 4a is associated to *CETP* expression, effect 4b with *CETP-202* proportion, 4c to *CETP-203* proportion in this model.

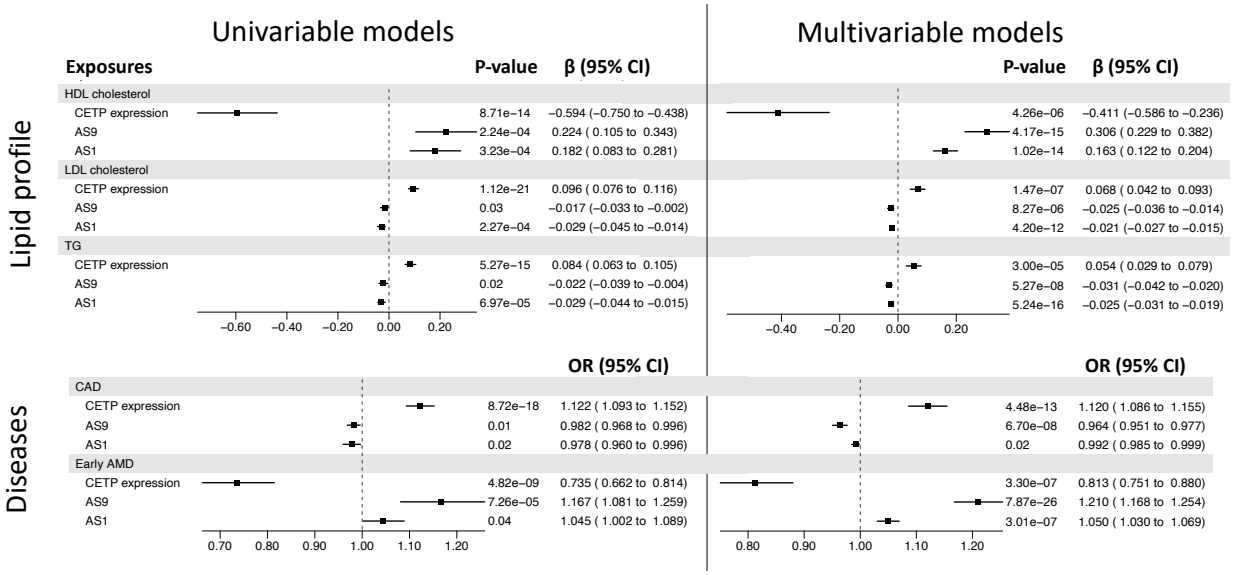

Fig. 3.10. Effects of change in the proportion of *CETP* isoforms using IVW univariable and multivariable Mendelian Randomisation on phenotypes previously associated with gene-level *CETP* expression

Results are from the IVW test. The multivariable MR takes into account gene-level *CETP* expression, alternative exon 1 (AS1) and alternative splicing of exon 9 (AS9). Estimates (β or Odd Ratio (OR) depending on the phenotype) represent the effect of a change of 1 standard deviation on the outcomes.

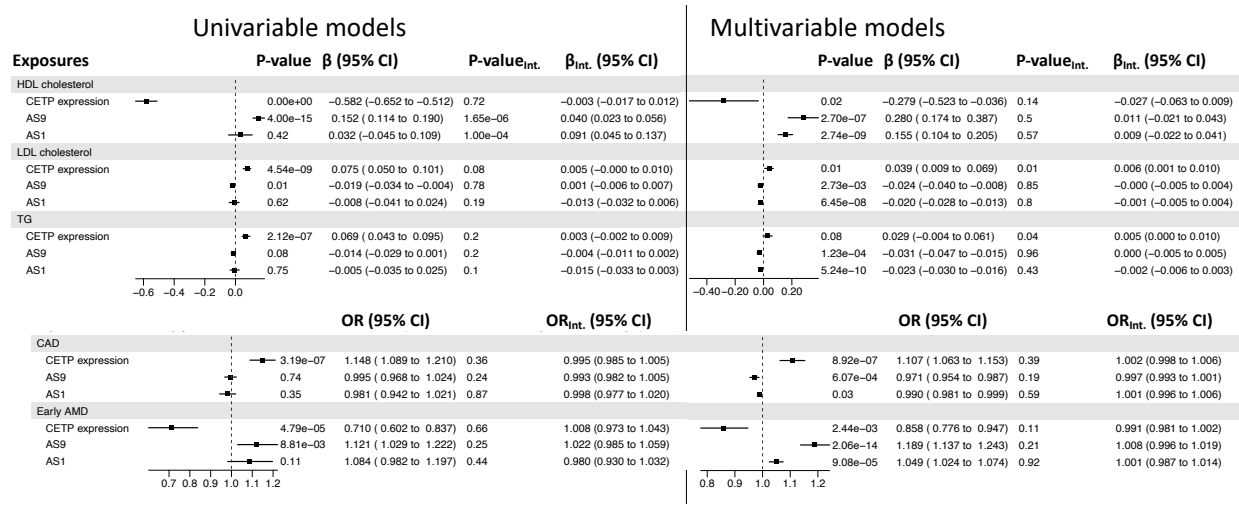

Fig. 3.11. Effects of change in the proportion of *CETP* isoforms using MR-Egger univariable and multivariable Mendelian Randomisation on phenotypes previously associated with gene-level *CETP* expression

Results are from the MR-Egger test. The multivariable MR takes into account gene-level *CETP* expression, alternative exon 1 (AS1) and alternative splicing of exon 9 (AS9). Estimates (β or Odd Ratio (OR) depending on the phenotype) represent the effect of a change of 1 standard deviation on the outcomes. P-value and estimate (β_{Int} or OR_{Int}) of the intercept (Int.) obtain by MR-Egger indicate the presence or absence of horizontal pleiotropy in the model.

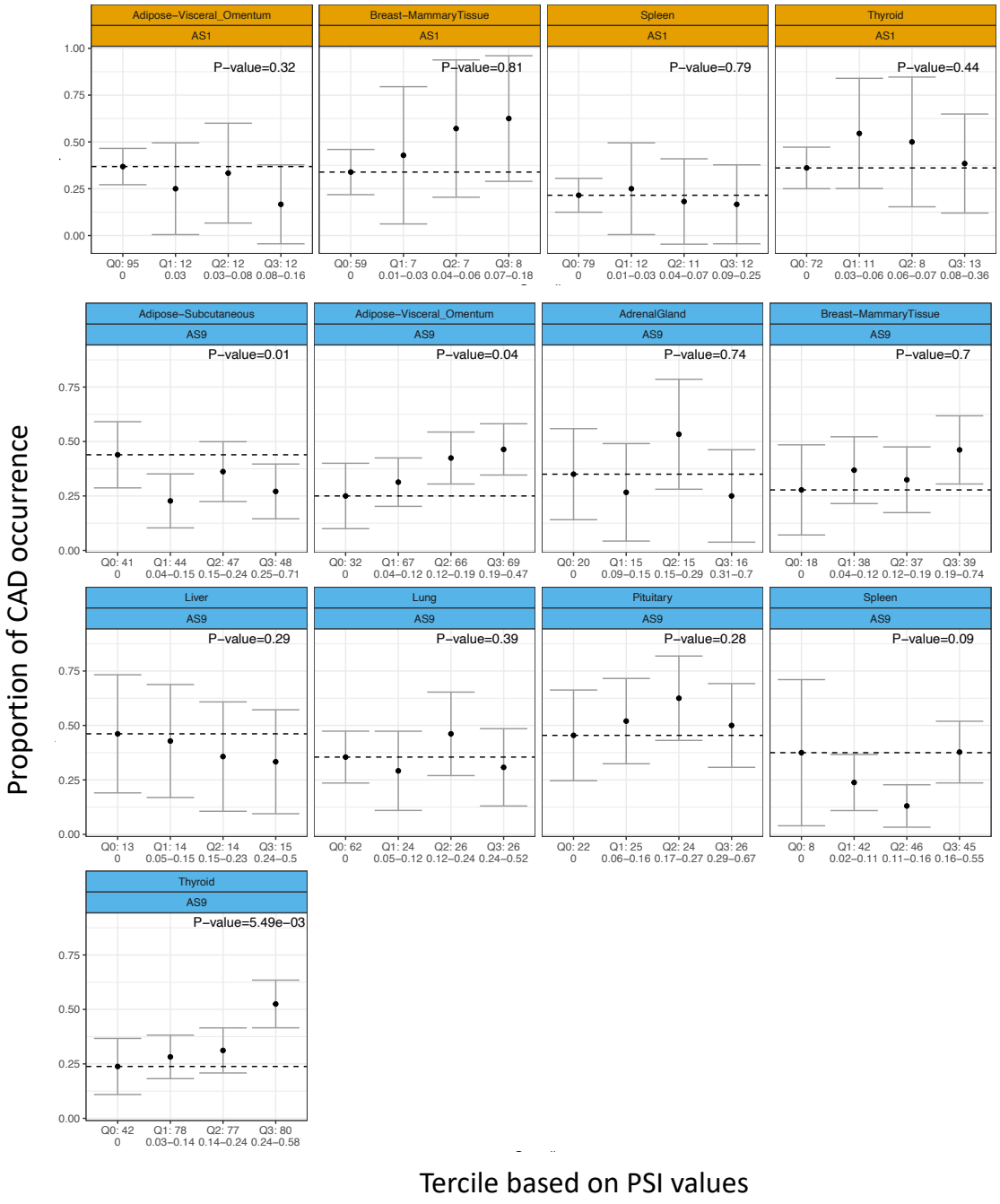

Fig. 3.12. Relationship between Proportion-Spliced-In (PSI) values and proportion of coronary artery disease (CAD) occurrence in GTEx individuals

PSI values were separated in four group (Q0 : PSI values equal 0; Q1 is the first tercile, Q2 the second and Q3 the third). The proportion of individuals with cardiovascular events in each group of PSI values for alternative exon 1 (AS1) alternative splicing of exon 9 (AS9), shown by tissues. Black dashed lines represent the proportion of Q0. P-value reported for each tissue results from a logistic regression model on CAD, using tercile group as categorial variable. The numbers on the x axis represents the number of samples in this group and the interval is the range of PSI values in the group.

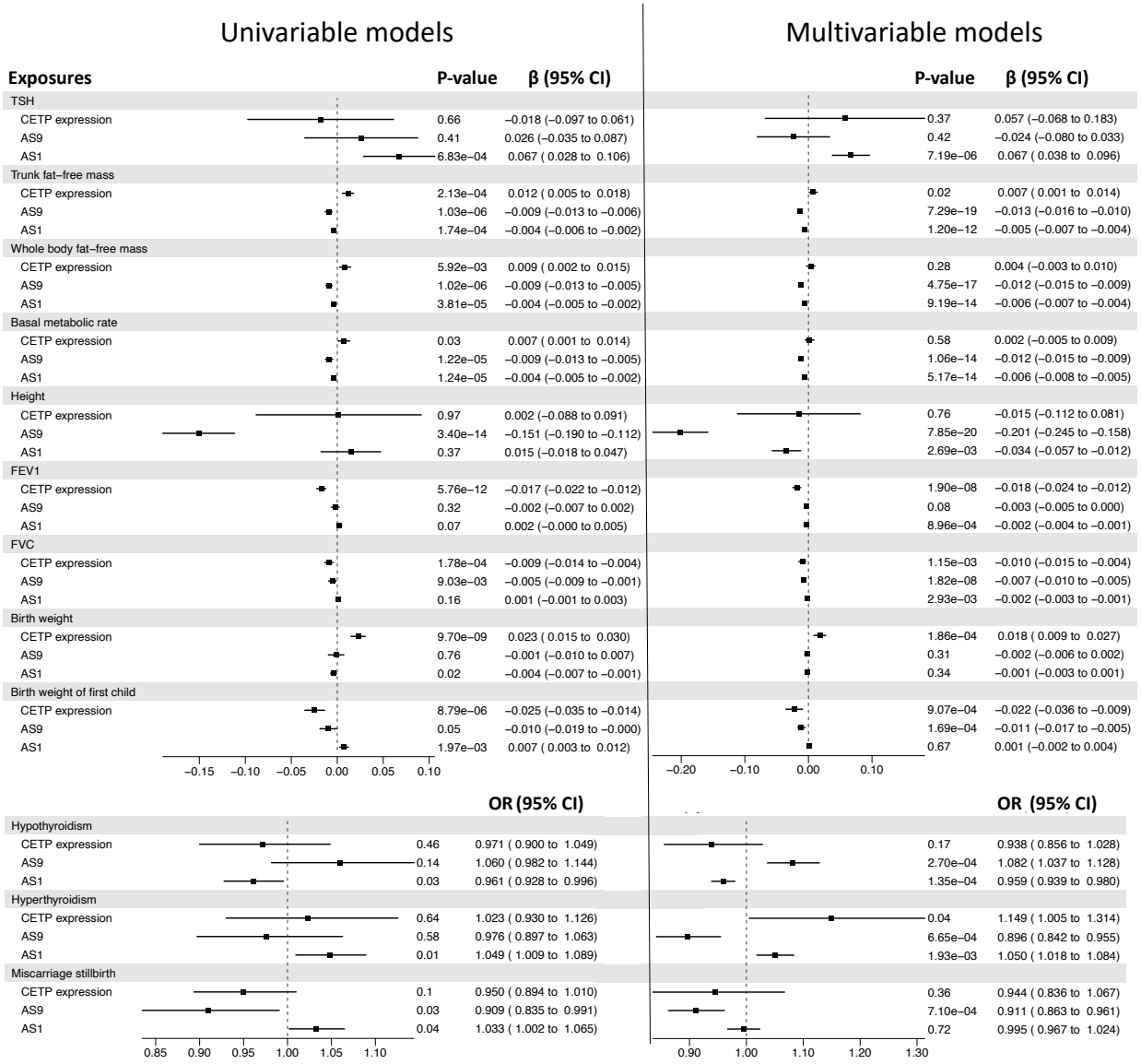

Fig. 3.13. Effects of change in the proportion of *CETP* isoforms using IVW univariable and multivariable Mendelian Randomisation on phenotypes associated with thyroid/pituitary gland or potentially under selective pressure
Results are from the IVW test. The multivariable MR takes into account gene-level *CETP* expression, alternative exon 1 (AS1) and alternative splicing of exon 9 (AS9). Estimates (β or Odd Ratio (OR) depending on the phenotype) represent the effect of a change of 1 standard deviation on the outcomes.

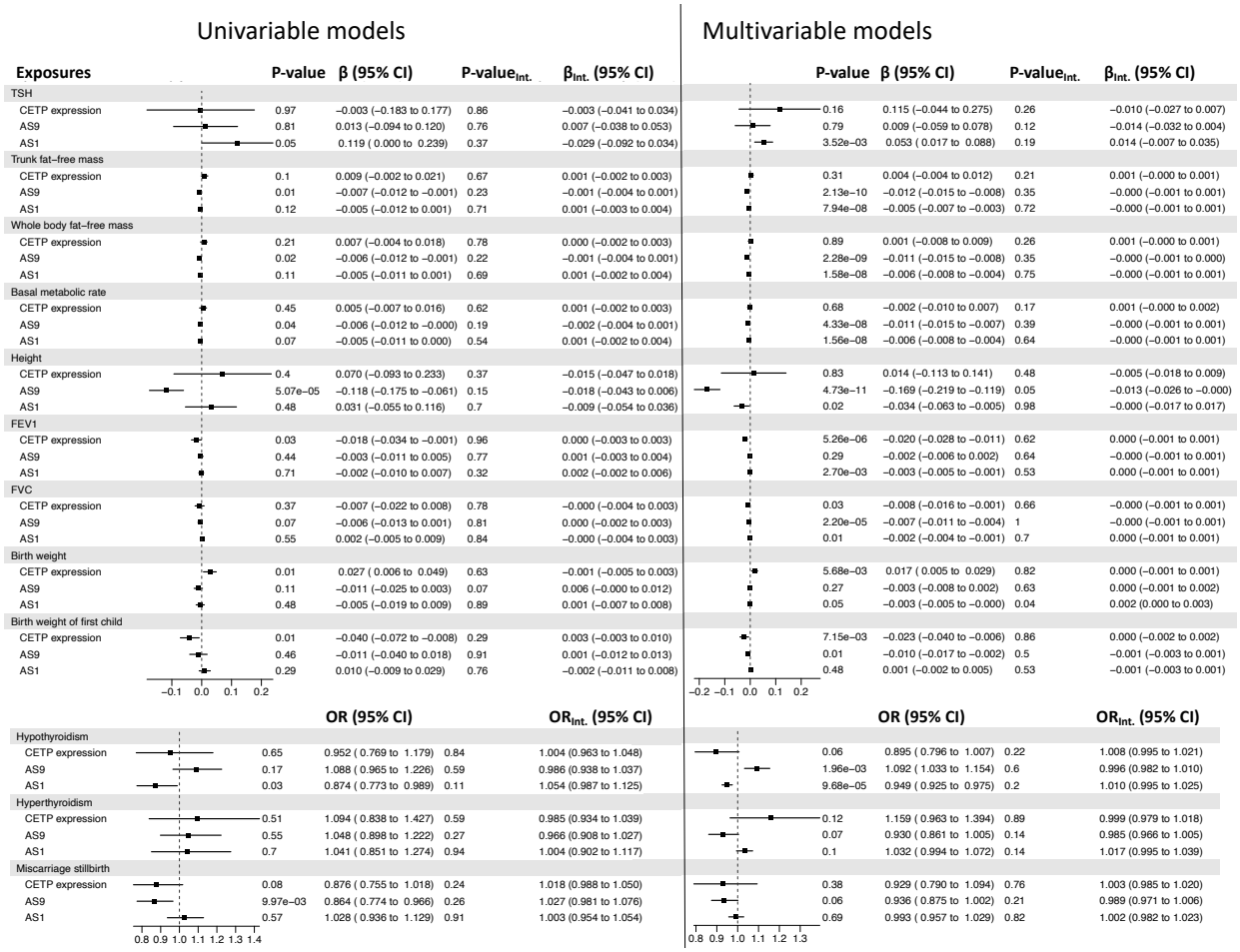

Fig. 3.14. Effects of change in the proportion of *CETP* isoforms using MR-Egger univariable and multivariable Mendelian Randomisation on phenotypes associated with thyroid/pituitary gland or potentially under selective pressure
Results are from the MR-Egger test. The multivariable MR takes into account gene-level *CETP* expression, alternative exon 1 (AS1) and alternative splicing of exon 9 (AS9). Estimates (β or Odd Ratio (OR) depending on the phenotype) represent the effect of a change of 1 standard deviation on the outcomes. P-value and estimate (β_{Int} or OR_{Int}) of the intercept (Int.) obtain by MR-Egger indicate the presence or absence of horizontal pleiotropy in the model.

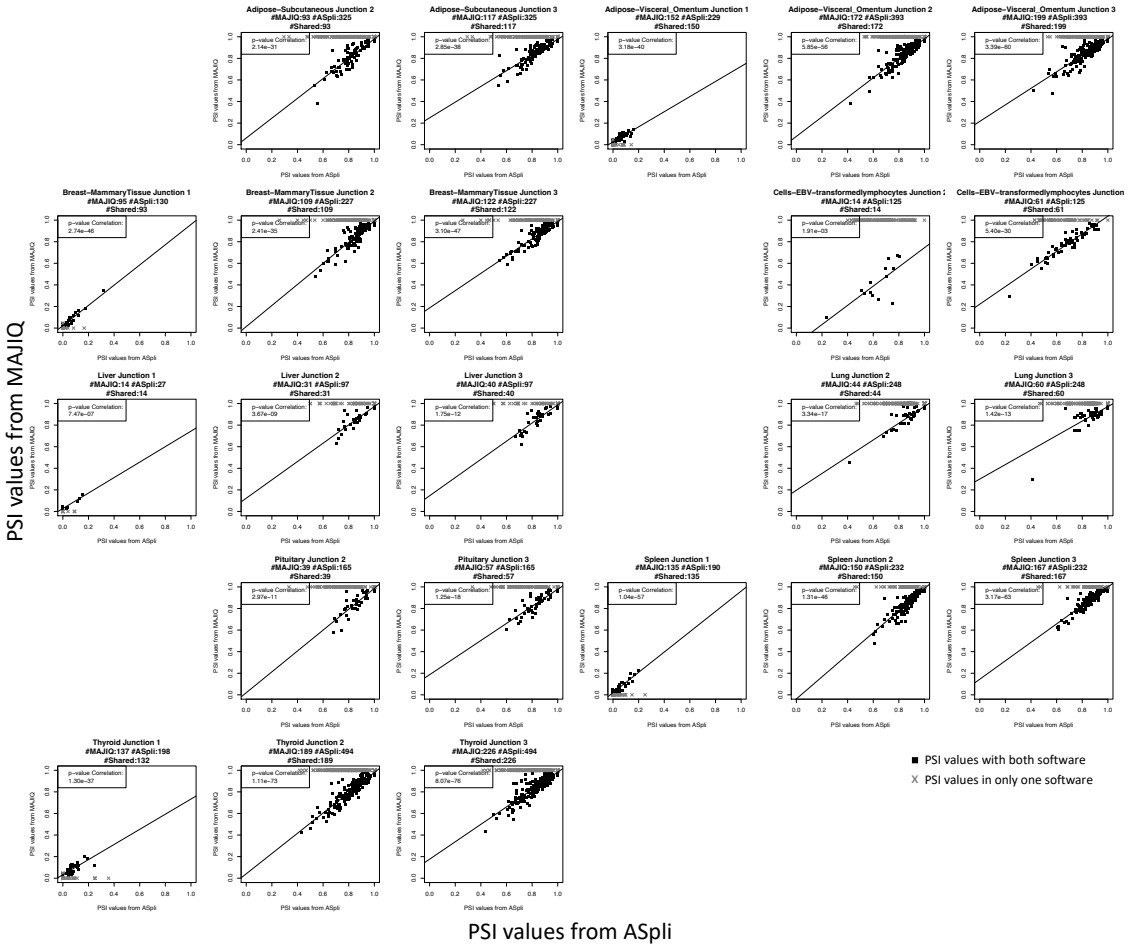

Fig. 3.15. Comparison of Percent Spliced-In (PSI) values obtain by MAJIQ (y axis) and ASpli (x axis) softwares

Junction 1 represents the splicing junction at the beginning of exon 2, quantifying alternative of exon 1 (AS1). Junction 2 represents the splicing junction at the end of exon 8, and junction 3 represents the splicing junction at the beginning of exon 10, both quantifying alternative splicing of exon 9 (AS9). Gray values indicate samples without PSI values in MAJIQ using default parameters, but with values in ASpli.

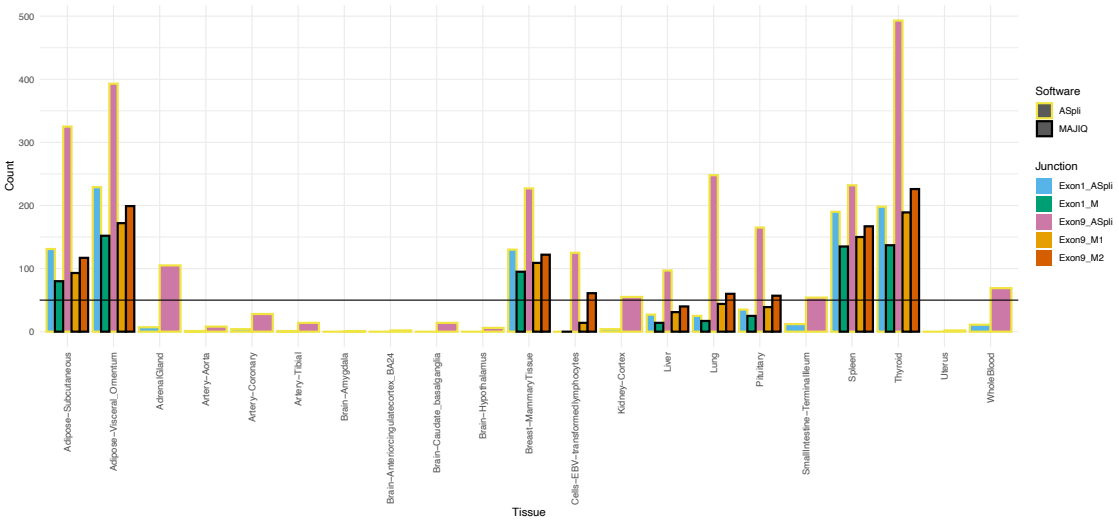

Fig. 3.16. Number of samples with PSI values obtained from MAJIQ and ASpli MAJIQ was used with default parameters, while samples with less than 10 reads of coverages for the junction were filtered out for ASpli. Alternative of the exon 1 are named Exon1_"Sofware" and alternative splicing of the exon 9 are named Exon9_"Software", where "Software" is either MAJIQ (M : AS1, M1 : Junction represents the splicing junction at the end of exon 8, M2 : Junction at the beginning of exon 10) or ASpli. Horizontal black lines indicate the threshold of 50 samples per junction used.

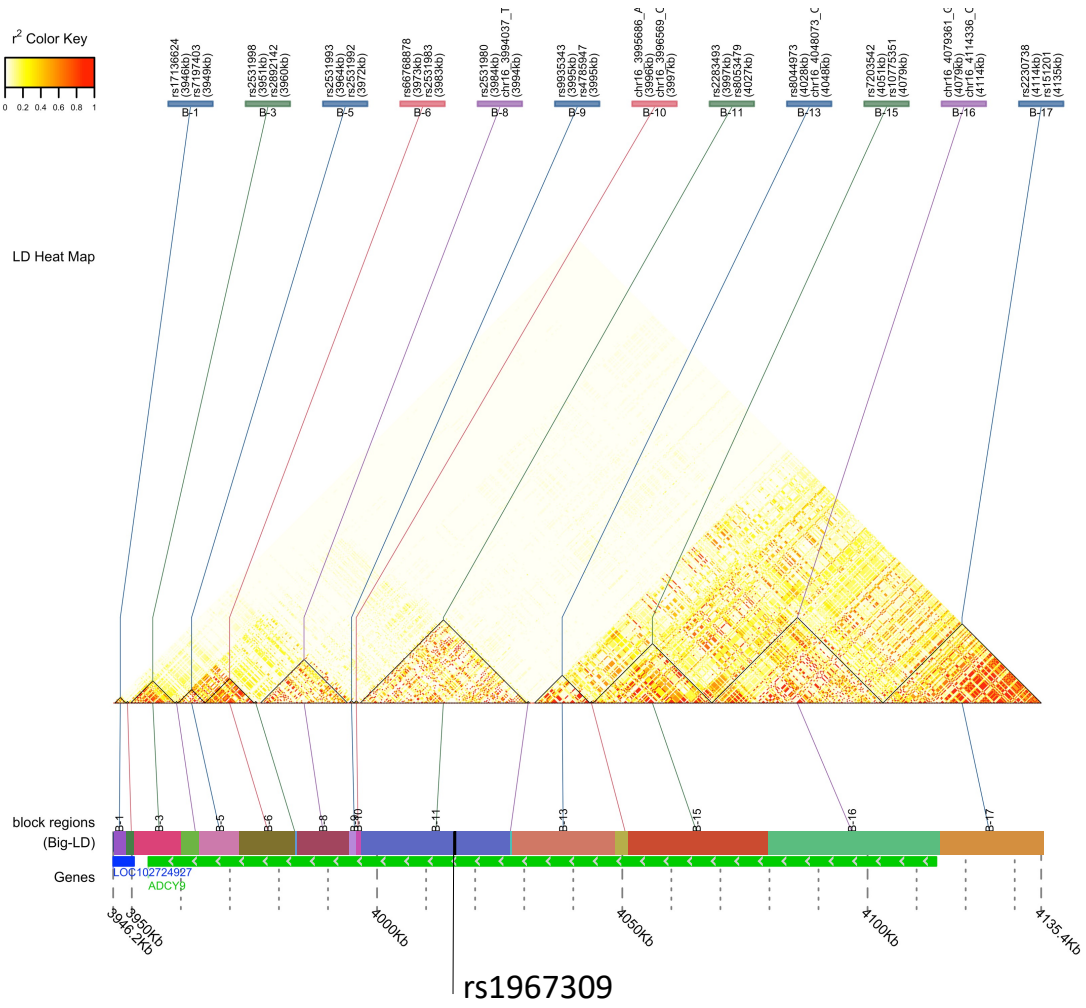

Fig. 3.17. Gene structure of *ADCY9* locus

Linkage Disequilibrium (LD) blocks within the *ADCY9* locus, with surrounding gene identified below and grey arrows indicating 5' to 3'. Blocks were estimated in GTEx participants from European descent using `BigLD` of gpart package with CLQ cut at 0.3 for SNPs having a MAF above 5%. IDs of SNPs delimiting each LD block are indicated above. Position of the mutation rs1967309 is indicated in the corresponding LD block.

Chapitre 4

Signatures of co-evolution and co-regulation in the *CYP3A* and *CYP4F* genes in humans

Contributions à ce chapitre

Mes contributions à l'article inclut dans ce chapitre sont les suivantes en tant que co-première auteure :

- Traitement des données d'expression du jeu de donnée de GTEx, calcul des covariables (facteurs PEER, PCA) et rédaction des scripts initiaux de ces analyses
- Analyses phénotypiques
  - Traitements et analyses des données phénotypiques de la base de données du UK biobank pour le PheWAS
  - Génération des données utilisées lors des analyses de randomisation mendélienne (Normalisation des données d'expression, Calcul de la force d'association entre la mutation et l'expression de *CYP3A5* (Statistique F-dérivée), Calcul des associations entre les mutations et les phénotypes)
  - Analyses de randomisation mendélienne
- Aide à l'évaluation des résultats obtenus par Alex Richard-St-Hilaire
- Aide à la révision du manuscrit, écriture des sections de l'article sur les analyses phénotypiques (résultats, méthodes, discussion)
- Réponses aux réviseurs
  - Correction des données de la statistique du D de Tajima
  - Génération des figures dans la réponse aux réviseurs (D de Tajima, Score Beta)
  - Ajustement du texte selon les commentaires des réviseurs

Ce travail n'aurait pas été possible sans l'aide des personnes suivantes :

- Alex Richard-St-Hilaire
  - Effectuer toutes les analyses initiales de pression de sélection
  - Analyser les données d'expression pour les gènes des sous-familles *CYP3A* et des *CYP4F* dans la base de données de GTEx
  - Rédaction du manuscrit et des figures, à l'exception des sections associées aux analyses phénotypiques (dans la section résultats, méthodes et discussion)
- Julie G Hussin :
  - Supervision du projet
  - Assistance à la préparation du manuscrit
- Justin Pelletier :
  - Génération des analyses préliminaires de sélection balancée
- Jean-Christophe Grenier :
  - Pré-traitement des données génétiques
  - Génération des valeurs iHS
  - Assistance dans la préparation du manuscrit et de sa révision
- Raphaël Poujol :
  - Assistance à la préparation du manuscrit

Signatures of co-evolution and co-regulation in the *CYP3A* and *CYP4F* genes in humans

by

Alex Richard-St-Hilaire[1,2], Isabel Gamache[1,3], Justin Pelletier[1,4], Jean-Christophe Grenier[3], Raphael Poujol[3], and Julie G Hussin[3,5]

([1]) Département de biochimie et médecine moléculaire, Université de Montréal, Montreal, Qc, Canada

([2]) Sainte-Justine hospital, Research Center, Montreal, Qc, Canada

([3]) Montreal Heart Institute, Research Center, Montreal, Qc, Canada

([4]) McGill University & McGill CERC Genomic Medicine, Montreal, Canada

([5]) Département de médecine, Université de Montréal, Montreal, Qc, Canada

This article was resubmitted to *Genome Biology and Evolution* after minor revisions.

1. Abstract

Cytochromes P450 (CYP450) are hemoproteins generally involved in the detoxification of the body of xenobiotic molecules. They participate in the metabolism of many drugs and genetic polymorphisms in humans have been found to impact drug responses and metabolic functions. In this study, we investigate the genetic diversity of *CYP450* genes. We found that two clusters, *CYP3A* and *CYP4F*, are notably differentiated across human populations with evidence for selective pressures acting on both clusters: we found signals of recent positive selection in *CYP3A* and *CYP4F* genes and signals of balancing selection in *CYP4F* genes. Furthermore, an extensive amount of unusual linkage disequilibrium is detected in this latter cluster, indicating co-evolution signatures among *CYP4F* genes. Several of the selective signals uncovered co-localize with expression quantitative trait loci (eQTL), which could suggest epistasis acting on co-regulation in these gene families. In particular, we detected a potential co-regulation event between *CYP3A5* and *CYP3A43*, a gene whose function

remains poorly characterized. We further identified a causal relationship between *CYP3A5* expression and reticulocyte count through mendelian randomization analyses, potentially involving a regulatory region displaying a selective signal specific to African populations. Our findings linking natural selection and gene expression in *CYP3A* and *CYP4F* subfamilies are of importance in understanding population differences in metabolism of nutrients and drugs.

Keywords: Cytochrome P450, Population genetics, Coevolution, PheWAS, Mendelian Randomisation, Malaria

2. Introduction

In the last decades, it has become clear that every individual has their own "fingerprint" of alleles encoding drug-metabolizing enzymes, playing central roles in the metabolism of endogenous and exogenous compounds. It was established that hydrophobic molecules are first modified by oxidation and subsequently excreted as water-soluble forms, two distinct steps now described as phases I and II. Phase I is performed mainly by Cytochromes P450 (CYP450) enzymes, able to catalyze a considerable variety of oxidation reactions for many structural classes of chemicals (including the majority of drugs) [405, 406]. They metabolically activate parent compounds to electrophilic intermediates, while Phase II enzymes conjugate these intermediates towards more easily excretable derivatives.

CYP450 genes are a super-family of genes which appeared more than 3.5 billion years ago [407], being present in fungi, plants, bacteria and animals. Genes are grouped into families and subfamilies based on sequence similarity: genes from the same family have sequence similarity greater than 40 % and, to be grouped into a subfamily, their sequence similarity must be greater than 55 % [408].

In humans, the *CYP450* family includes 57 genes and 58 pseudogenes [409] grouped in 18 families [410]. Several *CYP450* genes are found in clusters in the human genome but some members of the subfamilies can be spread out across the genome. For example, the CYP4F subfamily has genes on chromosome 19 and pseudogenes on multiple chromosomes. The *CYP2D6* gene is the most widely studied *CYP450* gene in humans, due to its role in the metabolism of many drugs [411, 412] along with *CYP3A4* and *CYP3A5*, members of the *CYP3A* subfamily[413, 414, 415, 416, 417]. However, not all *CYP450* genes or families have

been studied thoroughly, and details on the evolution and clinical significance are lacking for several families, such as the *CYP4F* subfamily.

Several *CYP450* genes are potential candidates that underwent natural selection in humans [153, 418]. Other studies of the genetic diversity for specific *CYP450* subfamilies in human populations confirmed the presence of positive [18, 419], balancing [148] or purifying selection signatures [420]. One example is *CYP2C19*, involved in the metabolism of clopidogrel [421, 422], where signals of positive selection on its alleles conferring slow metabolism (*CYP2C19**2 and *CYP2C19**3) were detected using relative extended haplotype homozygosity (REHH)[148]. *CYP2C19**2 is detected worldwide, but *CYP2C19**3 is only present in people of Asian descent. The selective advantages may have been caused by diet and environmental pollutants impacting humans over thousands of years and could differ between ethnic groups. Additionally, low F_{ST} values across *CYP2C19* SNPs suggest balancing selection in *CYP2C19* [148]. The excess of alleles at intermediate frequencies could reflect the evolution of balanced polymorphisms, which is to be expected in evolutionarily old enzymes responsible for numerous critical life functions.

Moreover, the detection of natural selection signals in the *CYP450* genes raises the possibility that the selective advantage acts on polymorphisms that modulate gene expression, widely known as expression quantitative trait loci (eQTL) [423]. Detecting eQTLs linked to selection signals helps clarifying how gene expression is regulated and can lead to a better understanding of variants' biological effects [424]. Furthermore, analysing eQTLs helps in the detection of gene-gene interaction [425] and co-regulation between genes [242]. Such gene-gene interactions can also be detected by looking at patterns of linkage disequilibrium (LD), as evolution will maintain co-evolving polymorphisms on the same haplotypes [426], which can also be detected as balancing selection signatures.

Here, we investigated genetic diversity and selective pressures across human populations in *CYP450* genes. Two subfamilies stood out in our analyses and were investigated in greater depth: the *CYP3A* and *CYP4F* families. Both subfamilies were generated by duplication events resulting in consecutive genes in the same genomic region, or gene cluster (Figure 4.7). The CYP3A subfamily contains four genes and four pseudogenes located in a genomic region of about 220 KB on chromosome 7. They metabolize around 50% of common drugs. The CYP4F subfamily has six genes located in a genomic region of about 430 KB on chromosome

19 and have mostly been associated with metabolism of lipids. We found that both families exhibit selective pressures in human populations and that the SNPs under selection could impact gene expression levels in several tissues. Furthermore, our results suggest interactions between the genes in both *CYP450* subfamilies, providing evidence of co-evolution and co-regulation within these gene clusters, that may vary between populations.

3. Methods

3.1. 1000 Genomes genetic data

The data analyzed is from the Phase III of the 1000 Genomes project (1000G) [252]. The 1000 Genomes Project includes 2,504 individuals from 26 populations. These populations can be split into 5 distinct genetic ancestries referred herein as super-populations, as defined by the 1000G consortium: African (AFR), European (EUR), South Asian (SAS), East Asian (EAS) and Admixed American (AMR). Data from the AMR population is not included in this study because the high degree of admixture may confound selection and linkage disequilibrium analyses. This left us with 22 sub-populations and four super-populations for study. The available variant call format (vcf) files of 1000G are under the GRCh37 genome build. VCFtools v0.1.14 [281] was used to filter the 1000G dataset. Indels and non-biallelic alleles were removed and only SNPs located in the 57 CYP450 genes were kept, extracted based on coordinates genomic coordinates obtained from the UCSC genes table using the UCSC Genome Browser (Supplementary file 5).After filtering, the CYP450 dataset included a total of 61,739 SNPs and 2157 individuals. We refer to this as the "1000G CYP450 dataset". A more recent dataset was also used as a validation dataset for the unusual linkage disequilibrium analysis, the re-sequencing dataset of 30X coverage, mapped on GRCh38 [427], which includes the 2157 individuals.

3.2. Genetic diversity and population differentiation

Both Tajima's D and F_{ST} statistics were obtained with VCFtools [281] using the 1000G CYP450 dataset. Tajima's D values were calculated in the super-population (AFR, EUR, EAS and SAS) separately on non-overlapping windows of 1 Kb. We also performed these analyses excluding positions and windows with low mappability (see Supplementary text

7.1, Figure 4.8). We computed the mean Tajima's D value for each gene by averaging the window-based values, and sorted genes according to their mean. To create a null distribution, we computed Tajima's D values for all SNPs associated with a gene name in the CADD (Combined Annotation Dependent Depletion) annotation file [428] on chromosome 22, so that all SNPs used to compute the empirical distribution are located in genes. We computed the 2.5 and 97.5th percentile on the window-based values of chromosome 22. Values above the 97.5th percentile and below the 2.5th percentile were considered to be statistically significant (two sided empirical p-value <0.05). To ensure that our results were not biased by fine-scale population structure, we also perform the analyses in each of the subpopulations of EUR (see Supplementary text 7.1, Figure 4.8). The F_{ST} values, from Weir and Cockerham derivation [429], were calculated using four super-populations (AFR, EUR, EAS and SAS) on a per-site basis. The per-gene mean was calculated on raw values and genes were sorted based on their mean F_{ST}. As in the previous analysis, chromosome 22 was used to create an empirical distribution. F_{ST} values were also computed on SNPs located in genes of the chromosome 22 (see above) and the per-gene mean F_{ST} was calculated.

3.3. Detecting natural selection

The method used to detect balancing selection is the β score [430]. This score has already been calculated on the whole 1000 Genomes project data for each sub-population. The approach used to detect signal of recent positive selection was iHS (integrated haplotype score) [431]. The iHS computation was performed by us on the 1000G dataset , filtered to exclude INDELs and CNVs. Reference alleles from filtered 1000 Genomes vcf files were changed to the ancestral alleles retrieved from 6 primates EPO pipeline (version e59) using the `fixref` plugin of bcftools [432]. The hapbin program v.1.3.0 [433] was then used to compute iHS using per population-specific genetic maps computed by Adam Auton on the 1000 Genomes OMNI dataset [252].. When the genetic map was not available for a sub-population, the genetic map from the closest sub-population was selected according to their global F_{ST} value computed on the 1000G dataset. For all natural selection analyses, SNPs annotated to be in a repetitive region were identified using the RepeatMasker track available on the UCSC genome browser [434] and were removed.

3.4. Unusual Linkage disequilibrium

Linkage disequilibrium between pairs of SNPs from the same cluster was assessed using the `geno-r2` option from VCFTools on SNPs with minor allele frequencies (MAF) above 0.05. The genetic position of each SNP was calculated with PLINK v1.90 [435] using the population-specific genetic maps the same was as described in previous section.

To compute a null distribution to detect unusual linkage disequilibrium (uLD), the Human GRCh38 Gene transfer format (GTF) file from Ensembl v87 was screened per autosomal chromosome using an in-house python script to find windows matching the *CYP4F* cluster: windows of 430 Kb containing 6 genes were kept. In these windows, we excluded INDELs and SNPs with MAF <0.05. The r^2 for each pair of SNPs located within a selected window was computed using VCFtools with the `geno-r2` option. We divided the genetic distance into bins of 0.01 cM and calculated the 99th percentile of r^2 values of each pair of SNPs lying in the bin. This process was done separately for each 1000G sub-population, yielding a null distribution per sub-population. r^2 values on pairs of SNPs in the extremes of the empirical distribution are considered to be significant for what we called unusual linkage disequilibrium (uLD).

To specifically confirm the signal seen between *CYP4F12* and other *CYP4F* genes, we extracted only the SNPs showing significant uLD in the previous analysis and kept only those pairs where one SNP was located in *CYP4F12*. Using VCFTools, *CYP4F* genetic data was extracted from the 1000 Genomes 30X on GRCh38 dataset [427] and r^2 values were calculated as described above.

3.5. eQTLs analysis of SNPs under selection

The Genotype-Tissue Expression v8 (GTEx)[436] was accessed through dbGaP (phs000424.v8.p2, dbgap project #19088) and contains gene expression across 54 tissues and 948 donors as well as genotyping information, compiled in a VCF file by GTEx on the GRCh38 genome build. The cohort comprises 67% males and 33% females, mainly of European descent (84.6%), aged between 20 and 79 years old. Analyses were done on 699 individuals of European descent, as described in Supplementary text (*Pre-processing of GTEx genetic data*). To take into account hidden factors, we calculated PEER factors on

the normalized expressions. We removed tissues with less than 50 samples, leaving samples from 50 different tissues.

For eQTL analyses, we selected only SNPs that were identified to be under positive or balancing selection in CYP3A and CYP4F clusters in previous analyses and with a MAF above 5%. Since the positions of these SNPs were in the GRCh37 genome build, we converted these positions to the GRCh38 genome build to match GTEx v8 data, using the `liftOver` function of the `rtracklayer` R library [437]. P-values of associations between each selected SNP and gene expression of every gene in the cluster were calculated with a linear model using the `lm` function in R. The linear regression was calculated on each SNP individually. The covariates include the first 5 principal components (PCs) (see Supplementary text), age, sex, PEER factors, the collection site (SMCENTER), the sequencing platform (SMGEBTCHT) and total ischemic time (TRISCHD). To report genome-wide significant eQTL signals, we used a p-value threshold for significance at 10^{-8}.

Lastly, we have searched for regulatory annotations of at eQTL signals using the UCSC Genome Browser, specifically looking at the data provided by the ReMap density database [438].

3.6. Phenotypic associations

The UK biobank (UKb) [115] was accessed through project 15357. We kept only individuals of European descent which were within 3 standard deviation of the mean for the top 3 PCs, removed one individual for each pair of related individuals, and removed individuals whose genetic sex did not match self-identified sex. We extracted positions for CYP3A (chr7:99-244-812-99-470-881, GRCh38) and for CYP4F (chr19:15-618-335-16-110-830, GRCh38) families. We then removed positions with more than 10% of missing genotype and with a MAF under 1%, then removed individuals with more than 5% missing genotypes, leaving us with 399,149 individuals and 3,092 variants for the CYP4F genes, and 400,504 individuals and 374 variants for the CYP3A genes

We used baseline values for continuous phenotypes. We selected phenotypes recommanded by the UKb, as well as blood cells measurements, a total of 90 and 11 phenotypes respectively (Table 4.3). When many values were available at the baseline, we took the mean

of those values. We also looked at diseases using phecode coding extracted from phewascatalog [439], which indicated ICD-10 to group and to exclude from controls. We kept only phecodes with more than 500 cases, leaving 603 for both sexes, 62 female-only and 11 male-only. Covariates used are the age at baseline, sex, top 10 PCs, deprivation index and the genotyping array. Analyses were done with `plink2` [435] with linear transformation of the quantitative covariates. We used a p-value threshold for significance at $6.44x10^{-5}$, based on Bonferonni correction for the number of phenotypes evaluated (0.05/(90+11+603+62+11)=6.44 x 10^{-5}).

We performed Mendelian randomisation analyses. As instrument variables, we selected SNPs in CYP3A cluster showing strong associations with CYP3A5 expression in lung (exposure), with a F-statistic above 10 and a p-value under 0.001 (0.05/50 tissues), then removed SNPs in pair with a correlation above r^2 >0.8, estimated on Europeans from GTEx using the function `ld_matrix` from `ieugwasr` package in R [440], leaving 8 SNPs for analyses. Furthermore, we used the scale function on continuous traits and gene expression to estimate the change of 1 standard deviation (SD) of the phenotypes for 1 SD of the gene expression. As outcomes, we used the 6 phenotypes for which SNPs under selection showed associations for both phenotype (p-value<6.44 x 10^{-5}) and CYP3A5 expression (p-value<0.001). Mendelian randomisation analyses were performed using `MendelianRandomization` package in R [367] and correlation matrix generated using `ld_matrix`[440] was given to the function to adjust for linkage desequilibrium. We performed Inverse Variance Weighted (IVW) as the main statistical test, and we performed MR-Egger to detect and correct for directional pleiotropy: we report MR-Egger results if the intercept was significant. Lastly, we performed weighted median test as a sensitivity test.

4. Results

We obtained genotypic data from the 1000 Genomes project phase 3 release[252] (1000G). A total of 2,157 individuals were analyzed from 22 populations, belongs to four of the five super-populations (ie. Africa, Europe, East Asia and South Asia).

4.1. Global genetic diversity across populations in CYP450 genes

First, we aimed to identify global genetic patterns by calculating Tajima's D values for each *CYP450* genes in each population of the 1000G dataset to provide insights into the

non-neutral forces that act on these genes. A total of 61,739 biallelic SNPs were analyzed in all of the 57 *CYP450* genes, and for each gene, we computed the mean Tajima's D per gene and also in 1 Kb windows. Significantly low Tajima's D values indicate an excess of rare alleles, whereas significantly high values of Tajima's D suggest an excess of intermediate frequency alleles, which can reflect the occurrence of balancing selection.

In European populations, nine genes had Tajima's D values consistently below 0 (Figure 4.1A). We assessed significance based on the empirical (null) distribution, which allows to determine whether any genes have values that are higher or lower than expected while taking population-specific demographic factors into account (see Methods). The proportion of 1 Kb-windows of each gene lying outside the null distribution is shown in Figure 4.1B. *CYP26A1*, *CYP27B1* and *CYP1A2* had the largest proportion of windows with significantly low D values, however these genes are quite small (4.4, 4.9 and 7.8 Kb, respectively), meaning that the signal is driven by one or two windows only. Interestingly, the four CYP3A genes in our dataset were all included in this group of nine genes, suggesting that strong purifying selection pressures may be acting, however complete selective sweeps driven by positive selection can also create this lack of diversity [441]. Notably, CYP3A5 has a low Tajima's D average but no 1 Kb-window is significantly lower than expected, whereas other CYP3A genes have several windows showing significantly low Tajima's D values. All *CYP450* genes show negative Tajima's D values, as expected in coding regions, but ten genes have a mean above 0, which suggests relaxation of purifying selection pressure. The presence of several 1 Kb-windows significantly enriched for high D values can also reflect the presence of localized balancing selection signatures within these genes. Of these ten genes, five are in the CYP4F subfamily: *CYP4F3*, *CYP4F11*, *CYP4F12*, *CYP4F8* and *CYP4F2*. The strongest of these signals is seen on *CYP4F12* (Figure 4.1B). Interestingly, the only CYP4F gene that does not show this specific signature is *CYP4F22*, which is the ancestral gene of the *CYP4F* cluster [442]. Notably, these analyses were also performed for each of the subpopulations, yielding similar results (Figure 4.8).

Because population differentiation can also help identifying natural selection signatures within genes, we calculated the mean fixation index (F_{ST}) across *CYP450* genes (Methods). F_{ST} measures the differentiation between populations using genotype frequencies, with high F_{ST} values indicating that the average pairwise heterozygosity is higher between than within

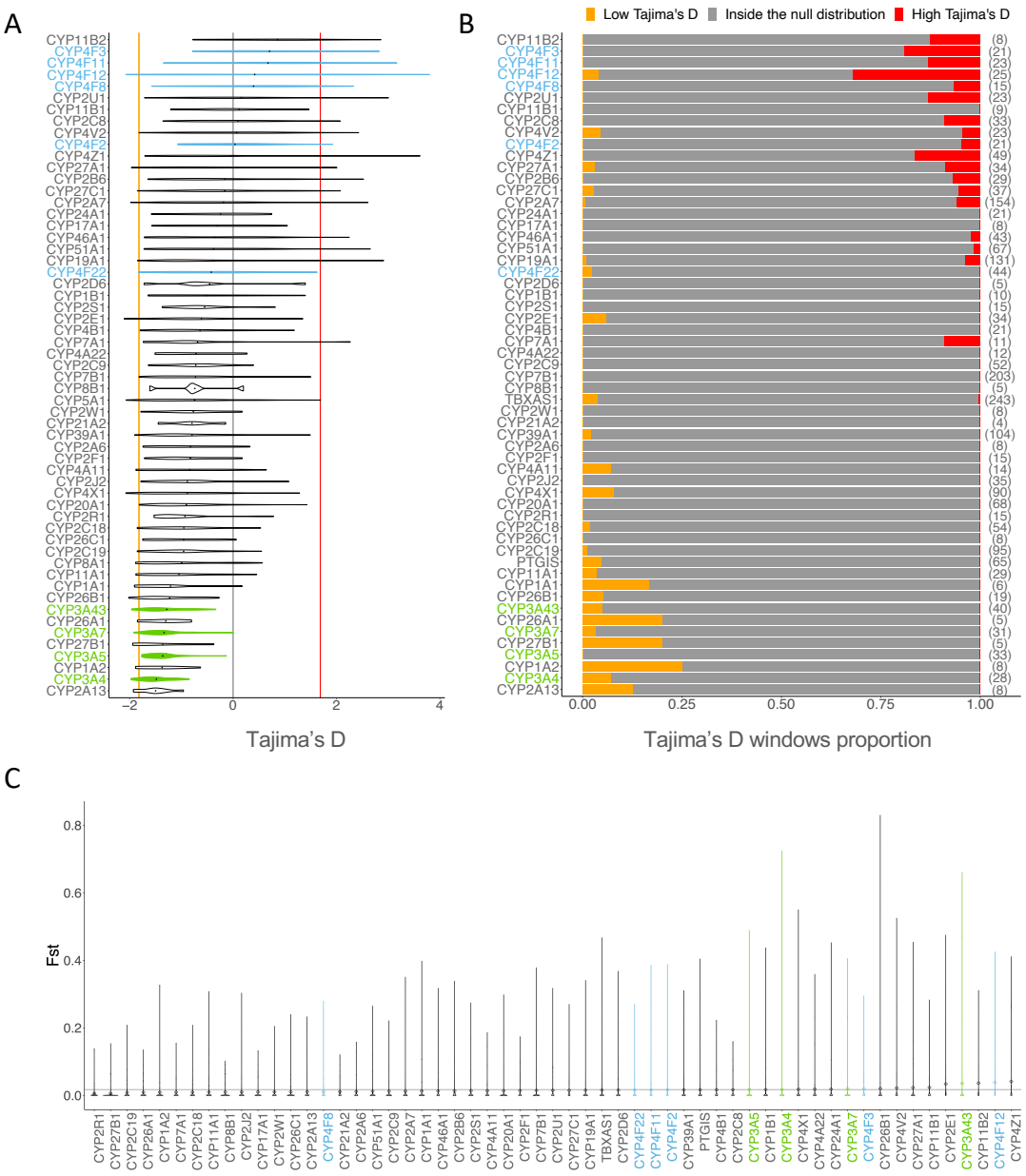

Fig. 4.1. Metrics of diversity and differentiation among *CYP* genes.
A) Distribution of Tajima's D values computed on windows of 1 Kb for each CYP450 genes in the European populations. The 2.5th percentile is marked by the orange vertical line and the 97.5th percentile is marked by the red vertical line, representing the significance threshold. B) Proportion of Tajima's D windows lying outside the null distribution for each CYP450 gene. For each gene, the total number of windows of Tajima's D is shown beside the proportions, between brackets. The windows with Tajima's D values below the 2.5th percentile is displayed in orange and over the 97.5th percentile is displayed in red. C) Distribution of F_{ST} values for each CYP450 gene calculated on 4 super-populations (AFR, EUR, EAS, SAS). The mean F_{ST} of chromosome 22, the null distribution, is displayed with the grey horizontal line.

populations. Figure 4.1C) shows the distribution of F_{ST} values for each *CYP450* gene calculated on 4 super-populations (AFR, EUR, EAS, SAS). CYP4F genes are scattered across the *CYP450* spectrum, with *CYP4F12* having the second highest mean F_{ST} while *CYP4F8* is in the bottom half of the distribution. Mean F_{ST} of genes of the *CYP3A* subfamily are in the highest values, meaning that these genes have a high divergence between population's genotype frequencies. This could indicate that the low Tajima's D in *CYP3A* reflects positive rather than extreme purifying selection.

4.2. Positive selection in CYP3A and CYP4F subfamilies

The global neutrality and differentiation analyses of *CYP450* genes suggest that positive selection, either directional (*CYP3A*) or balancing (*CYP4F*), may be acting on subfamilies of *CYP450* genes, possibly in a concerted fashion. To further validate positive selection signatures and identify specific putative sites, we used the integrated haplotype score (iHS), which leverages linkage disequilibrium (LD) patterns in a specific population [153]. Typically, an absolute value of iHS greater than 2 at a SNP suggests that the region around the SNP is under selection [153].

In the *CYP3A* cluster, significant iHS values are detected (Figure 4.2A), but signals of positive selection differ between populations. Many signals are detectable in Africans, in East Asians and in Europeans, while fewer signals are detectable in South Asians. Signals of positive selection are noticeable in *CYP3A5*, *CYP3A51P*, *CYP3A4* and *CYP3A43* among Africans. In particular, iHS values in *CYP3A5* are consistently below -2, indicating that the derived alleles have quickly increased in frequency, a signature of positive selection. Interestingly, unlike populations of European descent where the *CYP3A4* gene is typically the most expressed, the *CYP3A5* gene is the most expressed in the African individuals [20, 21]. Among East Asians, the selective sweep is located from *CYP3A51P* to *CYP3A4*, and among South Asians, in *CYP3A43*. Lastly, for Europeans, signals of positive selection are detectable in the region between *CYP3A7* and *CYP3A4*, a signal also present in the East Asian population. *CYP3A43* is the only gene with signals in all super-populations. With these results, we now have multiple lines of evidence pointing towards positive selection, corroborating and extending observations from previous studies [18, 151, 153, 443].

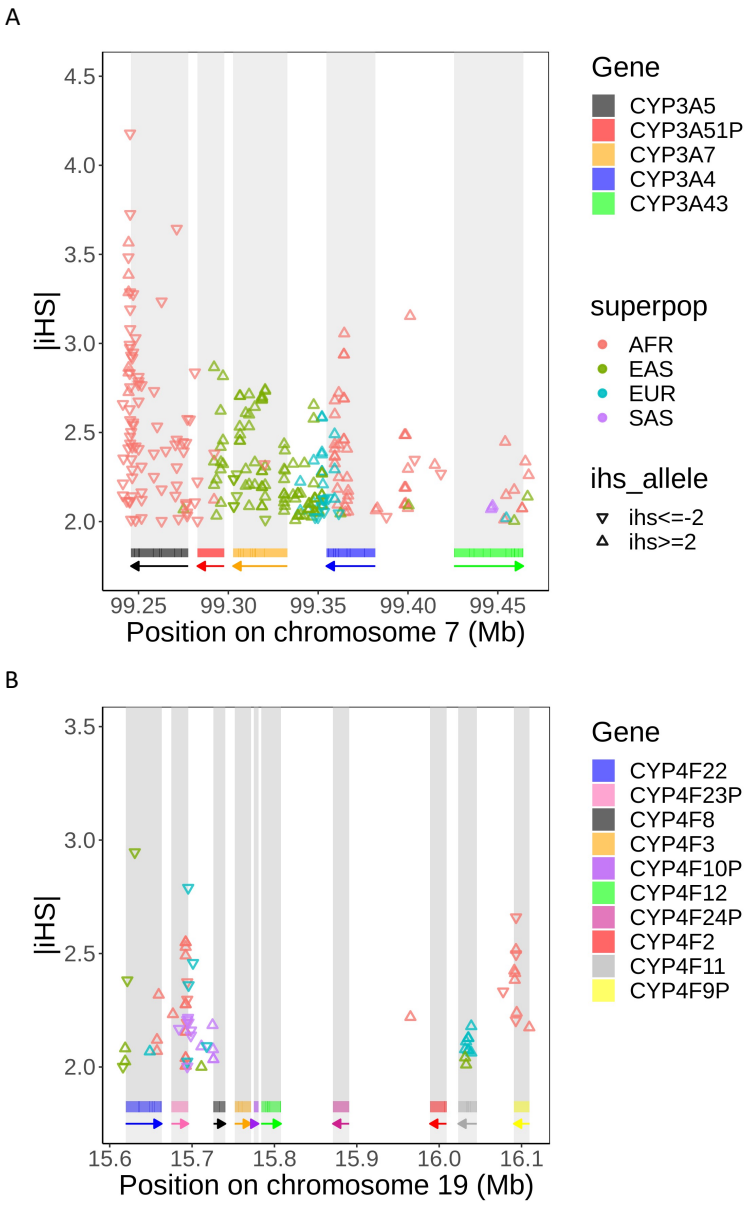

Fig. 4.2. Distribution of SNPs with high $|iHS|$ values ($|iHS| \geq 2$) in the A) CYP3A and B) CYP4F cluster.
A triangle standing on its base means an iHS value ≥ 2, indicating that the ancestral allele has increased in frequency, and a triangle standing on its point means an iHS value $\leq$ -2, indicating that the positive selection is acting on the derived allele. SNPs located in repetitive elements and sequences are masked. Rectangles below the plot show the position of each gene and arrows indicate on which strand the gene is located.

Positive selective pressure is also detected in the *CYP4F* cluster, but on a smaller scale. For the *CYP4F* cluster, signals of positive selection are visible in *CYP4F22*, *CYP4F23P*, *CYP4F11* and *CYP4F9P* (Figure 4.2B). The region between the pseudogene *CYP4F23P*

and the gene *CYP4F8* also shows high iHS values, indicating positive selection in every super-population. iHS values greater than 2 are present in *CYP4F11* in Europeans and East Asians, indicating positive selection acting on ancestral alleles. *CYP4F9P* has significant iHS values in Africans. Again, most iHS values are greater than 2, indicating selective pressures on ancestral alleles, but the 3 strongest signals are seen for derived alleles (iHS below -2), suggesting these SNPs may be driving the signal.

4.3. Balancing selection in CYP3A and CYP4F subfamilies

The Tajima's D analyses (Figure 4.1) suggested balancing selection in the *CYP4F* cluster. To confirm this finding, we used the Beta score [430], a statistic which detects clusters of alleles with similar allele frequencies, developed to specifically test whether balancing selection is present at specific loci.

We considered β score in the top 1% of the whole chromosome as significant β scores (empirical p-value < 0.01), which can vary between populations. In contrast to iHS, very few significant β score values are seen in the *CYP3A* cluster. Only one SNP in *CYP3A43* meets this criteria in Africans, in the same region where balancing selection was also identified in a previous study [444]. The same signal can been seen in the other populations, but it is weaker and do not pass our 1% threshold (Figure 4.3 A). Overall, these results show no clear evidence of balancing selection acting on the *CYP3A* cluster. In line with Tajima's D results, clearer signals are seen in the *CYP4F* cluster, which show larger β scores compared to the *CYP3A* cluster: the highest β score in the *CYP4F* cluster is almost twice as high as the highest *CYP3A*'s β score. SNPs in *CYP4F12* show highly significant β scores, replicated among Africans, Europeans and South Asians, but not in the East Asians. Also, the region between *CYP4F23P* and *CYP4F8* has the most extreme β score in the region, and the signal is visible in all super-populations (Figure 4.3 B). These consistent signals across populations provide convincing evidence of balancing selection acting around *CYP4F8* and *CYP4F12*. Weaker signals, which do not pass our significance threshold but are seen consistently between populations, are seen in *CYP4F23P* and *CYP4F11*. Taken together, these results demonstrate evidence supporting the presence of balancing selective pressures in the *CYP4F* cluster, but show a lack of evidence for balancing selection across the *CYP3A* cluster.

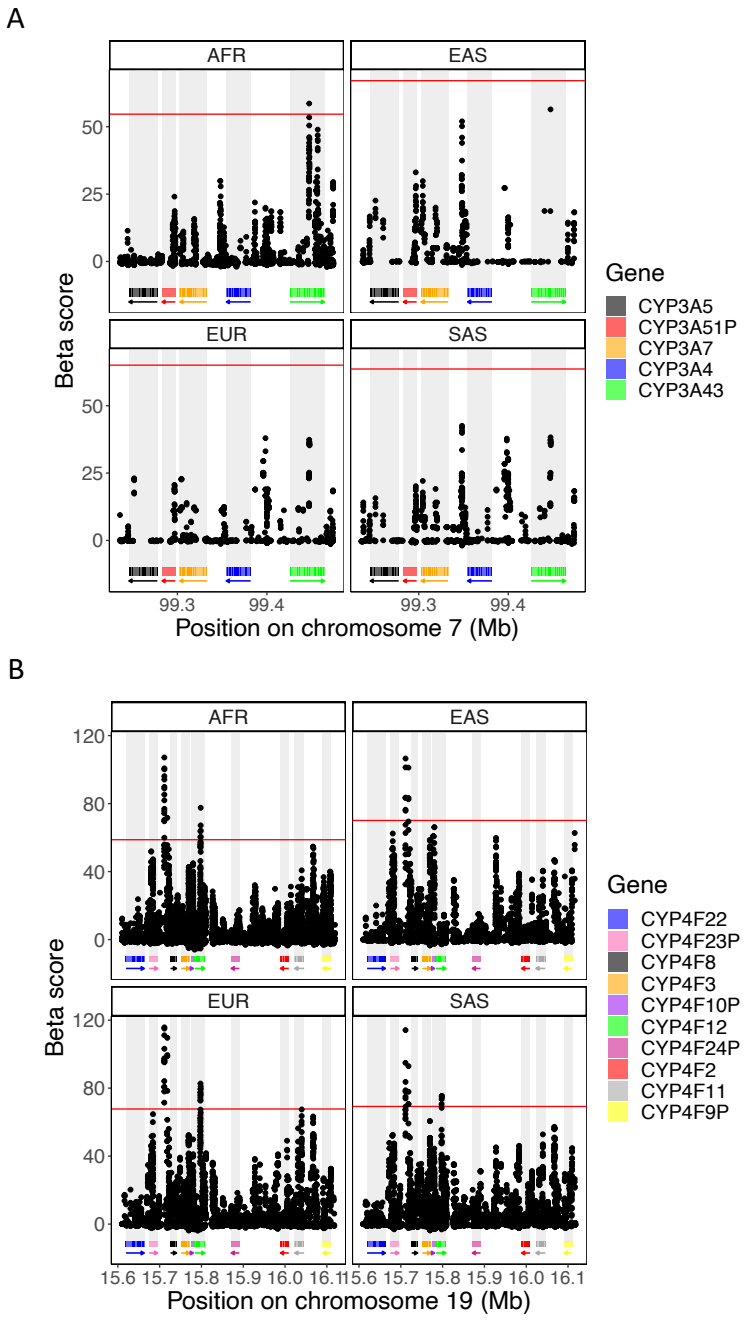

Fig. 4.3. β score in the chromosomal region of the A) CYP3A and B) CYP4F cluster for the 4 super-populations analyzed.
The β score was calculated on the 1000G dataset and the 99th percentile indicating the top 1% β score is displayed by the horizontal line in red. Rectangles below the plot show the position of each gene and arrows indicate on which strand the gene is located.

4.4. Detection of Unusual Linkage Disequilibrium

Since *CYP3A* and *CYP4F* genes are in a gene cluster and selective pressures are acting on these genes, co-evolution could be occurring. Indeed, the different combinations of alleles which co-occured during evolution can lead to concerted selective pressure, or co-evolution, depending on the resulting fitness of the individuals [426]. Such co-evolution signals can be revealed by analyzing patterns of linkage disequilibrium (LD) beyond local associations due to allelic proximity, in order to detect whether specific combinations of alleles (or genotypes) at two distinct loci are particularly overrepresented. To do so, we calculated the genotyped-based LD (r^2) between each pair of SNPs with minor allele frequency (MAF) above 0.05 in the two CYP450 clusters, across each 1000G subpopulation (Methods). Under neutrality, the LD association between SNPs is expected to decrease as genetic distance between the SNPs increases, allowing us to build an empirical distribution by considering clusters of genes of similar size genome-wide (Methods) to the clusters under investigation. Pairs of SNPs showing unusual LD (uLD) values, lying outside of this null distribution, are therefore likely transmitted together more often than expected, making it possible to identify candidate sites that are co-evolving.

In both clusters, strong signals of uLD are present (Figure 4.4, Figure 4.9) compared to matched gene clusters (Methods), with *CYP4F* showing much more extreme signals than *CYP3A* (8.1% vs 4.7% of pairs of SNPs in uLD), despite genetic distances in the *CYP4F* cluster being four times larger than in the *CYP3A* cluster (maximum distance of 0.60 cM vs 0.15 cM, respectively), whereas the physical size of the cluster is only twice (500 Kb vs 250 Kb, respectively). Significant uLD between *CYP3A5* and *CYP3A43* and between *CYP3A7* and *CYP3A43* can be seen in all European populations (Figure 4.9A). *CYP3A5* and *CYP3A43* are the opposite to each other in term of physical location in the cluster while *CYP3A7* and *CYP3A43* are next to each other. Finland (FIN) and Toscani (TSI) populations have the most uLD signals across European populations, with FIN uniquely showing uLD between *CYP3A5*-*CYP3A4*, and TSI showing uLD between *CYP3A4* and *CYP3A43*, a signal consistently seen in the East Asians. TSI also have the highest genetic distance interval in this region, likely due to a larger, more widespread, recombination rate in *CYP3A4* compared to other populations (Figure 4.10). Among East Asians, uLD signals are seen almost exclusively between SNPs in *CYP3A4* and *CYP3A43*, two genes that are

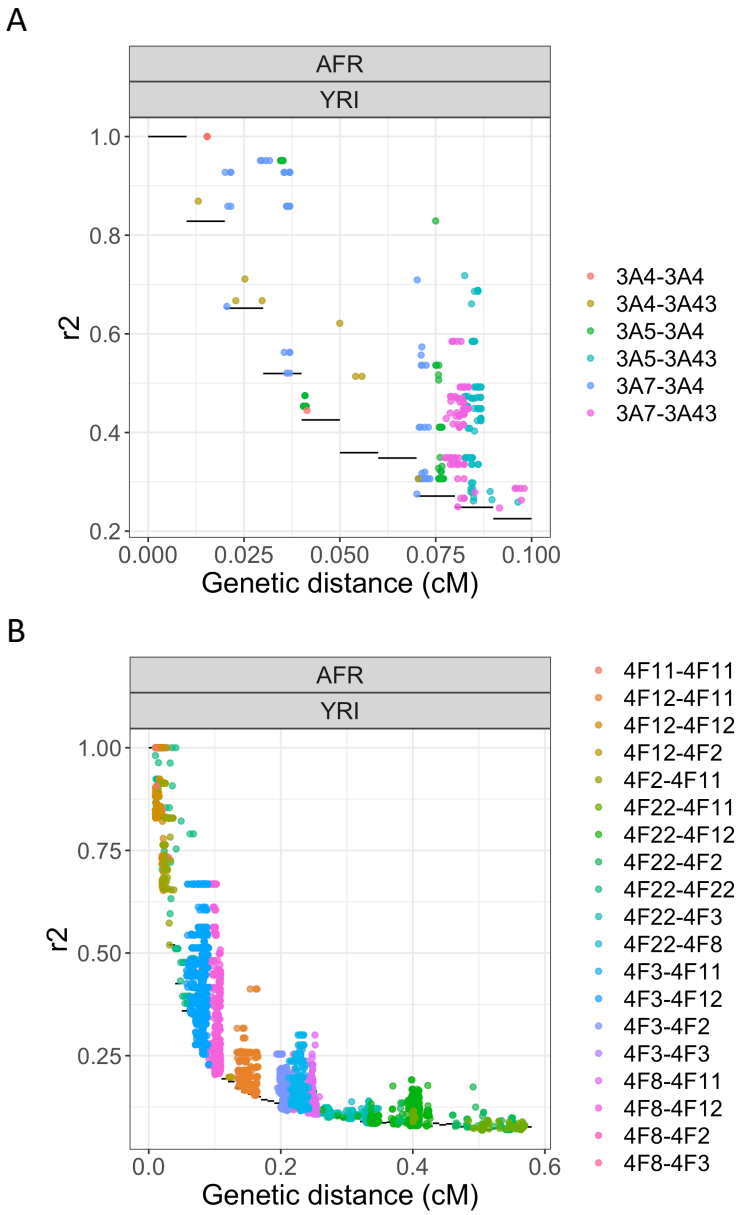

Fig. 4.4. r^2 values between each pairs of SNPs in the A) CYP3A and B) CYP4F cluster in the Yoruba (YRI, AFR) population.
The distance between the SNPs is in centimorgan (cM). Only r^2 values over the null distribution are shown. The null distribution is shown with black horizontal lines. Dots are colored according to which genes are involved in the pair.

next to each other, with no clear recombination hotspot separating them, meaning that linkage disequilibrium can be expected (Figure 4.10). SNPs in these genes are also in uLD in Gujarati Indian (GIH) population, but none of the other South Asian populations show any signal, which may be explained by the short genetic distances within this cluster in this super-population (SAS) (<0.05 cM). Finally, African populations show the most deviation

from the null (Figure 4.4A). SNPs in *CYP3A4* are in uLD with all other genes and the signal also replicates the observations from the European populations, with SNPs in *CYP3A43* in uLD with SNPs in all other genes (Figure 4.4A, Figure 4.9A).

In the *CYP4F* cluster, several pairs of SNPs have patterns of LD that deviate significantly from the empirical distribution (Figure 4.9B). There is uLD for *CYP4F22-CYP4F11* and *CYP4F22-CYP4F12* in almost every populations, even though these genes are far from each other (0.36 Mb and 0.12 Mb, respectively). *CYP4F22* and *CYP4F2* are also in uLD in AFR, EUR and EAS.

The African populations have more evidence of uLD than the other super-populations. One population in particular, the Yoruba (YRI) population, has even more extreme signals in comparison with other African populations and most uLD signal are driven by associations involving the *CYP4F12* gene (Figure 4.4B). Thus, we investigated whether a specific region in *CYP4F12* is in strong LD with the other genes. Indeed, in the YRI population, there is evidence of uLD between a region in *CYP4F12* (at 15.79 - 18.00 Mb on chromosome 19) and the *CYP4F3* (Figure 4.11A) and *CYP4F8* genes (Figure 4.11B). The extreme signals in this gene cluster are in line with the hypothesis that balancing selection acts via gene-gene interactions, or epistasis [445]. As these patterns could be due to sequencing errors [446], we used the latest 1000G dataset which has high-coverage sequencing and is aligned on GRCh38 (Methods). These results were replicated in this second dataset, greatly reducing the possibility that the observed signal is due to sequencing errors or spurious mapping. Finally, in the Europeans, the FIN population has a specific pattern between *CYP4F12* and *CYP4F2*, *CYP4F8*, *CYP4F3*. Looking more closely, many SNPs in *CYP4F12* are in uLD with one SNP in *CYP4F3* (Figure 4.11A) and two SNPs in *CYP4F8* (Figure 4.11B). No specific SNPs are in uLD with *CYP4F2*.

4.5. Detection of eQTLs

We next evaluated the effects of the SNPs identified as being under positive and balancing selection on the expression of the genes in each *CYP450* cluster to test if these are eQTLs.

In the *CYP3A* cluster, three SNPs are under positive selection in the Punjabi population from South Asia (PJL): rs487813, rs679320 and rs568859. These SNPs are located in *CYP3A43* and are significant eQTLs of *CYP3A5* in lung (Figure 4.5A). The SNP under

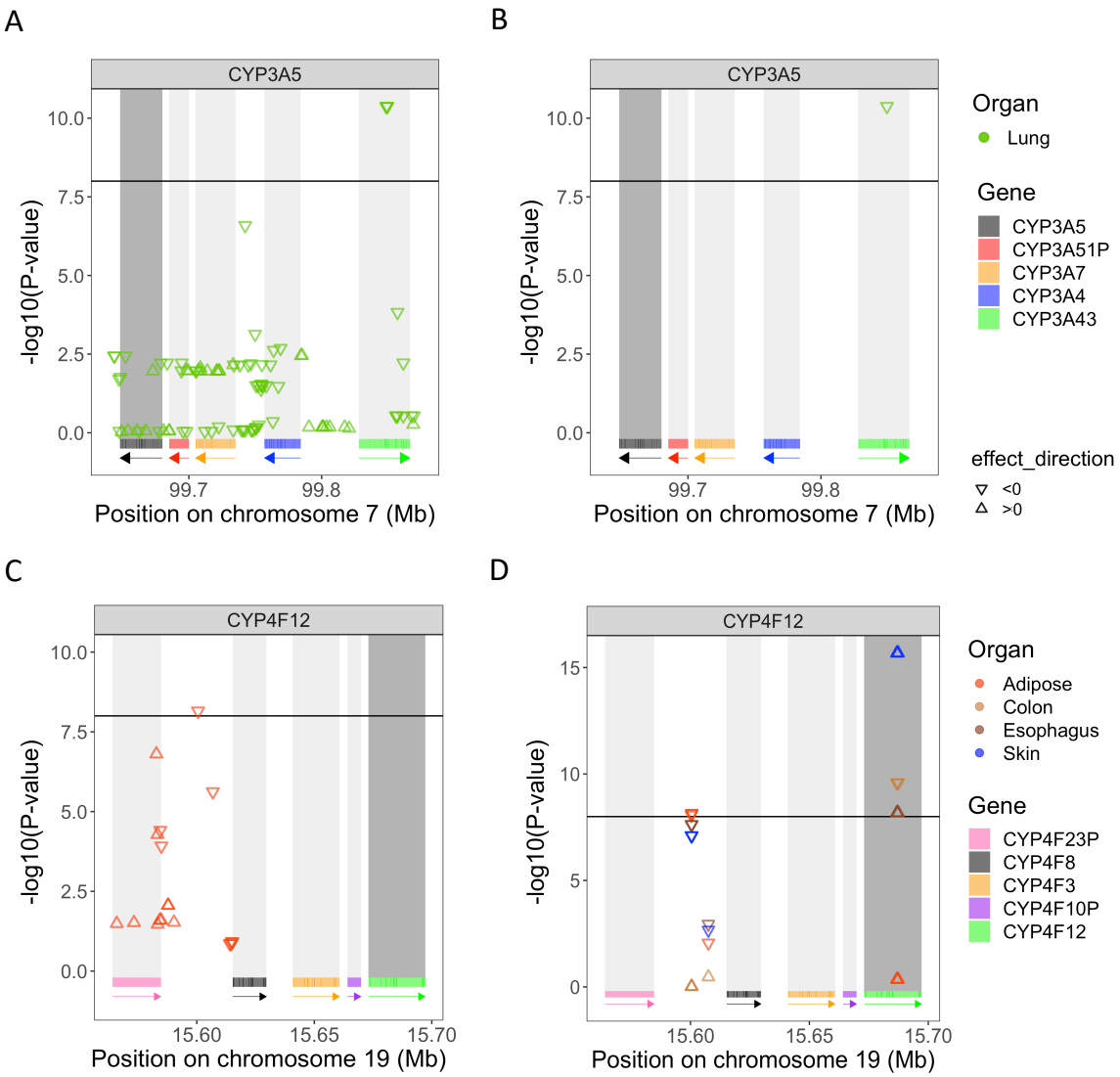

Fig. 4.5. P-values of the associations between SNPs under A) positive selection and B) balancing selection and CYP3A5's gene expression in lung and p-values associated with SNPs C) under positive selection and D) balancing selection and tissue-specific gene expression of CYP4F12.
CYP3A5 and CYP4F12 are shown in dark gray, as the expressions of these genes are tested. The triangle standing on its base indicates a positive effect size ($\beta_{eQTL} > 0$), while a triangle standing on its point indicates a negative effect size ($\beta_{eQTL} < 0$). The threshold, set to 10^{-8}, is represented by the horizontal black line, meaning that a $-log10(p-value) > 8$ is a significant eQTL. Only tissues with significant eQTLS are displayed. As before, rectangles below each plot show the position of each gene and arrows indicate on which strand the gene is located. Each gene has its own colour to indicate its location.

balancing selection in the Luhya population (LWK) in *CYP3A43*, rs800667, is also an eQTL of *CYP3A5* in lung (Figure 4.5B). The effect size estimate for these significant eQTL is negative, indicating a reduction in *CYP3A5* gene expression with each non-reference allele. This locus in *CYP3A43* thus impact *CYP3A5* expression in lung, even though *CYP3A5* and *CYP3A43* are at opposite ends of the cluster, 147.99 Kb apart. According to the ReMap density database [438], this locus also displays regulatory signals, supporting the importance of this region at the transcriptional regulatory level. This result is in line with the LD analyses (Figure 4.4A), which suggested uLD between SNPs in *CYP3A5* and *CYP3A43* in Europeans, Africans and the Japanese. Those four SNPs were all in uLD with 11 SNPs in the Toscani population (TSI) and five other SNPs in Americans of African Ancestry (ASW).

In the *CYP4F* cluster, a SNP under positive selection, rs74459786 (Table 4.1), located in the intergenic region between *CYP4F23P* and *CYP4F8*, is an eQTL of *CYP4F12* in adipose tissue (Figure 4.5C), with a negative effect size. SNPs under balancing selection (Table 4.2) within *CYP4F12* are eQTLs for *CYP4F12* expression in the colon, esophagus and skin, but interestingly, their effects in these tissues are in opposite directions, with positive effect sizes in the colon and skin, and negative ones for the esophagus. Furthermore, a SNP with a balancing selection signal is also an eQTL of *CYP4F12* expression in adipose-subcutaneous tissue (Figure 4.5D) with a negative effect size estimate. It lies in the intergenic region between *CYP4F23P* and *CYP4F8*, which is the same region as the SNP under positive selection (rs74459786) in Figure 4.5C.

Another SNP under positive selection in this intergenic region, rs62115147 (Table 4.1), is also associated with *CYP4F3* expression in one of the brain tissues (Brain-Spinalcord-cervicalc-1) and in nerve tissue (Figure 4.12A). The *CYP4F12* gene emerged repeatedly as a candidate in our balancing selection and uLD analyses, while the intergenic region between *CYP4F23P* and *CYP4F8* is seen only in the balancing selection analysis.

Even if less positive selection is present in the *CYP4F* cluster compared to the *CYP3A* cluster, many of the SNPs showing high iHS values in the *CYP4F* cluster show up as eQTLs for different genes. SNPs under positive selection located in *CYP4F11* (Table 4.1) are eQTLs of *CYP4F2* in brain and skin tissues (Figure 4.12B) with consistent, negative effect sizes. Additionally, the same SNPs under positive selection within *CYP4F11* are associated with

expression of *CYP4F11* itself in multiple tissues (Figure 4.12C). The direction of effect on gene expression is the same for all significant associations.

4.6. Phenotypic associations

Using the UK Biobank cohort (UKb), we did a Phenome-Wide Association Study (PheWAS) to identify phenotypes potentially under selective pressure (Methods), using the available variants with selective signals in the *CYP4F* genes (166 variants from the 180 found under selection) and in the *CYP3A* genes (62 from the 125 variants found under selection). No significant associations were found for SNPs under selective pressure in *CYP4F* cluster.

In the CYP3A cluster, however, SNPs under positive selection in at least one studied populations were found associated with six phenotypes (Figure 4.6A,B) in our PheWAS. Among the disease phenotypes, we found association with pelvic inflammatory disease (PID), which is female-specific, and for which the SNP with the strongest association (p-value$_{rs2014764} = 1.96x10^{-5}$) was under positive selection in European (CEU, GBR) and East Asian (CHB, CHS, CDX) populations. Among the continuous phenotypes investigated (Methods), we found association with pulse rate, for which the SNP with the strongest association (p-value$_{rs12536946} = 4.66x10^{-13}$) was also found under selective pressure in Europeans (CEU). Among the biomarker variables, the strongest associations with platelet count (p-value$_{rs503115} = 2.30x10^{-12}$) and erythrocyte count (p-value$_{rs10235630} = 1.83x10^{-7}$) were both found with SNPs under selective pressure in the Japanese population.

Lastly, for both high light scatter reticulocyte count and reticulocyte count, their strongest association (p-value$_{rs73713580} = 1.24x10^{-17}$; p-value$_{rs55830753} = 3.08x10^{-8}$ respectively) were both found under selective pressure in African population (MSL and ACB, respectively). Using Mendelian randomisation (Methods), we evaluated the causal relationship between *CYP3A5* expression in lung, for which eQTLs were found under selective pressure above, and the phenotypes found to be associated with SNPs in the *CYP3A* cluster. We identified a significant causal association between CYP3A5 expression and both high light scatter reticulocyte count (p-value$_{IVW} = 2.74x10^{-4}$) and reticulocyte count (p-value$_{IVW} = 1.66x10^{-4}$). We did not detect pleiotropy using MR-Egger and results were robust using the weighted median test (Methods).

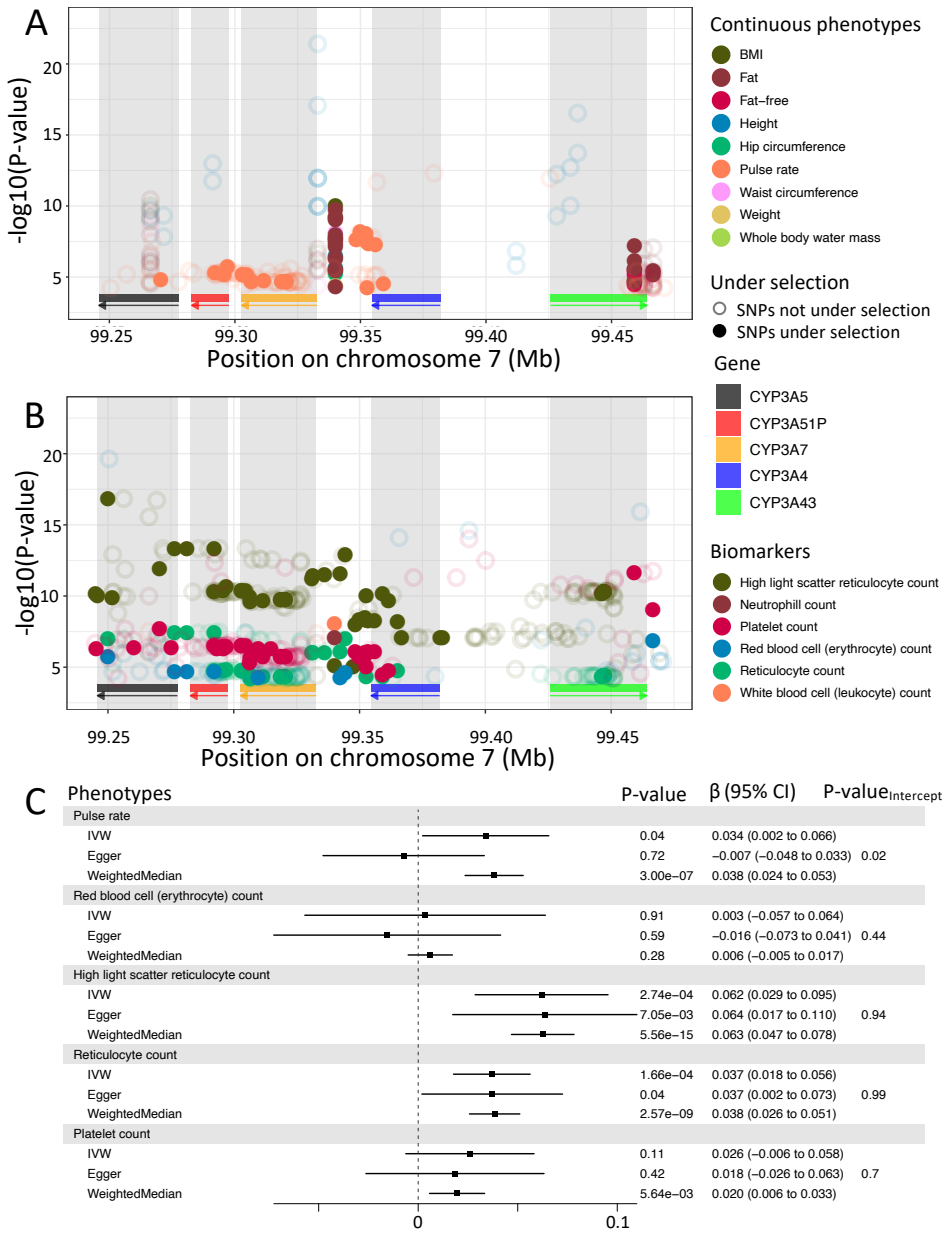

Fig. 4.6. Associations of CYP3A cluster with phenotypes in the UK biobank. Significant associations ($p<0.05/777$) for continuous traits (A) and plasmatic biomarkers (B) which are significant in at least one SNP under selection. SNPs under selection are represented as full dots, meanwhile other SNPs are represented as empty dots. As before, rectangles below each plot show the position of each gene and arrows indicate on which strand the gene is located. Each gene has its own colour to indicate its location. C) Causal relationship with CYP3A5 expression in lung for phenotypes showing significant association with its eQTLs. β represents the change of 1 standard deviation of CYP3A5 expressions on phenotypes, also in standard deviation units. P-value of three statistics (IVW, Egger, Weighted Median) are displayed with the β and the 95% confidence interval (CI) of the association for each phenotype in the grey box. For Egger, the p-value of the intercept is also displayed.

Altogether, these results indicate that the selective pressure in the *CYP3A* cluster could be driven by the production of reticulocyte through the expression levels of *CYP3A5*, and also suggest that pulse rate could be impacted by genetic variation in *CYP3A* genes.

Among other associations identified in this cluster, three SNPs showed strong associations with anthropometric traits (Figure 4.6A) and are under selective pressure in European population (CEU, IBS). Those SNPs were, however, found to be associated with expression of genes outside the *CYP3A* cluster (Supplementary text 7.3), prompting for further investigation of the relationship between this cluster and other neighbouring genes to understand the different drivers at play.

5. Discussion

Drug metabolism is a rather complex system with the *CYP450* genes metabolizing aroung 75% of common drugs. As shown by others [18, 151, 443, 447], we also found that selective pressure and genetic differentiation between populations were present in *CYP450* genes. Here, we provide a deeper analysis of two *CYP450* clusters, the widely studied CYP3A [21, 443, 447][448] and the less well-known *CYP4F* clusters, identified thanks to their outlier patterns in neutrality and population differentiation analyses. These two *CYP450* clusters exhibit multiple selective signatures (positive selection and balancing selection) and show population differentiation. We found that natural selection forces involved differ between the two clusters; the *CYP3A* cluster is evolving under positive selection, while the *CYP4F* cluster show signals of balancing selection. Furthermore, the *CYP4F* cluster shows strong evidence for co-evolution and co-regulation signals.

In the literature, the *CYP450* genes are often studied independently. In our study, we considered the evolution of the entire family cluster, mostly *CYP3A* and *CYP4F* genes, and detected signatures of coevolution between the paralogous genes, suggestive of potential epistatic interactions. As these clusters of genes are involved in drug metabolism [405, 406, 449, 450], it is important to understand the impact of genetic variants on their gene expression, to help understand how these variants might impact drug response and refine disease treatments in a personalized way. Our results also show that the impact of specific variants may differ between populations, which could lead to a deeper understanding of differences in individual drug response [451, 452].

The *CYP3A* cluster contains 4 genes: *CYP3A4*, *CYP3A5*, *CYP3A7* and *CYP3A43*. Signals of positive selection were detected in the *CYP3A* cluster, specifically in *CYP3A4* and *CYP3A7*, which have been under recent positive selection in African, European and the Chinese populations, while *CYP3A5* appears under positive selection in Europeans and *CYP3A43* in non-Africans [447]. Our analyses confirmed that *CYP3A* genes are evolving under positive selection as previously reported [18, 151, 153, 443].

We found that the locus known to cause non-expression of *CYP3A5* [20], rs10264272/ *CYP3A5**6, is under positive selection ($|iHS| \geq 2$) in African populations (YRI, GWD, LWK). A second locus, known to cause low *CYP3A5* expression, rs776746/ *CYP3A5**3, is under positive selection ($|iHS| \geq 2$) in two African populations (YRI, GWD). These derived allele have thus swept up in frequency in several African populations. In the Toscani population, rs776746/*CYP3A5**3 is found to be in uLD with the four SNPs under selective pressure in the *CYP3A* cluster, that are eQTLs of *CYP3A5* in lung.

CYP3A43 is the ancestor gene of this cluster [18, 19], however, its function is not well understood, unlike other *CYP3A* genes. Our analyses suggest that SNPs in *CYP3A43* regulate *CYP3A5* gene expression, at least in lung. Levels of expression of *CYP3A5* in lung were causally associated to reticylocytes count and many of its eQTLs were under selection in Africans. Since *Plasmodium vivax*, a parasite causing malaria, affect mainly young reticulocytes [453] and that malaria is present in Africa, the selective pressure found in this population could be associated to this disease. Further studies need to be done to validate this hypothesis.

In the *CYP4F* cluster, we found both positive and balancing selection pressures acting. Furthermore, the SNPs evolving under selective pressures are associated with gene expression levels across the cluster in several tissues. For instance, a cis-eQTL of *CYP4F12*, rs74459786, is detected to be under positive selection in the Kinh population in East Asia (KHV). We also found that several SNPs in *CYP4F11* are associated with *CYP4F2* expression. Both genes are implicated in common metabolic function, such as the synthesis of 20-HETE (20-hydroxyeicosatetraenoic acid) from arachidonic acid [454]. Thus, this could indicate a possible regulatory mechanism of common functions.

Finally, the region between *CYP4F23P* and *CYP4F8* emerged multiple times in our analyses. This intergenic region shows strong signals for selection, with the same SNPs also

being eQTLs of *CYP4F3* (nerve) and *CYP4F12* (adipose tissue). Given the implication of *CYP4F12* in fatty acid metabolism [455], our results may point towards the identification of new regulatory elements involved in this process in adipose tissues.

A potential limitation in the current study is that population genetic statistics can be biased in the presence of fine-scale population structure. However, to mitigate this issue, we performed our analyses not only at the broader population level but also within individual subpopulations, ensuring that the values obtained from subpopulations were consistent with those from the overall superpopulation.

An important limitation to consider is the methodology used for calculating Tajima's D using the `vcftools` software. This tool's current implementation does not account for mappability and callability in whole genome sequencing data. This approach introduces a bias by implicitly considering uncalled positions as non-variable, leading to an underestimation of diversity measures [456]. Although our strategy to exclude regions with high proportions of missing data likely minimizes this bias, we acknowledge that approaches that directly incorporate genomic accessibility considerations for a more precise estimation of genetic diversity should be used in future studies of CYP genes. While the impact on our results appears minimal, as evidenced by the high similarity of results after the removal of high-missing data regions, the potential for underestimation of diversity estimators should be considered when comparing these estimates across genomic regions.

In conclusion, our results demonstrate high heterogeneity across human populations, both in terms of selective signals and interaction between variants and expression levels, for the *CYP3A* and *CYP4F* genes. There could thus be important differences in metabolic regulation impacting drug response in individuals from different ethnicities. In particular, these variants could cause impaired efficacy, as well as side effects. As pharmacogenetic studies still typically focus on European populations, our results underline the importance of including individuals from several populations in order to capture all of the genetic diversity and its impact on disease treatment and metabolism.

6. Supplementary figures

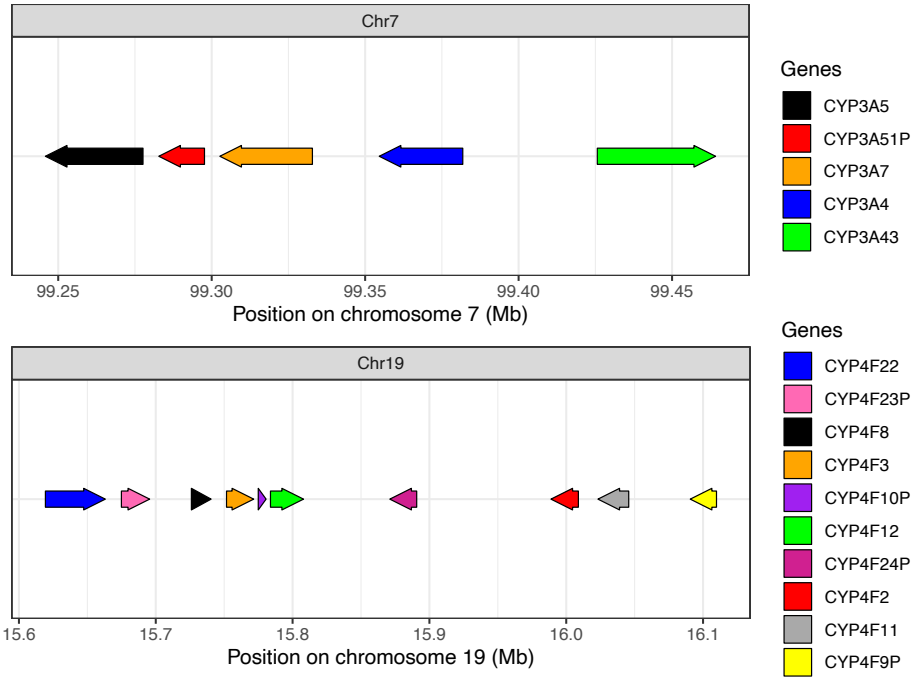

Fig. 4.7. Positions of genes in the cluster of *CYP3A* (top) and *CYP4F* (bottom) in GRCh37. The direction of the arrow indicate the direction of the gene.

Fig. 4.8. Distribution of Tajima's D values computed on intervals of 1 Kb for each CYP450 gene across each subpopulation in the European population
The 2.5th percentile is marked by the orange vertical line and the 97.5th percentile is marked by the red vertical line, representing the significance threshold. We kept Tajima's D value for intervals with a high level of mappability (>90% of interval overlaps high accessibility uniquely mapped regions). (CEU : Utah residents with Northern and Western European ancestry; TSI : Toscani in Italia; IBS : Iberian populations in Spain; FIN : Finnish in Finland; GBR : British in England and Scotland).

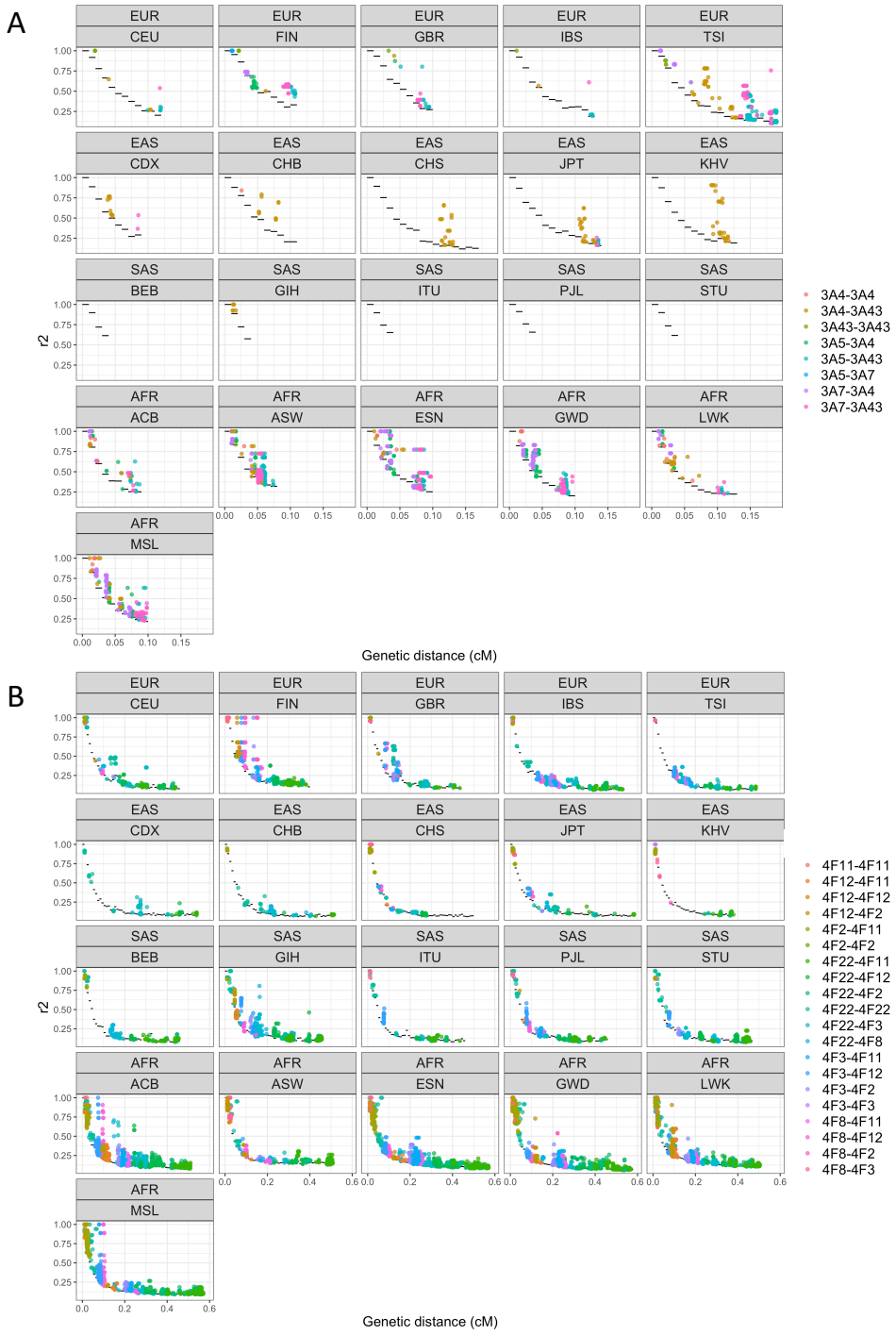

Fig. 4.9. r^2 values between each pairs of SNPs in the A) CYP3A and B) CYP4F cluster for each 1000G population, except YRI (AFR).

The genetic distance between the SNPs is in centimorgan (cM). Only r^2 values over the empirical threshold are shown. The empirical distribution is shown with black horizontal lines. Dots are colored according to which genes are involved in the pair.

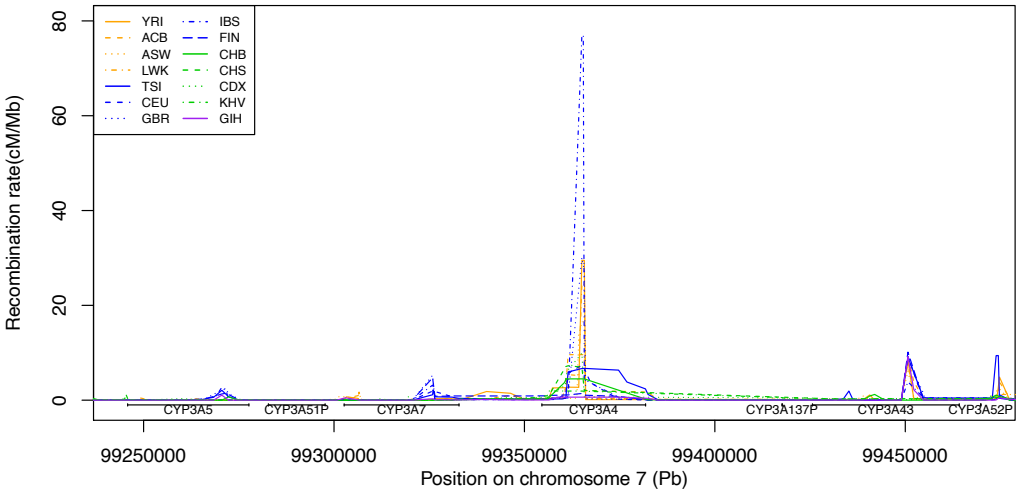

Fig. 4.10. Recombination map in the CYP3A gene cluster.
Each line, with a different line pattern, represents a population and is colored according to the super-population. Each gene and pseudogene are shown below the plot with horizontal line.

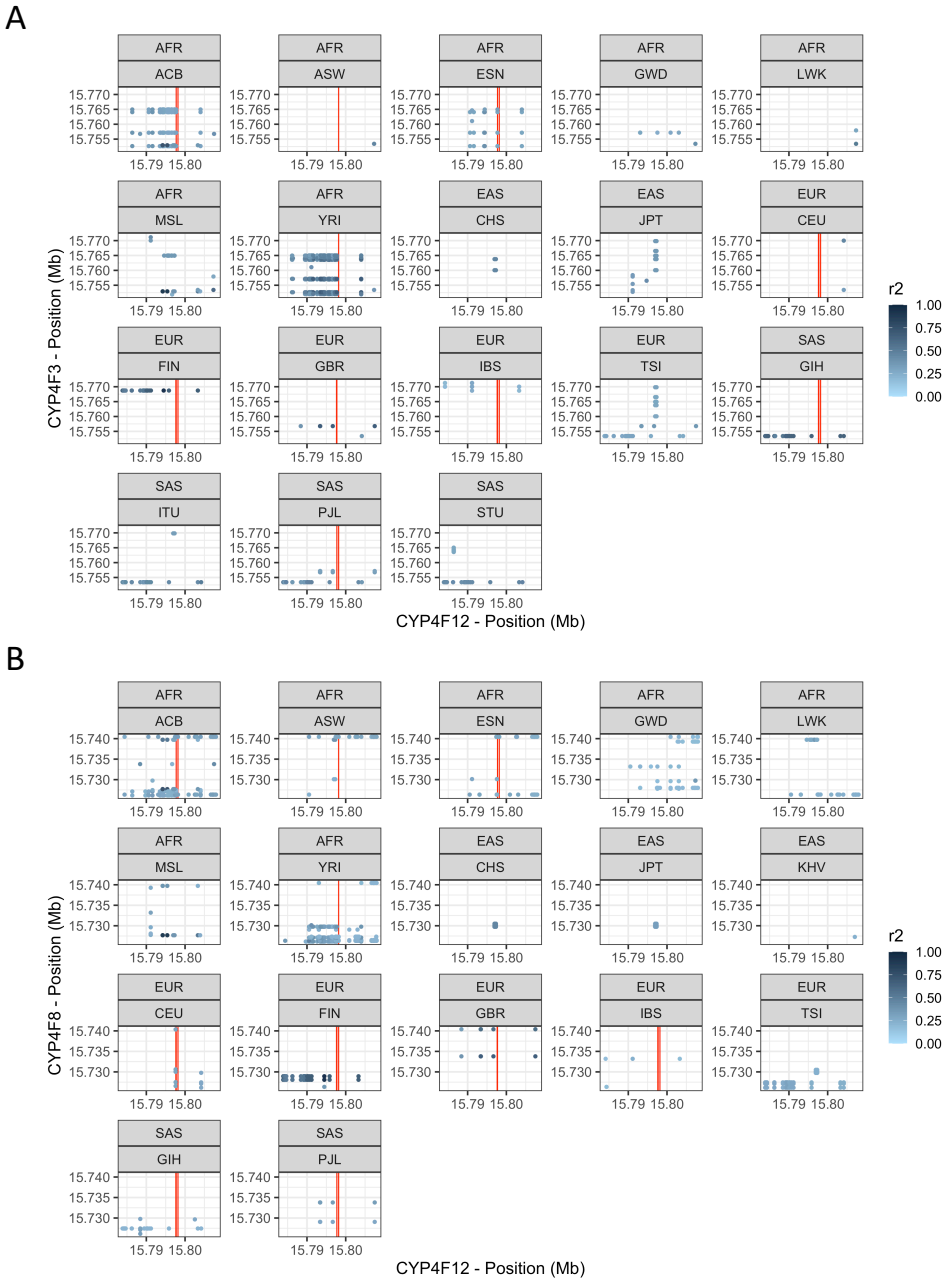

Fig. 4.11. Coordinates of each SNP that is in a pair of SNPs with r^2 values in the extremes of the empirical distribution for each subpopulation of 1000G.
The displayed SNPs pairs have one SNP in CYP4F12 and the other is in A) CYP4F3 and in B) CYP4F8. We took r^2 values from the previous analysis and filtered to keep only values where one SNP was located in CYP4F12. The graph is generated using the *ggplot2* library in R. The physical coordinates of each significant Beta signal, identified in the balancing selection analysis, are shown by the vertical red lines, which were created using *geom_vline*. Points were colored according to their respective r^2 values with *scale_color_gradient*.

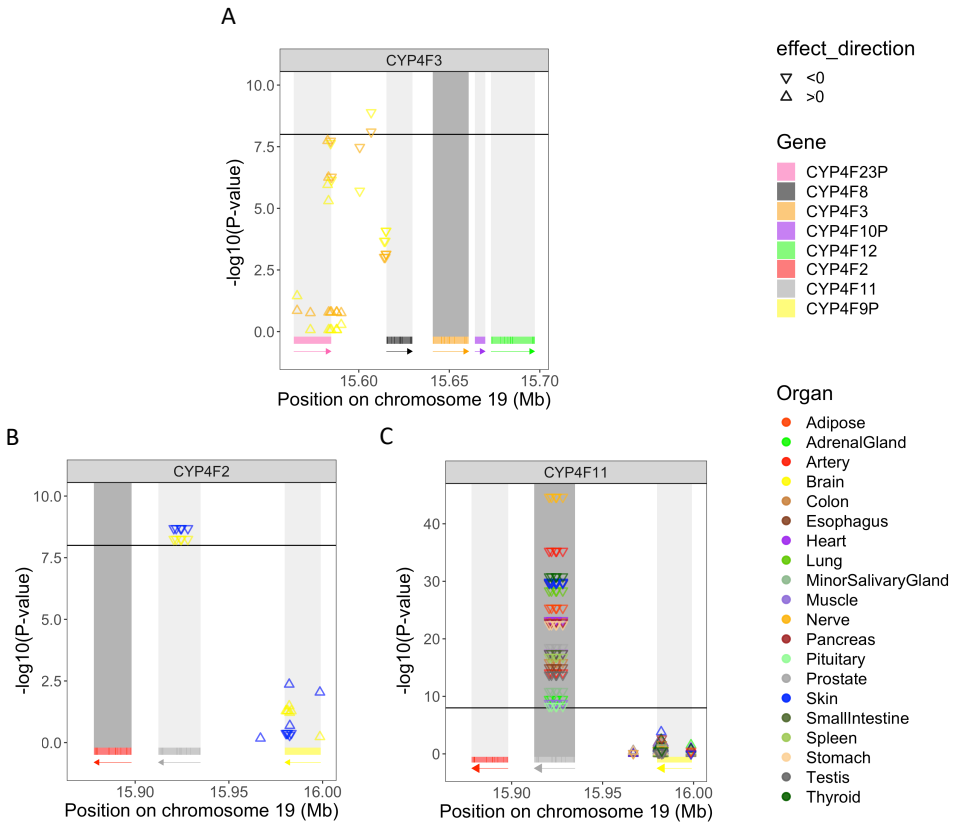

Fig. 4.12. P-values associated with SNPs under positive selection ($|iHS| \geq 2$) explaining variation of gene expression of A) CYP4F3 B) CYP4F2 and C) CYP4F11.
The tested gene is shown in dark gray and the effect size is represented either by a triangle standing on its base or a triangle standing on its point. The threshold, set to 10^{-8}, is represented by the horizontal black line, meaning that a $-log10(p-value) > 8$ is a significant eQTL.

Variant identifier	iHS	Population	Super-population	eQTL
rs74459786	2.00062	JPT	EAS	CYP4F12
	2.09063	STU	SAS	
rs62115147	-2.09205	IBS	EUR	CYP4F3
rs2365175	2.07818	TSI	EUR	CYP4F2, CYP4F11
	2.04181	KHV	EAS	
rs11086013	2.11270	TSI	EUR	CYP4F2, CYP4F11
	2.01056	KHV	EAS	
rs11881793	2.07352	TSI	EUR	CYP4F2, CYP4F11
	2.12697	IBS	EUR	
rs3746154	2.07395	TSI	EUR	CYP4F2, CYP4F11
	2.12768	IBS	EUR	
rs4808413	2.06351	TSI	EUR	CYP4F2, CYP4F11
	2.17967	IBS	EUR	

Table 4.1. SNPs under positive selection in the CYP4F cluster that are also eQTLs. Each significant SNP is reported with its iHS values ($|iHS| \geq 2$), specific population and RS variant identifier. The gene with differential expression is reported in the eQTL column.

Variant identifier	β score	Population	Super-population
rs644584	72.91370	CEU	EUR
	74.40043	FIN	
	72.64120	GBR	
	77.31849	IBS	
	70.56299	GIH	SAS
rs642322	67.17824	ACB	AFR
	64.12365	ASW	
	77.53712	ESN	
	60.42254	YRI	
	80.95814	CEU	EUR
	79.46902	FIN	
	76.20108	IBS	
	74.54502	GIH	SAS
	75.44200	PJL	
rs74459786	75.65607	ACB	AFR
	59.94361	ASW	
	96.13950	ESN	
	107.11148	GWD	
	89.83249	LWK	
	100.15158	MSL	
	107.11148	GWD	
	85.28441	YRI	
	111.07817	CEU	EUR
	95.36774	FIN	
	115.01944	GBR	
	97.56379	IBS	
	106.51889	CDX	EAS
	83.53252	CHS	
	76.40631	KHV	
	83.65641	BEB	SAS

Table 4.2 continued from previous page

Variant identifier	β score	Population	Super-population
	74.91468	ITU	SAS
	114.13019	GIH	
	78.72150	PJL	
rs75814017	70.44644	ACB	AFR
	88.88523	ESN	
	100.74132	GWD	
	81.90479	LWK	
	94.13457	MSL	
	76.91697	YRI	
	103.15354	CEU	EUR
	95.98925	FIN	
	115.75484	GBR	
	95.16955	IBS	
	95.82953	TSI	
	101.29256	CDX	EAS
	75.79028	CHS	
	77.93165	BEB	SAS
	73.87701	PJL	
rs73000014	69.60533	ESN	AFR
	74.28322	GWD	
	84.13158	CEU	EUR
	71.43819	FIN	
	78.00721	GBR	
	80.80257	IBS	
	80.54700	TSI	
	83.32819	CDX	EAS
	94.87801	GIH	
rs16980720	60.32447	ACB	AFR
	63.67802	ESN	

Table 4.2 continued from previous page

Variant identifier	β score	Population	Super-population
rs16980720	77.61884	CEU	EUR
	79.43404	FIN	
	75.75824	GBR	
	82.68951	IBS	
	75.28558	GIH	SAS
	73.43462	PJL	

Table 4.2. SNPs under balancing selection in the CYP4F cluster that are also eQTLs of CYP4F12. Each significant SNP is reported with its β values, specific population and RS variant identifier.

Description	Code field	Processing
Continuous phenotypes suggested by UKb		
Length of working week for main job	767	
Frequency of travelling from home to job workplace	777	
Age completed full time education	845	
Cooked vegetable intake	1289	
Salad / raw vegetable intake	1299	
Fresh fruit intake	1309	
Dried fruit intake	1319	
Bread intake	1438	
Cereal intake	1458	
Tea intake	1488	
Coffee intake	1498	
Water intake	1528	
Age started wearing glasses or contact lenses	2217	
Age high blood pressure diagnosed	2966	
Age diabetes diagnosed	2976	
Age angina diagnosed	3627	
Age hay fever, rhinitis or eczema diagnosed	3761	

Table 4.3 continued from previous page

Description	Code field	Processing
Age asthma diagnosed	3786	
Age heart attack diagnosed	3894	
Age emphysema/chronic bronchitis diagnosed	3992	
Age deep-vein thrombosis (DVT, blood clot in leg) diagnosed	4012	
Age pulmonary embolism (blood clot in lung) diagnosed	4022	
Age stroke diagnosed	4056	
Longest period of depression	4609	
Number of depression episodes	4620	
Age glaucoma diagnosed	4689	
Age cataract diagnosed	4700	
Longest period of unenthusiasm / disinterest	5375	
Number of unenthusiastic/disinterested episodes	5386	
Age when loss of vision due to injury or trauma diagnosed	5430	
Age when diabetes-related eye disease diagnosed	5901	
Age macular degeneration diagnosed	5923	
Age other serious eye condition diagnosed	5945	
Hand grip strength (left)	46	
Hand grip strength (right)	47	
Waist circumference	48	
Hip circumference	49	
Standing height	50	
Heel bone ultrasound T-score, manual entry	77	
Heel bone mineral density (BMD) T-score, automated	78	
Heel bone mineral density (BMD) T-score, automated (left)	4106	

Table 4.3 continued from previous page

Description	Code field	Processing
Heel bone mineral density (BMD) T-score, automated (right)	4125	
Heel bone mineral density (BMD) T-score, manual entry (left)	4138	
Heel bone mineral density (BMD) T-score, manual entry (right)	4143	
Pulse rate	4194	
Sitting height	20015	
Fluid intelligence score	20016	
Birth weight	20022	
Mean time to correctly identify matches	20023	
Cascot confidence score	20121	
Body mass index (BMI)	21001	
Weight	21002	
Body fat percentage	23099	
Whole body fat mass	23100	
Whole body fat-free mass	23101	
Whole body water mass	23102	
Basal metabolic rate	23105	
Impedance of whole body	23106	
Impedance of leg (right)	23107	
Impedance of leg (left)	23108	
Impedance of arm (right)	23109	
Impedance of arm (left)	23110	
Leg fat percentage (right)	23111	
Leg fat mass (right)	23112	
Leg fat-free mass (right)	23113	
Leg predicted mass (right)	23114	

Table 4.3 continued from previous page

Description	Code field	Processing
Leg fat percentage (left)	23115	
Leg fat mass (left)	23116	
Leg fat-free mass (left)	23117	
Leg predicted mass (left)	23118	
Arm fat percentage (right)	23119	
Arm fat mass (right)	23120	
Arm fat-free mass (right)	23121	
Arm predicted mass (right)	23122	
Arm fat percentage (left)	23123	
Arm fat mass (left)	23124	
Arm fat-free mass (left)	23125	
Arm predicted mass (left)	23126	
Trunk fat percentage	23127	
Trunk fat mass	23128	
Trunk fat-free mass	23129	
Trunk predicted mass	23130	
Systolic blood pressure, manual reading	93	Values based on the mean of the instance at the first visit
Diastolic blood pressure, manual reading	94	Values based on the mean of the instance at the first visit
Pulse rate (during blood-pressure measurement)	95	Values based on the mean of the instance at the first visit
Pulse rate, automated reading	102	Values based on the mean of the instance at the first visit

Table 4.3 continued from previous page

Description	Code field	Processing
Forced vital capacity (FVC)	3062	Values based on the mean of the instance at the first visit
Forced expiratory volume in 1-second (FEV1)	3063	Values based on the mean of the instance at the first visit
Peak expiratory flow (PEF)	3064	Values based on the mean of the instance at the first visit
Diastolic blood pressure, automated reading	4079	Values based on the mean of the instance at the first visit
Systolic blood pressure, automated reading	4080	Values based on the mean of the instance at the first visit
Blood cells		
White blood cell (leukocyte) count	30000	
Red blood cell (erythrocyte) count	30010	
Platelet count	30080	
Lymphocyte count	30120	
Monocyte count	30130	
Neutrophill count	30140	
Eosinophill count	30150	
Basophill count	30160	
Nucleated red blood cell count	30170	
Reticulocyte count	30250	
High light scatter reticulocyte count	30300	

Table 4.3 continued from previous page

Description	Code field	Processing

Table 4.3. Continuous phenotypes of the UKb

7. Supplementary text

7.1. D of Tajima additional filtering

To account for the high homology among many CYP450 genes potentially causing problems in callability, we implemented additional filtering steps in our analyses. We used a mappability mask provided by ENCODE (*wgEncodeCrgMapabilityAlign100mer* from UCSC Table Browser) and an accessibility mask (*20140520.combined_ mask.autosomes .bed* accessed here). We recalculated Tajima's D using after filtering the VCF using the mappability mask: specifically, we removed all SNPs for which mappability scores were different from 1 in wgEncodeCrgMapabilityAlign100mer (a score of 1 means unique mapping). This was done for each subpopulation, to ensure that population structure does not drive the results. We then kept only Taijma's D values estimated on 1Kb intervals within accessible genomic regions: the mask excludes regions where depth of coverage across all samples was higher or lower than the average depth by a factor of 2-fold. It also excludes sites where >20% of reads had mapping quality of zero. We required that at least 90% of the 1Kb interval overlaps the accessibility mask. All our results remained consistent with our initial observations (Figure 4.8).

7.2. Pre-processing of GTEx genetic data

Starting from the imputed genotyping dataset, we kept bi-allelic SNPs and removed positions with more than 5% missing genotypes, leaving 100,986 SNPs which were used to perform a PCA using flashPCA2 [362]. To retain the non-admixed individuals of European descent, we reduced the dimensionality of the top 10 PCs using the R package `UMAP` [327] (default parameters) to obtain a two dimensional representation of the genetic information contained within those PCs. We identified the largest homogeneous group (self-reported "white") and excluded outlier groups, used only these individuals for the rest of the analyses. We then reran a PCA on this group. We did all subsequent analyses with these 699 individuals. Next we separated each tissue, then removed tissues with fewer than 50 samples, leaving samples from 50 different tissues. We removed in our analyses genes that had fewer than 6 reads in at least 20% of the samples (as recommended by GTEx). We then normalized expression data using limma (TMM normalization) [457] and voom [458]. We calculated

PEER factors [273] on the normalized expressions. The suggested number of PEER factors for the GTEx tissues is 15 for N $<$ 150, 30 for 150 $\leq$ N $<$ 250, 45 for 250 $\leq$ N $<$ 350, and 60 for N $\geq$ 350 [459].

7.3. Additional analyses on phenotypes

In CYP4F cluster, no SNP under selection was found associated with one of the selected phenotypes. It is possible that the UKb, and more specifically the white British population, is not the appropriate population in which to investigate these selection signals. However, other SNPs in these loci, not found to be under selection in our previous analyses, are associated with forced expiratory volume in 1 second (FEV1) and forced vital capacity (FVC) in the gene *CYP4F2*, and eosinophil count in *CYP4F23P*.

In the *CYP3A* cluster, we identified significant eQTL signals for genes outside the *CYP3A* family, such as *GS1-259H13.2* (ENSG00000244219.6), *ARPC1B* (ENSG00000130429.12) and *ZKSCAN5* (ENSG00000196652.11), located between *CYP3A7* and *CYP3A4* genes and at the 3' end of *CYP3A43* (Figure 4.6A. These eQTLs are also associated with anthropometric traits, such as fat and height. None of the other SNPs under selection were found to be associated with the expression levels of these genes, meaning that the associations found could be associated with other members of the *CYP3A* family.

In Mendelian randomisation analyses, there were no causal relationship detected between CYP3A5 expression with neither PID (p_{IVW}=0.17), pulse rate (p_{IVW}=0.04, p_{Egger}=0.72,$p_{Intercept}$=0.02), erythrocyte count (p_{IVW}=0.91) and platelet count (p_{IVW}=0.11). For pulse rate, the strongest signals were located at the 3' end of CYP3A4, suggesting that this outcome could be associated with expression of another gene in the cluster and/or in another tissue, for which we did not have statistical power to detect appropriate eQTL instruments for Mendelian randomisation.

8. Supplementary file

Supplementary file 5. BED file containing the coordinates of the 57 genes used in the manuscript extracted from the UCSC genes table.

9. Competing interests

No competing interest is declared.

10. Author contributions statement

ARSH, IG and JP performed analyses. JCG, RP and IG pre-processed the data. ARSH, IG and JGH wrote the paper, revised by JCG, JP and RP. JGH initiated and supervised the project.

11. Acknowledgments

The authors thank the anonymous reviewers for their valuable suggestions. We also thank Claude Bherer and Marie-Pierre Dubé for their detailed review of this article. This work was completed thanks to computational resources provided by Calcul Quebec clusters Graham, Narval and Beluga. This study was supported by funding from the Canada Foundation for Innovation (CFI) (#40157) and the Montreal Heart Institute Foundation. ARSH received an internship scholarship from the Canadian Institutes of Health Research (CIHR). IG is a Robert-Cedergren Bioinformatics Awardee. JGH is a Fonds de Recherche du Québec en Santé (FRQS) Junior 2 research scholar.

12. Data availability

The 1000 Genomes Project, GEUVADIS is freely available. The GTEx v8 dataset was accessed through dbGaP under project number #19088. The UK Biobank was accessed through data access approval under the project number #15357. Information to apply for data access can be found here: https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access.

Chapitre 5

Autres contributions scientifiques

5.1. Publications en tant que co-auteur

5.1.1. Population Genomics Approaches for Genetic Characterization of SARS-CoV-2 Lineages

Dans le cadre de mon doctorat, j'ai participé aux travaux de mon laboratoire portant sur l'étude du coronavirus 2 du syndrome respiratoire aigu sévère (SARS-CoV-2, *severe acute respiratory syndrome coronavirus 2* en anglais), ce qui a résulté en la publication d'un article dont je suis le deuxième auteur, paru dans *Frontiers in Medicine* [460].

Cet article présente divers outils permettant de surveiller l'évolution du SARS-CoV-2.

L'un de ces outils est la fonction haploNet des progiciels 'Ape' [461] et 'pegas' [462] dans le logiciel R. Cette fonction permet de calculer le nombre de différences entre chaque paire de séquences afin de générer un réseau d'haplotypes. Cela offre une visualisation des haplotypes similaires et potentiellement un aperçu de leur évolution. En général, cette approche est utilisée sur des données à un moment précis et ne tient donc pas compte du temps. Cependant, avec la pandémie, il est possible d'estimer approximativement la date d'apparition d'un haplotype.

Ma principale contribution à cet article a été d'intégrer la dimension temporelle lors de la génération du réseau. J'ai utilisé la première date d'apparition d'un haplotype et généré un nouveau réseau d'haplotypes à chaque mois, en conservant le réseau généré le mois précédent. Les deux réseaux ont ensuite été combinés et, en cas de nouveaux haplotypes intermédiaires, soient qu'ils sont plus récents que les deux haplotypes auxquels ils sont liés, des cycles ont été générés. Pour éliminer ces cycles, le lien présent dans le réseau du mois précédent est

conservé et l'un des nouveaux liens a été supprimé. Ainsi, les connexions présentaient une continuité dans le temps. Le code est disponible sur github.

Une limitation majeure de cette approche réside dans l'imprécision de la date de première apparition. En effet, lors de l'enregistrement des informations par échantillon, des erreurs peuvent survenir, telles une date d'échantillonnage ultérieure à la date d'analyse. De plus, même si la date d'échantillonnage n'est pas erronée, ce n'est pas nécessairement le premier échantillon de cet haplotype, ce qui ne reflète pas réellement la date de première apparition, mais donne une estimation.

Malgré ces limitations, les réseaux générés étaient cohérents avec l'ordre d'apparition des variants majeurs. La figure 5.1 illustre un exemple d'un réseau d'haplotypes obtenu après cette méthode de correction par le temps.

5.1.2. Study of effect modifiers of genetically predicted CETP reduction

Avec mon expertise dans le domaine de la transcriptomique, j'ai également collaboré au projet de l'étudiant Marc-André Legault, qui a abouti à la publication d'un article dont je suis le troisième auteur et qui a paru dans *Genetic Epidemiology* [464].

Durant les études cliniques, certains facteurs démographiques sont souvent sous-représentés, ce qui pourrait mettre le doute sur certains résultats. Cet article présente comment le sexe et l'indice de masse corporelle (IMC) modifient les effet prédit génétiquement d'une réduction de la protéine CETP sur les biomarqueurs et les effets sur le système cardiovasculaire.

Une approche consistait à utiliser les niveaux d'expression de *CETP* dans certains tissus dans différents groupes de IMC et de sexe. Avec les données dont nous disposions, j'ai pu contribuer à ces analyses.

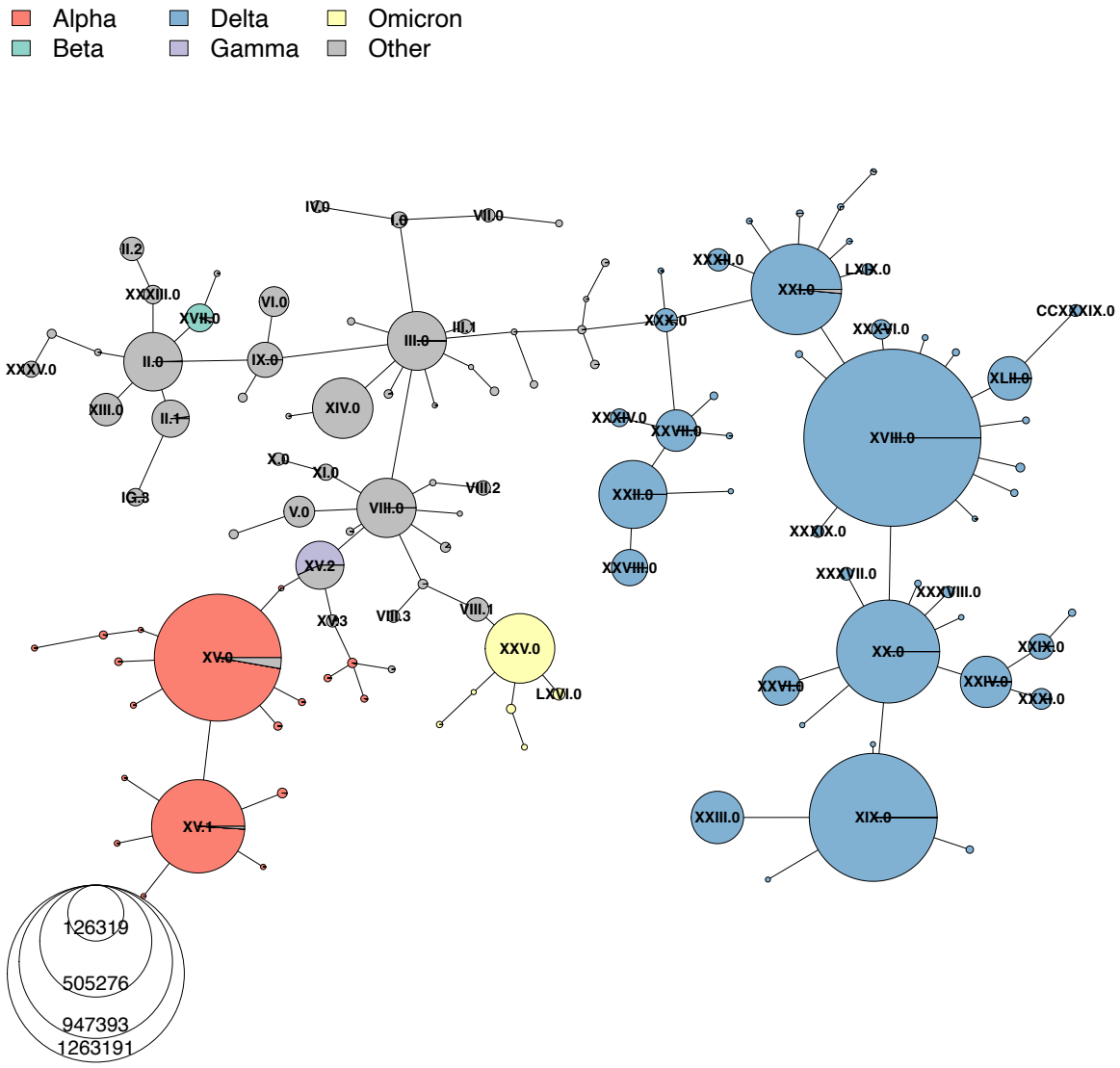

Fig. 5.1. Réseau d'haplotypes après correction par le temps pour les haplotypes du virus SARS-CoV-2 en date du 20 juillet 2022.
Les cercles représentent un haplotype formé par 25 positions génomiques, caractérisées par des changements dans le temps et une fréquence supérieure à 10% par mois. La taille des cercles correspond au nombre de séquences de la base de données GISAID [463] associées à chaque haplotype. La couleur indique la classification des souches du virus selon l'Organisation Mondiale de la Santé.

Chapitre 6

Discussion

Dans cette thèse, je présente le travail réalisé sur l'approfondissement des connaissances sur des gènes étudiés dans le domaine de la pharmacogénomique en utilisant des analyses de génétique des populations, de transcriptomique et d'étude d'association phénotypique.

6.1. Récapitulation des résultats du chapitre 2

L'objectif principal du premier projet, présenté dans les chapitres 2 et 3, était d'étudier les liens entre les gènes *ADCY9* et *CETP*, avec des associations parallèles par rapport à la réponse au dalcetrapib. Notre étude présentée au chapitre 2 a révélé la présence de signature de sélection au locus de rs1967309 du gène *ADCY9* (Figure 2.2), qui module la réponse au médicament dalcetrapib sur la survenue des événements cardiovasculaires, de sorte que les individus ayant le génotype AA avaient une réduction des événements cardiovasculaires et les individus GG avaient une augmentation [24].

Dans le chapitre 2, nous avons adopté une nouvelle approche pour identifier les événements de co-évolution, en nous basant sur le déséquilibre de liaison sur longue distance (LRLD). Grâce à cette méthode, nous avons identifié, au sein de la population péruvienne, un signal de co-évolution entre le locus rs1967309 dans le gène *ADCY9* (Figure 2.3) et des mutations dans le gène *CETP*, plus particulièrement avec la mutation rs158477. Ce cas représente l'une des rares détections de co-évolution observées chez l'être humain [159, 465]. L'application de notre nouvelle approche pourrait permettre d'identifier d'autres événements de co-évolution directement à partir de données génétiques, contribuant ainsi à une meilleure compréhension des relations entre les gènes.

La base de notre approche pour détecter la co-évolution consiste à évaluer si certaines combinaisons de génotypes sont significativement enrichies, ce qui pourrait signifier que la

combinaison est bénéfique à la survie et/ou la reproduction. Durant nos analyses, nous avons observé que la combinaison enrichie différait entre les hommes et les femmes. Même si les deux combinaisons incluaient le génotype protecteur avec le dalcetrapib, soit AA de la mutation rs1967309, l'autre génotype de la paire de cette combinaison était différente, soit que c'était avec les GG pour la mutation rs158477 chez hommes et avec les AA chez les femmes. Cela pourrait signifier que, chez les individus AA pour la mutation rs1967309, le génotype GG pour rs158477 chez les hommes est avantageux, tandis que ce serait le génotype AA pour rs158477 qui serait avantageux chez la femme. Étant donné que la direction est opposée, cela pourrait suggérer la présence d'une sélection sexuellement antagonistique. Cette hypothèse a été soutenue avec nos analyses d'association phénotypique (Figure 2.14), où l'interaction diffère entre les sexes. C'est la première fois qu'une co-évolution spécifique aux sexes a pu être identifiée chez l'humain seulement à l'aide de données génétiques.

6.2. Exploration des mécanismes potentiels de l'interaction entre les gènes *ADCY9* et *CETP*

Plusieurs mécanismes sont possibles afin d'expliquer la relation entre le gène *ADCY9* et le gène *CETP*.

Une première hypothèse possible serait une relation physique de la chromatine, potentiellement par une région régulatrice distante [466], nécessitant le repliement de la chromatine afin d'interagir avec la région promotrice du gène *CETP*. Afin de vérifier cela, j'ai regardé la base de données d'ENCODE en ligne, qui inclue la technique de Hi-C qui détecte les interactions physiques de la chromatine. Cependant, dans cette base de données, je n'ai pas observé d'enrichissement entre les deux loci. Cela n'exclut pas nécessairement cette hypothèse, étant donnée que l'interaction peut être circonstancielle, et donc, n'a été capturée lors de ces prises de données.

Autre qu'un contact physique de la chromatine, il est possible que cette interaction soit due à une voie métabolique commune [156]. En effet, des analyses chez les souris ont révélé la présence d'une relation épistasique fonctionnelle entre les deux protéines [89]. Dans le cadre de cette expérimentation, l'effet de l'interaction entre les deux gènes sur des phénotypes a été évalué en observant différentes combinaisons : l'absence ou la présence du gène murin *Adcy9* conjuguée à l'absence ou la présence du gène transgénique $CETP^{tg}$. Dans

cette expérimentation, seul l'absence des deux gènes conduit à des phénotypes différents, suggérant une redondance de fonction de ces gènes, générant de l'épistasie négative [242]. En cas de redondance, deux gènes ou voies métaboliques exercent une fonction similaire sur un phénotype. Tant que l'un des deux gènes ou voies métaboliques est présent, l'effet de l'absence de l'autre sur un phénotype est compensé par la fonction de celui-ci.

Cependant, les souris, qui n'expriment naturellement pas la protéine CETP [68, 69, 70] et le développement des plaques athérosclérotiques diffère du développement chez l'humain [467]. Cela suggérerait qu'elles pourraient avoir développé des voies signalétiques alternatives pour le développement de l'athérosclérose, où la protéine Adcy9 serait impliquée. En présence de la protéine $CETP^{tg}$, la voie signalétique à laquelle la protéine Adcy9 appartient serait alors compensée par la présence de $CETP^{tg}$. Cette voie métabolique alternative potentielle ne serait pas nécessairement présente chez l'humain, suggérant que la relation observée chez la souris n'est pas celle observable chez l'humain.

Néanmoins, cela n'exclut pas la possibilité que, chez l'humain, les protéines ADCY9 et CETP partagent une voie signalétique commune ayant une fonction sur le développement de l'athérosclérose. Dans ce contexte d'une voie signalétique commune, il serait possible qu'il existe un mécanisme de régulation entre ces deux gènes, par exemple, une régulation de CETP par ADCY9 ou vice-versa [242].

6.2.1. Expression génique

Pour examiner la relation entre les deux protéines, nous avons comparé le profil d'expression d'une lignée sauvage (WT, *Wild Type* en anglais) avec le profil d'expression résultant de l'inhibition d'ADCY9 (ADCY9-KD, *ADCY9-Knock Down* en anglais) dans des cellules hépatiques humaines (HepG2). Nous avons observé que l'inhibition d'ADCY9 avait un impact sur l'expression de plusieurs gènes, dont celle du gène *CETP* (Figure 6.1, Méthodes dans chapitre 2). Cela suggère que le gène *ADCY9* pourrait avoir un effet régulateur trans important sur le transcriptome, puisque les trans-eQTL affectent souvent des centaines de gènes [286, 468].

De plus, une mutation intronique dans le gène *ADCY9* influençant la transcription de $TGF-\beta$ a précédemment été trouvée [100], et il a également été découvert que le produit

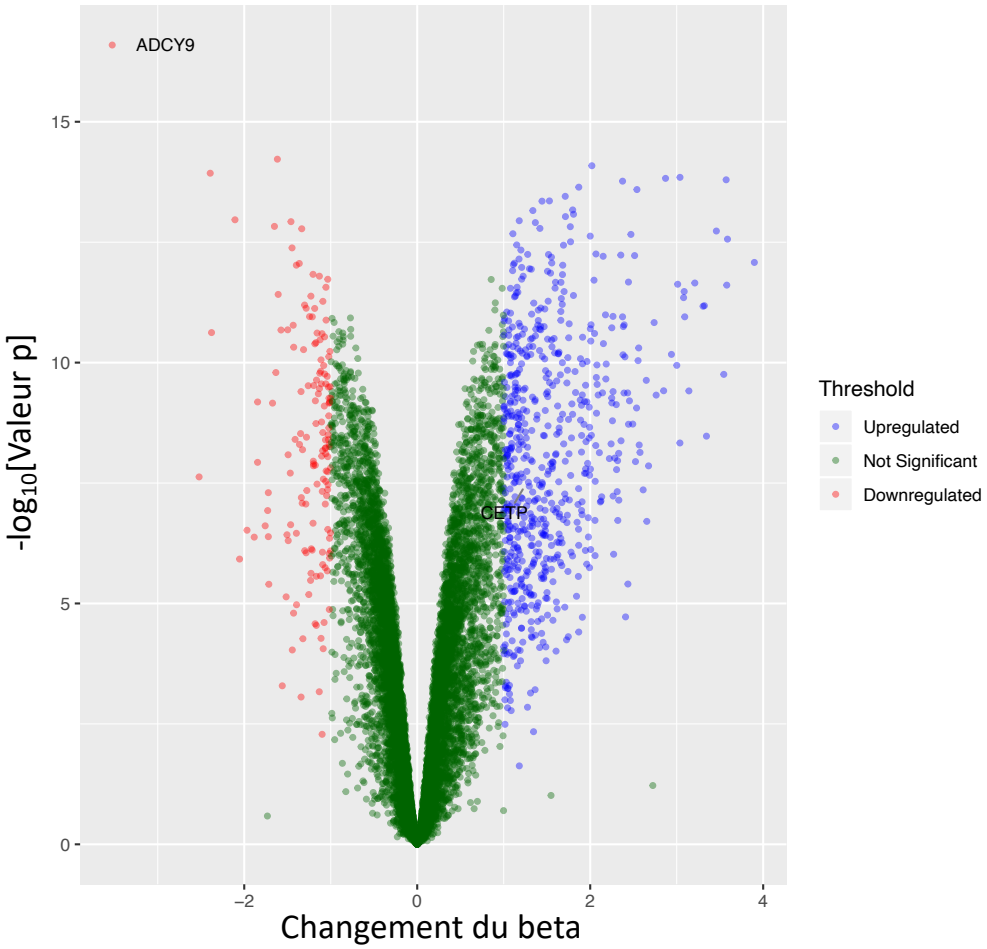

Fig. 6.1. Effet sur le transcriptome du *Knock-Down* d'ADCY9
Les gènes représentés en rouge (à la gauche de la figure) ont les niveaux d'expression significativement diminués et avec un facteur d'amplitude d'au moins 2. Les gènes représentés en bleu (à la droite de la figure) ont les niveaux d'expression significativement augmentés avec un facteur d'amplitude d'au moins 2. Les gènes *ADCY9* et *CETP* sont indiqués dans le graphique. Les données proviennent de lignée de cellules hépatiques humaines (HepG2) utilisées dans le chapitre 2 pour les analyses de l'effet de l'inhibition d'ADCY9 sur l'expression de *CETP*.

d'ADCY9, l'AMPc, influence la transcription de plusieurs autres gènes [469]. Ces observations indiquent que la relation entre ces deux gènes pourrait passer par la régulation de *CETP* via la protéine ADCY9. Des études approfondies des fonctions des gènes influencés par une modulation de la protéine ADCY9 pourraient permettre une meilleure compréhension de ses effets régulateurs.

Au niveau génétique, nous avons constaté que l'interaction épistasique entre les mutations rs1967309 dans le gène *ADCY9* et rs158477 dans le gène *CETP* influençait l'expression de

CETP dans plusieurs tissus, incluant des tissus du cerveau ainsi que de la peau (Figures 2.22 et 2.23). De plus, cette interaction semble également présenter un effet spécifique au sexe dans certains tissus, comme dans les artères tibiales (Figures 2.12, 2.22 et 2.24). Étant donné que les fonctions des ADCY dépendent des récepteurs qui les activent et du type cellulaire [92], il est possible que les conditions nécessaires à cette interaction ne soient pas présentes dans tous les tissus, ce qui expliquerait pourquoi elle n'est pas observée de manière globale.

Ces interactions significatives spécifiques à certains tissus entre les deux protéines suscitent un intérêt particulier en raison des phénotypes associés au gène *CETP*.

Par exemple, une interaction significative a été observée dans plusieurs régions du cerveau, dont l'hypothalamus (Figures 2.23 et 2.24), une région essentielle du cerveau impliquée dans le fonctionnement du système nerveux autonome. L'importance de cette région dans la relation entre les deux protéines est soulignée par l'observation de plusieurs phénotypes altérés chez la souris dans une étude précédente [89], liés au système nerveux autonome. Ces résultats suggèrent que leur interaction pourrait avoir un impact sur les fonctions cérébrales.

Cependant, le nombre d'échantillon était limité dans plusieurs tissus dans le jeu de données GTEx, principalement les tissus du cerveau, ce qui a limité la puissance statistique de nos analyses. Afin de valider nos observations, des analyses fonctionnelles ou des études utilisant des jeux de données d'expression plus volumineux sont nécessaires. De telles investigations contribueraient à éclairer les mécanismes sous-jacents à ces phénotypes et à mieux comprendre l'importance de cette interaction protéique.

6.2.2. Épissage alternatif

Bien que de nombreux résultats mettent en évidence une interaction au niveau de l'expression, la position de la mutation rs158477 dans le gène *CETP* a suscité notre intérêt, car elle se situe à moins de 200 paires de base après l'exon 9 de *CETP* qui est épissé dans l'isoforme *CETP-202*.

Les régions introniques situées avant et après un exon épissé jouent un rôle crucial dans la reconnaissance des facteurs d'épissage [190], ce qui suggère que la mutation rs158477 pourrait potentiellement influencer cette reconnaissance. En général, la majorité des isoformes alternatifs dans le transcriptome sont présents à faible fréquence, suggérant qu'ils ne soient pas fonctionnels [202]. Cependant, l'expression de l'isoforme alternatif *CETP-202* n'est pas

négligeable (Figure 3.5) et des analyses ont observé des effets de cet isoforme dans l'homéostasie lipidique intracellulaire [43, 51, 65], ce qui semble indiquer que cet isoforme soit fonctionnel.

Lorsque nous avons évalué l'effet de l'interaction épistasique entre les gènes *CETP* et *ADCY9* sur l'épissage alternatif de l'exon 9, nous avons observé une forte association entre la mutation rs158477 et une région fortement en LD avec la mutation rs1967309 dans le tissu mammaire. La mutation la plus fortement associée à l'épissage, soit rs4786452, est en parfait LD avec la mutation rs1967309 dans la population péruvienne (D'$_{PEL}$=1.0). Nos résultats montrent une absence de l'effet de rs158477 du gène *CETP* pour les individus qui possèdent les haplotypes avec le génotype délétère avec le dalcetrapib, soit GG de la mutation rs1967309 du gène *ADCY9*, tandis que chez les autres individus, la mutation continue d'avoir un effet sur la régulation de l'épissage alternatif de l'exon 9.

Implication de la protéine ADCY9 dans l'épissage alternatif. Lors des analyses de knock-down d'ADCY9, une augmentation des niveaux d'expression de *CETP* a été observée, ainsi qu'un changement dans l'épissage alternatif de l'exon 9 (AS9) de *CETP*, bien que non significatif (valeur p$_{Wilcoxon-T-Test}$=0.18, n=8 échantillons appariés) (Méthodes supplémentaires - Annexe A). La tendance montre une diminution des niveaux d'AS9 dans l'expérience ADCY9-KD par rapport à la lignée sauvage (WT) (Figure 6.2a), même si la couverture des fragments de lecture pour ADCY9-KD était plus élevée que pour la lignée sauvage (Figure 6.2b) (Nombre moyen de fragments de lecture couvrant la jonction dans WT=12.25 vs ADCY9-KD=27.75). Cela suggère que cette observation n'est pas due à une plus faible sensibilité de détection des événements d'épissage. Ainsi, ces résultats suggèrent un rôle régulateur potentiel de la protéine ADCY9 sur l'épissage alternatif de l'exon 9 de *CETP*.

Il existe peu de littérature sur le rôle des ADCY dans la régulation de l'épissage. Cependant, des études ont montré que la protéine kinase A (PKA), activée par le produit des ADCY, semble moduler la régulation de l'épissage d'au moins deux protéines [195, 383]. Cette régulation pourrait se faire potentiellement via les protéines SR, qui sont phosphorylées dynamiquement par des kinases lors de l'épissage [192, 193]. En effet, la région en aval des exons, où se trouve la mutation rs158447, est importante pour la reconnaissance par le snRNP U1 et cette liaison est stabilisée par différents facteurs, y compris les protéines SR [187, 190]. Les mutations de rs1967309 et de rs4786452 se trouvent dans l'intron 2 du

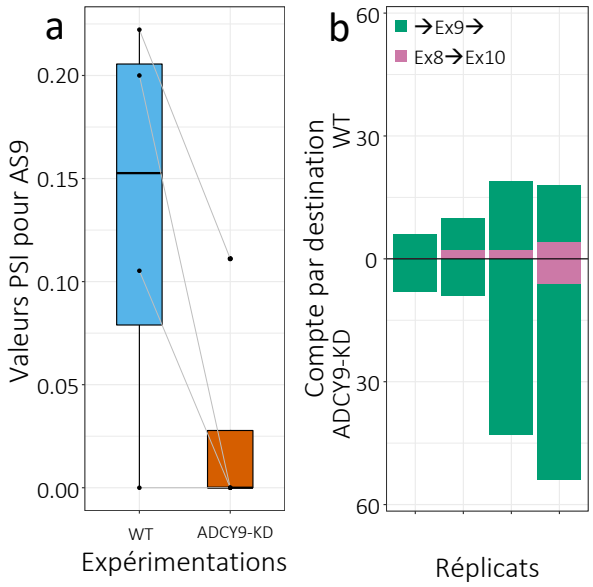

Fig. 6.2. Effet du *Knock-Down* d'ADCY9 sur l'épissage alternatif de l'exon 9
(a) Comparaison des valeurs PSI pour l'épissage alternatif de l'exon 9 (AS9) dans quatre réplicats expérimentaux appairés dans la lignée sauvage (WT) et avec le *Knock-Down* d'ADCY9 (ADCY9-KD) dans la lignée cellulaire de HepG2 (b) Couverture par destination des fragments de lecture couvrant la jonction associée à l'exon 9 dans les quatre réplicats appairés. Les fragments de lecture couvrant l'inclusion de l'exon 9 sont indiqués par "$\rightarrow Ex9 \rightarrow$", tandis que ceux appuyant l'épissage de l'exon 9 sont indiqués par "$Ex8 \rightarrow Ex10$".

gène *ADCY9* et la régulation génétique du gène *ADCY9* varie énormément entre les tissus (Figure 6.3). Une hypothèse serait que cette région affecte l'expression d'une protéine spécifique, potentiellement un isoforme spécifique d'*ADCY9*, qui permet la reconnaissance de la région de rs158477. Sans cette protéine, la mutation rs158477 ne modulerait pas l'épissage de *CETP*.

Ces observations suggèrent donc un rôle de la protéine ADCY9 dans la régulation du mécanisme d'épissage alternatif de certains gènes, dont potentiellement *CETP*.

Une prochaine étape serait d'étudier les protéines affectées par le gène *ADCY9*, particulièrement la région de rs1967309, afin d'identifier celles qui pourraient être associées à l'épissage, ainsi qu'à la reconnaissance de la région de rs158477.

6.2.3. Impacts potentiels

Nos analyses ont révélé de nouvelles fonctions potentielles du gène *ADCY9*, l'un des moins explorés au sein de la famille des *ADCY* [90], dans la régulation des niveaux d'expression du

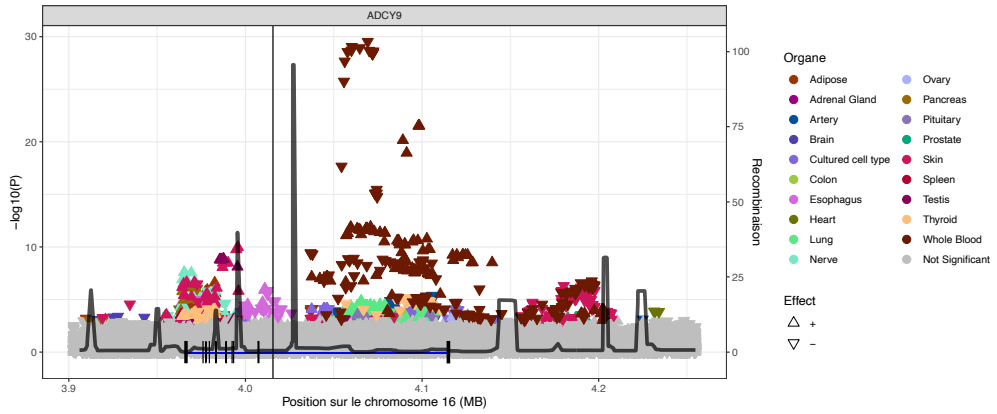

Fig. 6.3. Régulations génétiques du gène *ADCY9* à travers les tissus de GTEx
Les cis-eQTL pour le gène *ADCY9* ont été analysés dans le locus d'*ADCY9* pour 49 tissus provenant de GTEx. Les noms des tissus sont regroupés par type, et seules les mutations qui ont atteint un seuil statistique (valeur p $< 0.05/49$ tissus) sont colorées. Le gène *ADCY9*, composé de 11 exons, est représenté en dessous de la figure. La ligne noire verticale sur la figure représente la position de la mutation rs1967309. La ligne noire variable traversant horizontalement la figure représente le taux de recombinaison dans la population CEU de 1000G. La forme du triangle indique la direction de l'effet de la mutation sur l'expression d'*ADCY9*.

transcriptome, notamment en ce qui concerne l'épissage. L'étude des gènes dont les niveaux d'expression sont influencés par la protéine ADCY9 ouvrira de nouvelles perspectives pour mieux comprendre les réseaux de régulation génique, ainsi que les liens entre ce gène et divers phénotypes.

Nos résultats pointent vers des pistes de mécanismes potentiels expliquant la relation entre le gène *ADCY9* et le gène *CETP*, cible du médicament dalcetrapib. Cette compréhension pourrait éclairer sur le mode d'action spécifique de ce modulateur de CETP dans des événements cardiovasculaires survenant uniquement dans un sous-groupe d'individus, ce qui pourrait permettre de développer des traitements de médecine de précision.

La majorité des interactions épistasiques connues se trouvent chez les espèces telles qu'*E.coli* [470, 471], mais les cas chez l'humain sont moins bien répertoriés [472, 473], souvent identifiés à l'aide d'approches utilisant des lignées cellulaires ou des organismes vivants. Nos résultats ont pu démontrer l'utilité d'une nouvelle approche, basée sur la génétique des populations, pour identifier des interactions épistasiques uniquement à l'aide des données génétiques.

6.3. Pression de sélection co-évolutive en réponse à l'environnement en haute altitude

Les populations vivant en haute altitude, telles que la population andéenne, doivent faire face à des défis environnementaux comme une pression partielle d'oxygène réduite et divers autres stress. La population péruvienne, qui a été identifiée comme étant soumise à une pression de sélection co-évolutive dans le chapitre 2, provient de la population andéenne. Il est possible que cette signature de sélection soit due à l'impact environnemental de l'altitude sur cette population.

Des voies signalétiques communes ont été identifiées entre l'hypoxie causée par la haute altitude et les perturbations du système cardiovasculaire [305, 306, 307, 308, 309], et, dans la population des Andes, une pression de sélection en réponse à l'hypoxie a été associée avec le système cardiovasculaire [269]. Dans nos analyses, l'interaction entre les mutations rs1967309 dans le gène *ADCY9* et rs158477 dans le gène *CETP* est associée avec des événements cardiovasculaires et des phénotypes respiratoires. De plus, nos analyses de co-évolution dans le chapitre 2 n'ont pas montré de différence entre les différents groupes d'âge, ce qui suggère que la pression de sélection pourrait arriver en bas âge, voir même pendant la période de grossesse. Durant nos analyses de randomisation mendélienne dans le chapitre 3, nous avons également pu observer des associations entre l'épissage de l'exon 9 et le phénotype des complications de grossesse (Figure 3.4), ce qui appuie l'hypothèse que la pression de sélection surviendrait pendant la période de grossesse.

6.3.1. Lien potentiel entre les gènes *ADCY9*, *CETP* et le stress oxydatif

Un processus biologique par lequel cette pression de sélection pourrait se produire est le mécanisme de stress oxydatif. Ce processus survient lorsque la production d'espèces réactives de l'oxygène (ROS, *Reactive Oxygen Species* en anglais) par la mitochondrie est excessive et/ou lorsque la production d'antioxydant est réduite. Le stress oxydatif se produit dans diverses situations, telles que l'hypoxie causée par la haute altitude [474] et pendant la grossesse [475]. Il est également associé, entre autres, à des dysfonctionnements endothéliaux qui

peuvent avoir un impact sur le développement des maladies cardiovasculaires et la stabilité des plaques athérosclérotiques [476].

Les gènes *ADCY9* et *CETP* semblent tous deux être associés à ce processus métabolique. En effet, le gène *ADCY9* est impliqué dans des voies de signalisation associées au stress oxydatif [477]. Concernant la protéine CETP, elle a été associée avec le stress oxydatif dans les cellules endothéliales aortiques chez la souris [478], et les particules HDL-c, dont la concentration est modulée par CETP, possèdent des propriétés antioxydantes [479]. Enfin, des études chez la souris [89] ont montré que l'interaction entre les protéines $CETP^{tg}$ et Adcy9 modulait les fonctions endothéliales et le développement des plaques athérosclérotiques.

6.3.1.1. Interaction génétique, stress oxydatif et risque de complications de grossesse

Dans l'éventualité où l'interaction affecterait le stress oxydatif, cela pourrait expliquer en partie l'effet spécifique au sexe que nous avons observé. En effet, le corps des femmes enceintes est soumis à un stress oxydatif important, particulièrement lors de grossesses présentant une prééclampsie [480]. Cette complication hypertensive de la grossesse est de deux à trois fois plus fréquente dans les populations vivant en haute altitude, telles que les populations péruviennes et boliviennes, et elle constitue une cause importante de mortalité maternelle et fœtale [475, 481, 482].

Pendant la grossesse, le placenta, où l'expression de *CETP* est élevée (Selon *Protein Atlas*, accès le 9 juin 2023), augmente significativement la production des ROS afin de déclencher une réaction inflammatoire systémique. Pour éviter les dommages tissulaires causés par cette augmentation, il y a également une augmentation de la synthèse des antioxidants, impliquant potentiellement les HDL. Tout au long de la grossesse, mais particulièrement pendant le deuxième et troisième trimestre, les profils lipidiques, incluant les niveaux des HDL-c, et l'activité de CETP fluctuent de façon importante [378, 483].

Lors de nos analyses de randomisation mendélienne, nous avons observé une corrélation entre une augmentation de l'épissage alternatif de l'exon 9 et une réduction du risque de fausse couche (Figure 3.4). Ensuite, lors de nos analyses d'interaction, nous avons observé que les individus ayant l'allèle C pour le locus rs4786452 dans le gène *ADCY9* présentaient une variation du niveaux d'épissage alternatif de l'exon 9 selon l'allèle présent au locus rs158477 (Figure 3.4). Il est important de se rappeler que l'allèle C de rs4786452 (*ADCY9*) se trouve

toujours sur le même haplotype que l'allèle A du locus rs1967309 dans le gène *ADCY9*, bénéfique pour la réponse au dalcetrapib. Chez ces individus, ceux ayant le génotype AA-rs158477 dans le gène *CETP* présentaient un niveau d'épissage plus élevé que ceux ayant le génotype GG-rs158477, et donc potentiellement un risque de fausse couche plus faible. Cette combinaison est la plus fréquente chez les femmes péruviennes, ce qui pourrait expliquer l'enrichissement observé chez elles.

Cependant, l'association causale entre l'épissage alternatif de l'exon 9 et le phénotype de complication de grossesse, obtenu de la *UK Biobank*, n'est pas significatif au niveau pan-génomique, cela pourrait être causé par un nombre limité de cas (N=5214 cas/212 254 contrôles) et que les femmes de cette cohorte ne vivent pas dans un environnement avec du stress oxydatif chronique. Si l'effet de l'interaction est protecteur dans un environnement exposé au stress oxydatif chronique, les effets bénéfiques pourraient être limités dans un environnement faible en stress oxydatif. En revanche, la combinaison enrichie chez les hommes est celle augmentant le risque de fausse couche, mais cela ne les affecterait pas directement car ils ne peuvent pas être enceintes.

Les grossesses avec un fœtus de sexe masculin et féminin diffèrent, et les fœtus ne réagissent pas de la même manière à l'environnement maternel. Il a été constaté que les fœtus masculins sont exposés à un environnement plus pro-inflammatoire que les fœtus féminins et ils semblent également être plus vulnérables à l'environnement maternel [484, 485], y compris aux niveaux de stress oxydatif de la mère [481]. Étant plus vulnérable, ils pourraient être plus exposés au risque de fausse couche. Pendant la grossesse, les niveaux des sous-types de HDL, qui ont différentes propriétés antioxydantes et anti-inflammatoires [28, 31, 32, 33, 34, 35], varient entre les trimestres et entre le foetus et la mère [486, 487], ainsi qu'entre les sexes à l'âge adulte [488]. Les variations des sous-types de HDL selon le sexe du fœtus et l'impact de l'épissage de l'exon 9 sur les sous-types de HDL demeurent des domaines qui restent à être explorés.

Ayant pu observer une différence de direction des effets entre les sexes pour les maladies cardiovasculaires, il est possible que les fœtus masculins portant le génotype GG-rs158477 produisent un environnement moins inflammatoire ou qu'ils soient moins sensibles à un environnement pro-inflammatoire, ce qui pourrait réduire le risque de résulter en une fausse couche.

Lien avec la prééclampsie. En lien avec nos découvertes concernant la sélection spécifique aux sexes au sein de la population péruvienne, la prééclampsie, très répandue dans cette communauté, représente un risque significatif de mortalité. Cette condition pourrait donc être étroitement liée à la signature de sélection que nous avons identifiée, puisque cela influence la survie et la reproduction [489] de la mère et du foetus. La prééclampsie est associée au stress oxydatif, mais également au transport inverse du cholestérol [490]. Une des fonctions de CETP est de performer des transferts entre les sous-types de HDL, qui ont également différentes propriétés antioxydantes [31]. Potentiellement que cette fonction pourrait être particulièrement bénéfique dans ces grossesses. De plus, des études ont démontré que l'impact de la prééclampsie sur le fœtus diffère entre les sexes [482, 491]. Ces éléments suggèrent que, pendant la grossesse, les combinaisons de génotypes favorisant la survie pourraient varier. Étant donné que la prééclampsie est moins courante dans les autres populations que nous avons étudiées, il est possible qu'aucune combinaison spécifique de génotypes ne soit nécessaire pour assurer la survie du fœtus dans ces populations, ce qui pourrait expliquer l'absence d'enrichissement de combinaisons observés. Cependant, dans les populations présentant une prévalence élevée de prééclampsie, ces combinaisons pourraient devenir essentielles en fonction du sexe, ce qui pourrait expliquer les résultats observés lors de nos analyses.

Étude prospective visant à évaluer les biomarqueurs du stress oxydatif chez les nouveau-nés et les fausses couches. Ces hypothèses nécessitent cependant des études supplémentaires. Pour les vérifier, il serait idéal de constituer une biobanque prospective regroupant des échantillons d'ADN de mères enceintes ainsi que du fœtus, qu'ils soient nés à terme ou résultant en une fausse couche, idéalement dans la population péruvienne, et avec des données biologiques comme de la transcriptomique et de la métabolomique, afin d'obtenir des biomarqueurs pour quantifier les niveaux de stress oxydatif. Cela nous permettrait de déterminer s'il est possible d'identifier une combinaison génétique enrichie chez les nouveau-nés par rapport à ceux ayant résulté en une fausse couche. Cette base de données nous permettrait également d'étudier l'effet de l'interaction entre nos mutations, mais aussi d'autres combinaisons génétiques létales, fournissant ainsi des informations supplémentaires pour les conseils génétiques, tout en améliorant notre compréhension de la relation génétique entre la mère et le fœtus.

6.3.2. Interaction génétique et capacités pulmonaires chez les femmes

Durant nos analyses, nous avons également pu observer un effet de l'interaction entre les deux mutations, soit rs1967309 dans le gène *ADCY9* et rs158477 dans le gène *CETP*, sur les capacités pulmonaires chez les femmes (Figure 2.14). Il est connu que le développement des poumons diffère entre les sexes pendant la grossesse [492]. Des différences dans la capacité pulmonaire entre les sexes apparaissent dès la naissance et persistent tout au long de la vie, et la prévalence des maladies pulmonaires varie également entre les sexes.

Étant donné que les niveaux d'expression des deux gènes sont élevés dans les poumons, il est possible que leur interaction puisse affecter le développement pulmonaire au stade embryonnaire ou le maintien des capacités pulmonaires à l'âge adulte. Le développement pulmonaire se produit principalement pendant la deuxième moitié de la grossesse, une période où les niveaux de CETP maternel fluctuent significativement [378, 483]. Il a aussi été observé que l'expression des gènes *ADCY* est modulée par les hormones sexuelles [493, 494], qui varient également pendant la grossesse. Ainsi, l'interaction entre ces protéines pourraient potentiellement influencer le développement pulmonaire chez les fœtus.

En lien avec la signature de sélection identifiée, en situation d'hypoxie, telle que celle induite par une haute altitude, le corps humain doit s'adapter. Des différences dans ces mécanismes d'adaptation ont été observées entre les sexes [495]. Il serait donc pertinent de mener des études supplémentaires pour évaluer l'effet de l'interaction entre ces deux gènes en situation d'hypoxie, que ce soit sur les niveaux d'expression ou d'épissage, ou sur les capacités pulmonaires en haute altitude. De telles études permettraient de mieux comprendre le comportement de cette interaction dans des conditions d'hypoxie.

Limitations et perspectives de l'association phénotypique dans la population péruvienne. Nous avons pu formuler des hypothèses concernant les phénotypes potentiellement soumis à une pression de sélection dans la population péruvienne. Cependant, la principale limitation de cette association réside dans l'ancestralité génétique de la population étudiée, qui est d'ascendance européenne. Bien que les effets biologiques puissent être présents, il est possible que leur amplitude diffère ou que des mécanismes biologiques

spécifiques, tels que ceux déclenchés en cas d'hypoxie, puissent manquer pour observer l'association véritable [496]. Pour remédier à cette limitation, il serait nécessaire de reproduire nos analyses d'association phénotypique dans des biobanques comprenant des individus d'origine péruvienne. De plus, si cette association se manifeste en cas d'hypoxie, il serait également pertinent d'étudier d'autres populations vivant en haute altitude, comme les Boliviens et les Tibétains. Cela permettrait de mieux comprendre l'interaction épistasique et d'évaluer si elle est spécifique à une population ou si elle se manifeste dans des conditions d'hypoxie plus largement.

6.3.3. Sous-populations et signatures de sélection associées à rs1967309

Nos analyses se sont concentrées sur la population péruvienne, qui montrait une différence dans la fréquence allélique entre les populations. Cependant, dans nos analyses d'iHS (Figures 2.2b et 2.16), nous avons observé des signaux pour rs1967309 dans plusieurs autres populations, mais principalement des populations d'Asie du Sud. Nous n'avons pas poussé plus loin nos analyses dans ces populations, mais il pourrait être intéressant de mener des études similaires à ce que nous avons fait afin d'identifier les potentielles pressions de sélection associées à ce locus dans ces populations.

Une autre observation intéressante est que, dans la population du Kenya en Afrique (LWK), les fréquences de l'allèle A-rs1967309 (79%) et G-rs158477 (73%) semblent également plus élevées que le reste de la population africaine (70% et 58% sans LWK respectivement). Cette population se trouve également dans une région montagneuse, où l'altitude de la région d'échantillonnage est d'environ 1500 mètres au dessus du niveau de la mer. Même si nous n'avons pas observé de LRLD entre les deux mutations dans cette population (valeur p >0.05), cela n'exclut pas nécessairement la possibilité qu'une relation existe, mais qu'elle n'est pas observable avec notre approche actuelle. En effet, l'approche initiale utilisée ne permet pas de capturer la présence d'enrichissement pour tous les types de combinaison de génotypes, et l'approche par χ^2_4 est limitée par sa puissance statistique avec des petites tailles d'échantillons comme dans le cas du jeu de données de 1000G. L'utilisation de bases de données plus volumineuses, telles que le projet H3Africa [497] comprenant plus de 70 000

participants à travers l'Afrique, permettrait d'augmenter la puissance statistique et d'évaluer la possibilité d'événements de co-évolution entre les deux gènes.

En étudiant l'origine des différentes signatures de sélection sur le gène *ADCY9* dans les autres populations, cela pourrait aider à comprendre son association derrière les divers phénotypes auquel il est associé.

6.3.4. Impacts potentiels

Nos analyses ont révélé des signatures de sélection spécifique aux sexes dans la population péruvienne, ainsi que suggérer des hypothèses sur les phénotypes associés.

La détection d'événements de co-évolution demeure complexe. Quelques cas de co-évolution ont été observés chez l'humain [465], mais, à ma connaissance, aucune co-évolution sexuellement antagonistique sur un chromosome autosomal n'a été rapportée dans la littérature chez cette espèce. Le développement de notre méthodologie pourrait faciliter l'identification d'autres cas de co-évolution, potentiellement spécifiques aux sexes, ce qui permettrait une meilleure compréhension des réseaux d'interaction entre les gènes et leurs effets sur les phénotypes, ainsi que les différences entre les sexes.

Nous avons émis l'hypothèse que la relation entre les gènes *ADCY9* et *CETP* pourrait être liée au stress oxydatif, avec un lien avec des phénotypes de grossesse. Cela pourrait permettre de mieux comprendre le mécanisme derrière les risques de complication de grossesse, incluant les risques de prééclampsie. Avec une meilleure compréhension de ce qui cause ce risque, il serait possible de développer des thérapies qui viseraient à réduire ces incidences dans des groupes d'individus spécifiques.

6.4. Implications potentielles des isoformes de *CETP* dans l'efficacité des inhibiteurs

Parmi les essais cliniques portant sur les inhibiteurs de la protéine CETP, seul le modulateur dalcetrapib a été démontré comme étant modulé par la mutation rs1967309 dans le gène *ADCY9* [24]. La raison est encore inconnue, mais nos analyses ont permis d'établir certaines hypothèses.

La majorité des études concentrent leur analyse d'association entre de la protéine CETP et les maladies cardiovasculaires sur son effet sur le profil lipidique sanguin. Cependant,

nos analyses suggèrent qu'une partie de l'association avec la protéine CETP passerait par les isoformes présents dans les cellules. Les inhibiteurs de la protéine CETP, en raison de leur caractéristique lipophile, ont la capacité d'entrer dans les cellules [348, 349] et, par conséquent, il est possible qu'ils ciblent les isoformes de *CETP* dans la cellule, avec des affinités variables.

Nos résultats suggèrent que les changements dans les niveaux d'épissage alternatif, en d'autres mots la proportion des différents isoformes, pourraient également influencer les maladies cardiovasculaires, potentiellement avec des effets spécifiques aux tissus. Selon les effets des inhibiteurs sur chaque isoforme et dans différents tissus, il est possible que l'inhibition de la protéine CETP ne produise pas uniquement les effets bénéfiques attendus.

6.4.1. Particularité du dalcetrapib

Même si des pistes ont été émises quant au lien entre le gène *ADCY9* et la réponse au dalcetrapib, plusieurs questions demeurent sans réponse, notamment : pourquoi le dalcetrapib est-il le seul qui est influencé par la mutation dans le gène *ADCY9*?

Pour explorer cette interrogation, une meilleure compréhension de l'effet des inhibiteurs de CETP avec la protéine est nécessaire. Tous les inhibiteurs de CETP, à l'exception du dalcetrapib, bloquent l'intégralité des fonctions de transfert de CETP [36, 348], augmentant ainsi la concentration des HDL-c au détriment des lipoprotéines à plus faible densité. Cependant, l'étape de transfert entre les sous-types de HDL est également inhibée, qui est pourtant une étape importante dans le transport inverse du cholestérol [498]. En revanche, le dalcetrapib n'inhibe que partiellement les fonctions de CETP, permettant le maintien du transfert du cholestérol des HDL_3 vers les HDL_2 [498], avec les HDL_2 pouvant diminuer le risque des maladies cardiovasculaires [28, 31, 32, 33, 34, 35]. Chez les individus avec un déficit de CETP, une augmentation des niveaux de HDL_2 a pu être observée, avec une diminution des risques de maladies cardiovasculaires [73, 74], ce qui ressemble aux effets du dalcetrapib lorsqu'on considère le gène *ADCY9*. Il est donc possible que le maintien du transfert plasmatique ne soit pas seul responsable des effets bénéfiques de l'inhibition de CETP.

Étant donné que seul le transfert des HDL vers les lipoprotéines à plus faible densité est inhibé par le dalcetrapib, il est possible que les fonctions de CETP à l'intérieur des cellules, régies par les isoformes, soient préservées. Nous avons constaté que des modifications dans

la proportion des isoformes sont associées à différents phénotypes, y compris les maladies cardiovasculaires (Figure 3.3). Il est donc envisageable que l'impact des mutations du gène *ADCY9* sur la régulation de l'épissage de l'exon 9 puisse être observé sur ces phénotypes dans certains groupes d'individus. Par conséquent, les effets du dalcetrapib sur les événements cardiovasculaires pourraient dépendre des niveaux d'épissage, lesquels semblent être régulés par le gène *ADCY9*. Puisque les autres inhibiteurs pourraient également affecter les fonctions des isoformes, cette modulation par *ADCY9* pourrait ne pas être observable dans leur cas.

Une autre possibilité est que l'inhibition seule des fonctions de CETP ne soit pas suffisante, mais qu'une diminution de la concentration soit également nécessaire. En effet, à la fois le gène *ADCY9* et l'isoforme CETP-202 diminuent les niveaux de CETP plasmatique, soit en modulant son expression (Figure 2.11), soit en inhibant sa sécrétion dans le plasma [42, 66]. Il est possible qu'une diminution de ses niveaux, combinée au maintien des fonctions de transferts homotypiques, soit nécessaire pour générer des effets bénéfiques. Des études supplémentaires sont nécessaires pour déterminer la cause de ces effets bénéfiques et des relations avec les isoformes de CETP.

Pour parvenir à une compréhension précise des mécanismes sous-jacents et des conséquences cliniques des interactions entre les inhibiteurs de CETP et les isoformes de *CETP*, des études approfondies supplémentaires seront nécessaires. Une meilleure caractérisation de leurs affinités avec les isoformes et de leurs effets spécifiques sur les isoformes contribuera à une meilleure évaluation des implications thérapeutiques des isoformes de *CETP* par rapport à ses inhibiteurs.

6.4.2. Épistasie signée

Dans l'étude dal-OUTCOMES, ainsi que dans nos propres analyses sur l'interaction des mutations sur les événements cardiovasculaires (Figure 2.14), nous avons observé des inversions d'effet en fonction du génotype du locus rs1967309 dans le gène *ADCY9*. Cette inversion pourrait être associée à de l'épistasie signée, c'est-à-dire une interaction entre les gènes qui conduit à des effets opposés selon les combinaisons génotypiques.

Plusieurs mécanismes peuvent expliquer ce type d'interaction. Une hypothèse amenée pour expliquer cette interaction est basée sur l'idée d'une voie signalétique commune, où les paramètres optimaux d'un gène dépendent des paramètres d'un autre gène indépendant

[470]. Cela signifie que, par exemple, en l'absence d'un gène, il serait plus favorable d'avoir un niveau d'expression plus élevé du deuxième gène, tandis qu'en présence du premier gène, il serait plus favorable d'avoir un niveau d'expression plus faible du deuxième gène. Cette hypothèse pourrait également être en accord avec l'hypothèse d'évolution sexuellement antagoniste, où l'effet optimal chez les femmes pourrait être délétère chez les hommes, et vice-versa. Par exemple, les effets protecteurs ou délétères de l'interaction pourraient varier selon les niveaux des hormones sexuelles, qui ont un impact important sur divers phénotypes, comme les maladies cardiovasculaires [499] où une inversion d'effet a été observée dans nos analyses (Figure 2.14).

Une hypothèse alternative pourrait être liée à la présence d'un sommet de valeur adaptative dans le paysage évolutif (Figure 1.10). Cela suggère que le phénotype optimal ne serait pas un extrême, comme une expression fortement diminuée de *CETP*, mais plutôt un niveau intermédiaire. Dépasser ce sommet aurait un impact négatif sur la valeur adaptative [500]. Dans le cas de *CETP*, étant donné que l'isoforme CETP-202 régule l'excrétion de CETP-201 et que la région de rs1967309 ou l'expression d'*ADCY9* régulent l'épissage de *CETP*, il est possible qu'un niveau d'épissage spécifique soit nécessaire pour atteindre ce sommet. Cela suggère que le maintien d'un équilibre dans l'épissage de *CETP* pourrait être crucial pour la fonction optimale de cette protéine et de son rôle dans le métabolisme lipidique.

L'identification de l'épistasie signée a été rapporté plusieurs fois, principalement dans les bactéries [470, 471, 501], mais plus rarement chez l'humain [473].

6.4.3. Impacts potentiels

Nos analyses ont permis de faire un lien entre le mystère de la relation entre le gène *ADCY9* et la modulation de la réponse au dalcetrapib. Cette relation pourrait être médiée par la régulation des niveaux des isoformes de CETP. En comprenant plus précisément leur interaction, il pourrait être envisageable de développer de nouveaux inhibiteurs prenant en compte les isoformes dans leur mécanisme d'inhibition et qui visent des groupes plus spécifiques.

6.5. Quantification des isoformes

Deux approches ont été utilisées afin de quantifier les isoformes de *CETP*.

6.5.1. Quantification des niveaux des isoformes

La première méthode est celle via la reconstruction des transcrits avec le logiciel RSEM [218]. Il a été possible d'observer des associations génétiques significatives avec l'expression de *CETP-201* et de *CETP-203*, mais pas pour l'expression de *CETP-202*. Le manque de précision lors de la quantification des isoformes pourrait diminuer la puissance de détection des eQTL associés à des variants dont la fréquence allélique est plus rare. De plus, nous avons observé que le logiciel attribuait faussement de l'expression à *CETP-202* et *CETP-203* dans les LCL, dans deux bases de données distinctes, soient GTEx et GEUVADIS. Cela nous a permis d'observer une source de biais potentiel causant ce manque de précision, principalement pour *CETP-202*. Dans cette lignée cellulaire, il y avait presque une absence complète de couverture entre les exons 1 et 6, avec une augmentation importante au milieu de l'exon 9 (Figure 3.6). Étant donné que la forme de *CETP* détectée était courte, la couverture était attribuée à l'isoforme le plus court avec l'exon 9, soit l'isoforme *CETP-203*. La forme *CETP-202* était détectée à cause de la présence de l'épissage alternatif de l'exon 9.

6.5.2. Quantification des événements d'épissage alternatif

La deuxième approche utilisée se base sur les informations des jonctions d'épissage. Cette approche permet de diminuer les erreurs de classification causées par une faible couverture. Pour y parvenir, les individus présentant une couverture insuffisante pour ces jonctions sont exclus, ne conservant ainsi que ceux offrant une quantification plus fiable. Pour cette approche, nous avons utilisé deux logiciels, soient MAJIQ [229] et ASpli [231].

Les deux outils utilisent des méthodes de calcul différentes, mais fournissent des valeurs PSI similaires (Figure 3.15), comme on peut s'y attendre lors de la comparaison de logiciels axés sur la détection des événements d'épissage alternatif [191]. Dans les deux cas, la régulation génétique pour l'épissage liée à la première jonction correspondait à la région associée aux eQTL de *CETP-203*, ce qui confirme que la quantification par RSEM de l'isoforme *CETP-203* n'était pas significativement biaisée en raison d'une plus faible couverture, du moins dans les tissus avec un niveau d'expression suffisant.

Contrairement à la méthode de quantification des niveaux par reconstruction des isoformes pour *CETP-202*, cette approche a permis d'identifier une région génétique qui régule

la proportion de cet isoforme. Cependant, il n'aurait pas été possible d'observer la particularité dans les LCL avec cette méthode uniquement. De plus, cette approche présente une limitation majeure : elle permet uniquement de quantifier les proportions des isoformes et non leurs niveaux d'expression respectifs.

Limitation du séquençage de deuxième génération. Ces résultats soulignent, une fois de plus [211], la limitation majeure du séquençage à ARN utilisant des courts fragments de lecture pour la quantification des niveaux des isoformes. En effet, ces fragments sont généralement courts, allant de seulement 50 à 300 paires de bases, ce qui ne couvre qu'une petite partie du gène, ne permettant pas de recouvrir tous les éléments nécessaires pour une distinction précise entre les isoformes. L'utilisation du séquençage de troisième génération permettrait une couverture complète du gène et de chacun de ses isoformes, ce qui aurait pu éviter cette erreur d'attribution. Cependant, avec la technologie actuelle, le séquençage de longs fragments présente plusieurs limitations, telles qu'une couverture moindre limitant la quantification des isoformes, un haut niveau d'erreur de séquençage, ainsi qu'une disponibilité encore restreinte des ensembles de données [213].

6.5.3. Nouveaux isoformes

Dans les LCL, nous avons observé deux formes tronquées du gène *CETP* (Figure 3.6a), potentiellement de nouveaux isoformes de *CETP* qui n'ont jamais été rapportés dans la littérature jusqu'à maintenant. Les LCL proviennent d'une transformation cellulaire à l'aide du virus EBV. Le gène *CETP* a déjà été trouvé comme ayant des relations avec ce virus, soit avec des sites d'interaction avec l'épisome du virus (Figure 3.6b, [502]), soit en ayant une activité augmenté chez les individus infectés [48], signifiant que l'infection pourrait affecter la transcription de *CETP*. Il n'est cependant pas connu si les nouveaux isoformes n'existent que lors de l'infection en laboratoire ou s'ils pourrait également être générés chez les humains qui sont infectés par ce virus. Puisque les formes tronquées conservent plusieurs domaines fonctionnels, il pourrait être intéressant de vérifier si elles sont exprimées dans les individus affectés, si elles ont des fonctions régulatrices et si elles produisent des formes protéiques. Il faudrait également pousser notre recherche afin de déterminer pourquoi la forme complète de *CETP* n'est pas observable dans cette lignée cellulaire.

Perspectives d'analyses *single-cell* et études en haute altitude. Les régulations de l'expression et de l'épissage alternatif sont connues pour être spécifique au tissu et même au type cellulaire [191]. Nos analyses ont montré que nos résultats sur l'interaction entre le gène *ADCY9* et le gène *CETP* sur l'expression et l'épissage différaient selon les tissus. Cependant, les tissus où cette interaction a été identifiée représentent un mélange de plusieurs types cellulaires, rendant impossible la détermination des lignées cellulaires spécifiques où cette interaction se produit. Des analyses de *single-cell* pourraient permettre de répondre à cette question. Cependant, dans les données actuellement disponibles, y compris la base de données du *Human Cell Atlas* [503], la puissance statistique pour étudier les isoformes, évaluer les interactions épistasiques et les effets spécifiques aux sexes demeure limitée.

6.6. Application de la méthodologie à d'autres gènes - Le cas des cytochromes P450

Dans le chapitre 2, nous avons développé une méthode permettant d'identifier les événements de co-évolution entre deux gènes spécifiques, à savoir les gènes *ADCY9* et *CETP*. Cette méthode repose sur le calcul du r^2 génotypique pour détecter le déséquilibre de liaison sur longue distance. Pour évaluer sa possible applicabilité à d'autres paires de gènes, nous l'avons utilisée pour étudier les gènes de la superfamille des *CYP450*.

Toujours dans le domaine de la pharmacogénomique, la famille des cytochromes P450 est très impliquée dans la réponse aux médicaments, de par leur faculté à métaboliser les substances provenant de l'extérieur du corps. Les membres de cette famille sont séparés en sous-familles, telles que la sous-famille cytochrome P450 3A (CYP3A). L'expansion de cette famille de gènes est causée par la duplication de gènes. Ces duplications peuvent entraîner l'inactivation de certains gènes et les convertir en pseudogènes [504]. Cependant, les duplications peuvent également donner lieu à des relations d'épistasie fonctionnelle, car elles peuvent engendrer un effet de redondance [202, 242]. En effet, cette redondance peut compenser la perte d'un gène par la présence d'un autre ayant une fonction similaire. Cela crée une relation d'épistasie négative, où la perte des deux gènes est nécessaire pour l'absence complète d'un phénotype [242].

Avec des analyses de génétique des populations, nous avons identifié des enrichissement de signaux de sélection (Figure 4.1) dans les sous-familles *CYP3A* et *CYP4F*, suggérant que les

gènes de ces sous-familles ont évolué ensemble. La sous-famille des *CYP3A* a principalement des signaux de sélection positive (Figure 4.2A), avec des différences entre les populations, mais également un signal de sélection balancé dans le gène ancestral *CYP3A43* dans la population africaine (Figure 4.3A). La sous-famille des *CYP4F* présente principalement des signaux de sélection balancée (Figure 4.3B).

La sous-famille des *CYP3A* a été générée à partir de multiples événements de duplication [505]. La méthodologie développée dans le chapitre 2, basée sur le déséquilibre de liaison, a été appliquée et à permis d'identifier de la co-évolution entre les membres de ces sous-familles. Étant donnée la proximité des gènes dans cette sous-famille, nous utilisons le terme LD inhabituel à la place de LRLD. Un LD inhabituel a été observé entre plusieurs paires de gènes de cette sous-famille, telles qu'entre *CYP3A5* et *CYP3A43* dans les populations européenne et africaine (Figures 4.4 et 4.9), ce qui pourrait suggérer des événements de co-évolution. Cette co-évolution semble être générée par de la co-régulation, où les mutations sous sélection dans le gène *CYP3A43* régulent significativement les niveaux d'expression du gène *CYP3A5* (Figure 4.5), situés à chaque extrémité de ce regroupement de gènes. De plus, les eQTL dans le gène *CYP3A43* sont les plus fortement associés que les signaux d'expression de la région promotrice précédant le gène *CYP3A5*. Le mécanisme par lequel cette régulation se fait n'est pas encore connu. Il se pourrait que ce soit un élément régulateur de la région promotrice, mais plus d'analyses fonctionnelles restent à être effectuées.

Un des objectifs principaux du deuxième projet, présenté au chapitre 4, visait à identifier les phénotypes qui pourraient être liés aux signatures de sélection que nous avions identifiées. Afin de découvrir leurs origines potentielles, nous avons effectué des analyses de PheWAS et de randomisation mendélienne à travers les deux sous-familles. À travers la majorité de la région génomique des *CYP3A*, nous avons observé de fortes associations avec le compte des réticulocytes et ce phénotype a également permis de découvrir une nouvelle relation causale avec l'expression de *CYP3A5* (Figure 4.6). De plus, plusieurs mutations présentant des signatures de pression de sélection dans la population africain étaient également significativement associée avec ce phénotype.

Cette association est intéressante en raison de certains facteurs environnementaux, principalement présents sur le continent africain. L'un de ces facteurs est associé à une maladie causée par un pathogène responsable de la malaria. Il existe cinq types de *Plasmodium*

causant la malaria chez l'humain, parmi lesquels *Plasmodium vivax* (*P.vivax*) affecte particulièrement les jeunes réticulocytes [453]. Originaire d'Afrique, cette espèce de *Plasmodium* a été confrontée à plusieurs mécanismes de résistance au sein de la population humaine, parmi lesquels le plus connu est le groupe sanguin Duffy-négatif [506]. Cette particularité se caractérise par l'absence des récepteurs Duffy sur les réticulocytes et les érythrocytes, qui sont normalement utilisés par le *P.vivax* pour entrer dans les cellules [506]. En conséquence, cela pourrait faire en sorte que la fréquence de ce parasite est moindre que dans les autres populations, comme l'Asie et les Amériques Latines. Néanmoins, des lacunes subsistent quant aux connaissances sur l'immunité acquise contre le *P.vivax*. Il est envisageable que l'association entre l'implication des réticulocytes par les *CYP3A* soit liée à cette résistance, mais cela nécessite des investigations plus approfondies.

6.6.1. Généralisation de la détection d'événements de co-évolution

Notre méthodologie a permis d'identifier deux phénomènes de co-évolution, soit un entre les gènes *ADCY9* et *CETP*, et l'autre dans les sous-familles des *CYP3A* et *CYP4F*. Cependant, cette approche reste encore très restreinte à des gènes ciblés.

Les multiples approches permettant la détection de signature de sélection utilisent l'environnement génomique autour des mutations afin de détecter ces signatures évolutives. Une prochaine étape serait de développer une statistique, potentiellement avec des approches d'apprentissage automatique, qui permettrait de détecter une région en co-évolution basée sur les motifs de déséquilibre de liaison sur un des loci. Même sans connaître les deux loci sous co-évolution, les analyses initiales ont détecté la présence de signature de sélection dans nos gènes, ce qui pourrait indiquer qu'il n'est pas nécessaire de pré-spécifier les deux loci pour identifier ces régions. Cependant, cela ne permet pas de départager si l'événement est lié à une co-évolution.

Avec une approche à l'aide de simulations, nous pourrions générer des jeux de données avec de l'épistasie, où nous pourrions intégrer une composante de valeur adaptative pour générer de la co-évolution, et peut-être même un effet spécifique au sexe. Avec ces jeux de données, on pourrait potentiellement générer une approche globale, utilisant de l'apprentissage automatique, qui déterminerait si une position est sous une pression co-évolutive grâce à

la détection de traces de sélection à proximité génomique de la mutation, ce qui permettrait de restreindre nos analyses de LRLD afin de trouver des paires de sites sous co-évolution.

6.7. Conclusion

Le corps humain est le résultat de son évolution, et notre compréhension restreinte de son évolution représente une limitation. Quand on le comprendra, on pourra mieux intervenir afin de le soigner. C'est pourquoi il est clair que la médecine évolutive joue un rôle crucial dans notre compréhension des mécanismes sous-jacents aux maladies.

Le premier projet présenté dans cette thèse portait sur la relation entre le gène *CETP*, qui est la cible thérapeutique du médicament dalcetrapib, et le gène *ADCY9*, où une mutation a été trouvée comme modulant la réponse à ce médicament. Cette relation entre ces deux gènes n'était pas bien définie. En utilisant des analyses de génétique des populations, nous avons observé que cette relation affectait spécifiquement les populations péruviennes, et que cette association variait en fonction du sexe. Nous avons également postulé que le mécanisme sous-jacent à cette relation pourrait être lié à la modulation de l'épissage alternatif, un mécanisme fréquemment impliqué dans les processus évolutifs. Les voies métaboliques impliquant cette interaction reste encore à être identifiées, mais nous avons émis l'hypothèse qu'elles pourraient être liées au stress oxydatif.

Dans le cadre de ce projet, nous avons développé une nouvelle méthodologie pour détecter les événements de co-évolution entre les gènes. Cette approche a également été appliquée dans un deuxième projet portant sur les gènes de la superfamille des cytochromes P450. Grâce à cette méthodologie, nous avons identifié des événements potentiels de co-évolution, ainsi que d'identifier une cause environnementale potentielle à cette co-évolution. En effet, les gènes *CYP3A5* et *CYP3A43* ont été identifiés comme co-évoluant dans la population africaine, avec des mutations dans le gène *CYP3A43* modulant l'expression du gène *CYP3A5*. Cette relation pourrait être liée aux mécanismes de résistance contre la malaria *Plasmodium vivax*, qui affecte les réticulocytes et pour lesquels l'expression du gène *CYP3A5* module le nombre de ces cellules.

Notre méthodologie a donc permis d'obtenir des résultats préliminaires, qui ont par la suite été approfondis en utilisant différentes données "-omiques", ce qui a permis d'obtenir une meilleure compréhension des relations entre les gènes.

Ce projet a contribué à une meilleure compréhension de gènes importants dans le domaine de la pharmacogénomique, en examinant leur évolution, leur régulation génétique et leurs effets sur les phénotypes. Cette avancée permet une meilleure compréhension de l'impact de ces gènes sur les traitements dans différentes populations, ouvrant ainsi la voie à des améliorations des traitements adaptés aux diverses populations humaines.

Références bibliographiques

[1] J. S. Papadopoulos and R. Agarwala. Cobalt: constraint-based alignment tool for multiple protein sequences. *Bioinformatics*, 23(9):1073–1079, 2007.

[2] G. Heldmaier and D. Werner. *Environmental Signal Processing and Adaptation*. Springer Berlin Heidelberg, 2002.

[3] J. Licinio and M.L. Wong. *Pharmacogenomics: The Search for Individualized Therapies*. Wiley, 2009.

[4] A.T. Presanna et al. Pharmacogenomics: the right drug to the right person. *Journal of clinical medicine research*, 1(4):191–194, 2009.

[5] C. J. Lord and A. Ashworth. The dna damage response and cancer therapy. *Nature*, 481(7381):287–294, 2012.

[6] M. P. Doogue and T. M. Polasek. The ABCD of clinical pharmacokinetics. *Therapeutic advances in drug safety*, 4(1):5–7, 2013.

[7] S.C. Khojasteh, H. Wong, and C.E.C.A. Hop. *Drug Metabolism and Pharmacokinetics Quick Guide*. SpringerLink : Bücher. Springer New York, 2011.

[8] C. D. Bruno et al. Effect of lipophilicity on drug distribution and elimination: Influence of obesity. *British Journal of Clinical Pharmacology*, 87(8):3197–3205, 2021.

[9] M. Ingelman-Sundberg. Human drug metabolising cytochrome P450 enzymes: properties and polymorphisms. *Naunyn-Schmiedeberg's archives of pharmacology*, 369(1):89–104, 2004.

[10] U.M. Zanger and M. Matthias Schwab. Cytochrome p450 enzymes in drug metabolism: Regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacology & Therapeutics*, 138(1):103–141, 2013.

[11] S. Crettol, N. Petrovic, and M. Murray. Pharmacogenetics of phase i and phase ii drug metabolism. *Current pharmaceutical design*, 16(2):204–219, 2010.

[12] M. A. Cerny. Prevalence of Non-Cytochrome P450-Mediated Metabolism in Food and Drug Administration-Approved Oral and Intravenous Drugs: 2006-2015. *Drug Metabolism & Disposition*, 44(8):1246–1252, 2016.

[13] W. E. Evans and M. V. Relling. Pharmacogenomics: translating functional genomics into rational therapeutics. *Science*, 286(5439):487–491, 1999.

[14] L. S. Klyushova, M. L. Perepechaeva, and A. Y. Grishanova. The role of cyp3a in health and disease. *Biomedicines*, 10(11), 2022.

[15] M. Ingelman-Sundberg. Genetic polymorphisms of cytochrome P450 2D6 (CYP2D6): clinical consequences, evolutionary aspects and functional diversity. *The pharmacogenomics journal*, 5(1):6–13, 2005.

[16] A. Mahgoub et al. Polymorphic hydroxylation of debrisoquine in man. *Lancet (London, England)*, 2(8038):584–586, 1977.

[17] C. Taylor et al. A review of the important role of cyp2d6 in pharmacogenomics. *Genes (Basel)*, 11(11), 2020.

[18] H. Qiu et al. Cyp3 phylogenomics: evidence for positive selection of cyp3a4 and cyp3a7. *Pharmacogenetics and Genomics*, 18(1):53–66, 2008.

[19] A. G. McArthur et al. Phylogenetic Analysis of the Cytochrome P450 3 (CYP3) Gene family. *Journal of Molecular Evolution*, 57(2):200–211, 2003.

[20] P. Kuehl et al. Sequence diversity in cyp3a promoters and characterization of the genetic basis of polymorphic cyp3a5 expression. *Nature Genetics*, 27(4):383–391, 2001.

[21] O. Burk and L. Wojnowski. Cytochrome p450 3a and their regulation. *Naunyn-Schmiedeberg's Archives of Pharmacology*, 369(1):105–124, 2004.

[22] K. Z. Edson and A. E. Rettie. Cyp4 enzymes as potential drug targets: focus on enzyme multiplicity, inducers and inhibitors, and therapeutic modulation of 20-hydroxyeicosatetraenoic acid (20-hete) synthase and fatty acid ω-hydroxylase activities. *Curr Top Med Chem*, 13(12):1429–1440, 2013.

[23] Y. B. Jarrar and S.-J. Lee. Molecular functionality of cytochrome p450 4 (cyp4) genetic polymorphisms and their clinical implications. *International Journal of Molecular Sciences*, 20(17), 2019.

[24] J.C. Tardif et al. Pharmacogenomic determinants of the cardiovascular effects of dalcetrapib. *Circulation: Cardiovascular Genetics*, 8(2):372–382, 2015.

[25] L. Larifla. *Athérosclérose, hypertension, thrombose*. Abrégés. Modules transversaux. Masson, 2002.

[26] M.F. Linton, P.G. Yancey, S.S. Davies, W.G.J. Jerome, E.F. Linton, and K.C. Vickers. The role of lipids and lipoproteins in atherosclerosis, 2019.

[27] J.R. Harris. *Cholesterol Binding and Cholesterol Transport Proteins:: Structure and Function in Health and Disease*. Subcellular Biochemistry. Springer Netherlands, 2010.

[28] M. Rysz-Górzyńska and M. Banach. Subfractions of high-density lipoprotein (HDL) and dysfunctional HDL in chronic kidney disease patients. *Archives of medical science : AMS*, 12(4):844–849, 2016.

[29] F. M. Sacks et al. Protein-defined subspecies of hdls (high-density lipoproteins) and differential risk of coronary heart disease in 4 prospective studies. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 40(11):2714–2727, 2020.

[30] P. Piko et al. The profile of hdl-c subfractions and their association with cardiovascular risk in the hungarian general and roma populations. *Scientific reports*, 12(1):10915, 2022.

[31] A. Kontush. HDL particle number and size as predictors of cardiovascular disease. *Frontiers in pharmacology*, 6:218, 2015.

[32] R. K. Mutharasan et al. HDL efflux capacity, HDL particle size, and high-risk carotid atherosclerosis in a cohort of asymptomatic older adults: the Chicago Healthy Aging Study. *Journal of lipid research*, 58(3):600–606, 2017.

[33] S. Fazio and N. Pamir. HDL Particle Size and Functional Heterogeneity. *Circulation research*, 119(6):704–707, 2016.

[34] S. A. Didichenko et al. Enhanced HDL Functionality in Small HDL Species Produced Upon Remodeling of HDL by Reconstituted HDL, CSL112 : Effects on Cholesterol Efflux, Anti-Inflammatory and Antioxidative Activity. *Circulation research*, 119(6):751–763, 2016.

[35] Y. Chen et al. Evacetrapib reduces preβ-1 hdl in patients with atherosclerotic cardiovascular disease or diabetes. *Atherosclerosis*, 285:147–152, 2019.

[36] H. Shinkai. Cholesteryl ester transfer-protein modulator and inhibitors and their potential for the treatment of cardiovascular diseases. *Vasc Health Risk Manag*, 8:323–331, 2012.

[37] M.R. Wilkins. *Cardiovascular Pharmacogenetics*. Handbook of Experimental Pharmacology. Springer Berlin Heidelberg, 2013.

[38] R. Arakawa et al. Pharmacological inhibition of abca1 degradation increases hdl biogenesis and exhibits antiatherogenesis. *Journal of Lipid Research*, 50(11):2299–2305, 2009.

[39] L. Yang et al. Lcat- targeted therapies: Progress, failures and future. *Biomedicine & Pharmacotherapy*, 147:112677, 2022.

[40] Y. Wang and al. Plasma cholesteryl ester transfer protein is predominantly derived from kupffer cells. *Hepatology*, 62(6):1710–1722, 2015.

[41] K. G. Santana and al. Cholesterol-ester transfer protein alters m1 and m2 macrophage polarization and worsens experimental elastase-induced pulmonary emphysema. *Frontiers in Immunology*, 12, 2021.

[42] A. Inazu and al. Alternative splicing of the mrna encoding the human cholesteryl ester transfer protein. *Biochemistry*, 31(8):2352–2358, 1992.

[43] Y. Liu, D. Mihna, L. Izem, and R.E. Morton. Both full length-cholesteryl ester transfer protein and exon 9-deleted cholesteryl ester transfer protein promote triacylglycerol storage in cultured hepatocytes. *Lipids*, 57(1):69–79, 2022.

[44] L. Lagrost. Regulation of cholesteryl ester transfer protein (cetp) activity: review of in vitro and in vivo studies. *Biochimica et Biophysica Acta (BBA) - Lipids and Lipid Metabolism*, 1215(3):209 – 236, 1994.

[45] J. J. Albers, J. H. Tollefson, G. Wolfbauer, and R. E. Jr. Albright. Cholesteryl ester transfer protein in human brain. *International journal of clinical & laboratory research*, 21(3):264–266, 1992.

[46] X. C. Jiang et al. Dietary cholesterol increases transcription of the human cholesteryl ester transfer protein gene in transgenic mice. dependence on natural flanking sequences. *The Journal of clinical investigation*, 90(4):1290–1295, 1992.

[47] H. C. Oliveira et al. Human cholesteryl ester transfer protein gene proximal promoter contains dietary cholesterol positive responsive elements and mediates expression in small intestine and periphery while predominant liver and spleen expression is controlled by 5'-distal sequences. cis-acting sequences mapped in transgenic mice. *The Journal of clinical investigation*, 271(50):31831–31838, 1996.

[48] F. Apostolou et al. Acute infection with epstein-barr virus is associated with atherogenic lipid changes. *Atherosclerosis*, 212(2):607–613, 2010.

[49] L. Izem, D.J. Greene, K. Bialkowska, and R.E. Morton. Overexpression of full-length cholesteryl ester transfer protein in sw872 cells reduces lipid accumulation. *Journal of Lipid Research*, 56(3):515–525, 2015.

[50] F. Oestereich et al. The cholesteryl ester transfer protein (cetp) raises cholesterol levels in the brain. *Journal of lipid research*, 69(9):100260, 2022.

[51] L. Izem and R. E. Morton. Possible role for intracellular cholesteryl ester transfer protein in adipocyte lipid metabolism and storage. *Journal of Biological Chemistry*, 282(30):21856–21865, 2007.

[52] Z. Zhang and al. Expression of cholesteryl ester transfer protein in human atherosclerotic lesions and its implication in reverse cholesterol transport. *Atherosclerosis*, 159(1):67–75, 2001.

[53] A. Thompson et al. Association of Cholesteryl Ester Transfer Protein Genotypes With CETP Mass and Activity, Lipid Levels, and Coronary Risk. *JAMA*, 299(23):2777–2788, 2008.

[54] H. C. Oliveira and E. C. de Faria. Cholesteryl ester transfer protein: the controversial relation to atherosclerosis and emerging new biological roles. *IUBMB Life*, 63(4):248–257, 2011.

[55] C. M. Grion et al. Lipoproteins and cetp levels as risk factors for severe sepsis in hospitalized patients. *European journal of clinical investigation*, 40(4):330–338, 2010.

[56] A. C. Reisinger et al. Impact of sepsis on high-density lipoprotein metabolism. *Frontiers in cell and developmental biology*, 9:795460, 2021.

[57] L. L. Blauw, Y. Wang, K.W.v. Dijk, and P.C.N. Rensen. A novel role for cetp as immunological gatekeeper: Raising hdl to cure sepsis? *Trends in Endocrinology & Metabolism*, 31(5):334–343, 2020.

[58] M. Terán-García, J. P. Després, A. Tremblay, and C. Bouchard. Effects of cholesterol ester transfer protein (cetp) gene on adiposity in response to long-term overfeeding. *Atherosclerosis*, 196(1):455–460, 2008.

[59] Y. F. Wang et al. Cetp/lpl/lipc gene polymorphisms and susceptibility to age-related macular degeneration. *Scientific Reports*, 5(1):15711, 2015.

[60] A. Cougnard-Grégoire et al. Elevated high-density lipoprotein cholesterol and age-related macular degeneration: The alienor study. *PLOS ONE*, 9(3):1–11, 2014.

[61] C. C. Paun et al. Genetic variants and systemic complement activation levels are associated with serum lipoprotein levels in age-related macular degeneration. *Investigative ophthalmology & visual science*, 56(13):7766–7773, 2015.

[62] R. Liutkeviciene et al. Associations of cholesteryl ester transfer protein (cetp) gene variants with predisposition to age-related macular degeneration. *Gene*, 636:30–35, 2017.

[63] C. Bruce, D.S. Sharp, and A.R. Tall. Relationship of hdl and coronary heart disease to a common amino acid polymorphism in the cholesteryl ester transfer protein in men with and without hypertriglyceridemia. *Journal of Lipid Research*, 39(5):1071–1078, 1998.

[64] F. Cunningham et al. Ensembl 2022. *Nucleic Acids Research*, 50(D1):D988–D995, 2022.

[65] L. Izem, Y. Liu, and R.E. Morton. Exon 9-deleted cetp inhibits full length-cetp synthesis and promotes cellular triglyceride storage. *Journal of Lipid Research*, 61(3):422–431, 2020.

[66] M. E. Lira, A. K. Loomis, S. A. Paciga, D. B. Lloyd, and J. F. Thompson. Expression of cetp and of splice variants induces the same level of er stress despite secretion efficiency differences. *Journal of Lipid Research*, 49(9):1955–1962, 2008.

[67] T. P. Yang, L. B. Agellon, A. Walsh, J. L. Breslow, and A. R. Tall. Alternative splicing of the human cholesteryl ester transfer protein gene in transgenic mice: Exon exclusion modulates gene expression in response to dietary or developmental change. *Journal of Biological Chemistry*, 271(21):12603–12609, 1996.

[68] K. Tsutsumi, A. Hagi, and Y. Inoue. The relationship between plasma high density lipoprotein cholesterol levels and cholesteryl ester transfer protein activity in six species of healthy experimental animals. *Biological & Pharmaceutical Bulletin*, 24(5):579–581, 2001.

[69] D.J. Zea, H. Richard, and E. Laine. Ases: visualizing evolutionary conservation of alternative splicing in proteins. *Bioinformatics*, 38(9):2615–2616, 2022.

[70] C. A. Hogarth, A. Roy, and D. L. Ebert. Genomic evidence for the absence of a functional cholesteryl ester transfer protein gene in mice and rats. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology*, 135(2):219–229, 2003.

[71] R. Albalat and C. Cañestro. Evolution by gene loss. *Nature Reviews Genetics*, 17(7):379–391, 2016.

[72] A. Pirillo, A. L. Catapano, and G.D. Norata. *HDL in Infectious Diseases and Sepsis*. Springer International Publishing, 2015.

[73] F. Matsuura, N. Wang, W. Chen, X. C. Jiang, and A. R. Tall. Hdl from cetp-deficient subjects shows enhanced ability to promote cholesterol efflux from macrophages in an apoe- and abcg1-dependent pathway. *The Journal of clinical investigation*, 116(5):1435–1442, 2006.

[74] L.T. Nordestgaard et al. Long-term Benefits and Harms Associated With Genetic Cholesteryl Ester Transfer Protein Deficiency in the General Population. *JAMA Cardiology*, 7(1):55–64, 2022.

[75] C.J. Fielding and R.J. Havel. Cholesteryl ester transfer protein: friend or foe? *The Journal of clinical investigation*, 97(12):2987–2688, 1996.

[76] P. J. Barter et al. Effects of torcetrapib in patients at high risk for coronary events. *New England Journal of Medicine*, 357(21):2109–2122, 2007.

[77] G. G. Schwartz et al. Effects of dalcetrapib in patients with a recent acute coronary syndrome. *New England Journal of Medicine*, 367(22):2089–2099, 2012.

[78] The HPS3TIMI55–REVEAL Collaborative Group. Effects of anacetrapib in patients with atherosclerotic vascular disease. *New England Journal of Medicine*, 377(13):1217–1227, 2017.

[79] A. M. Lincoff et al. Evacetrapib and cardiovascular outcomes in high-risk vascular disease. *New England Journal of Medicine*, 376(20):1933–1942, 2017.

[80] G. K. Hovingh, J. J. Kastelein, S. J. van Deventer, and al. Cholesterol ester transfer protein inhibition by ta-8995 in patients with mild dyslipidaemia (tulip): a randomised, double-blind, placebo-controlled phase 2 trial. *Lancet*, 286(9992):452–460, 2015.

[81] S. J. Nicholls and al. Lipid lowering effects of the cetp inhibitor obicetrapib in combination with high-intensity statins: a randomized phase 2 trial. *Nature Medicine*, 28(8):1672–1678, 2022.

[82] C. M. Ballantyne et al. Obicetrapib plus ezetimibe as an adjunct to high-intensity statin therapy: A randomized phase 2 trial. *Journal of Clinical Lipidology*, 2023.

[83] J.D. Furtado et al. Pharmacological inhibition of cetp (cholesteryl ester transfer protein) increases hdl (high-density lipoprotein) that contains apoc3 and other hdl subspecies associated with higher risk of coronary heart disease. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 42(2):227–237, 2022.

[84] S.J. Nicholls et al. Comparative effects of cholesteryl ester transfer protein inhibition, statin or ezetimibe on lipid factors: The accentuate trial. *Atherosclerosis*, 261:12–18, 2017.

[85] J. C. Tardif, M. A. Pfeffer, S. Kouz, and al. Pharmacogenetics-guided dalcetrapib therapy after an acute coronary syndrome: the dal-GenE trial. *European Heart Journal*, 43(39):3947–3956, 2022.

[86] M. D. Solomon, E. J. McNulty, J. S. Rana, and al. The covid-19 pandemic and the incidence of acute myocardial infarction. *New England Journal of Medicine*, 383(7):691–693, 2020.

[87] S. Bhatt Ankeet, A. Moscone, E. McElrath Erin, and al. Fewer hospitalizations for acute cardiovascular conditions during the covid-19 pandemic. *Journal of the American College of Cardiology*, 76(3):280–288, 2020.

[88] S.E. Nissen et al. Adcy9 genetic variants and cardiovascular outcomes with evacetrapib in patients with high-risk vascular disease. *JAMA Cardiology*, 3(5):401–408, 2018.

[89] Y. Rautureau et al. Adcy9 (adenylate cyclase type 9) inactivation protects from atherosclerosis only in the absence of cetp (cholesteryl ester transfer protein). *Circulation*, 138(16):1677–1692, 2018.

[90] I. Gomes, J.H. Wardman, and S.D. Stockton. *Neuropeptide Receptors*. Colloquium Lectures on Neuropeptides. Biota Publishing, 2013.

[91] C. K. Billington and R. B. Penn. Signaling and regulation of g protein-coupled receptors in airway smooth muscle. *Respiratory research*, 4(1):2, 2003.

[92] H. Lodish et al. *Section 20.3, G Protein–Coupled Receptors and Their Effectors*. Molecular Cell Biology, New York, 4 edition, 2000.

[93] H. M. P. Teixeira, N. M. Alcantara-Neves, M. Barreto, C. A. Figueiredo, and R. S. Costa. Adenylyl cyclase type 9 gene polymorphisms are associated with asthma and

allergy in brazilian children. *Molecular Immunology*, 82:137 – 145, 2017.

[94] E. M. A. Slob et al. Pharmacogenetics of inhaled long-acting beta2-agonists in asthma: A systematic review. *Pediatric Allergy and Immunology*, 29(7):705–714, 2018.

[95] D. C. Mahadeo et al. A chemoattractant-mediated gi-coupled pathway activates adenylyl cyclase in human neutrophils. *Molecular biology of the cell*, 18(2):512–522, 2007.

[96] L. Liu, S. Das, W. Losert, and C. A. Parent. mtorc2 regulates neutrophil chemotaxis in a camp- and rhoa-dependent fashion. *Developmental cell*, 19(6):845–857, 2010.

[97] A. Ray and J.K. Kolls. Neutrophilic inflammation in asthma and association with disease severity. *Trends in immunology*, 38(12):942–954, 2017.

[98] Oganisation Mondial de la Santé. World malaria report, 2022.

[99] A. Manjurano et al. Candidate human genetic polymorphisms and severe malaria in a tanzanian population. *PLOS ONE*, 7(10):1–8, 2012.

[100] T. O. others Apinjoh. Association of candidate gene polymorphisms and tgf-beta/il-10 levels with malaria in three regions of cameroon: a case-control study. *Malaria journal*, 13:236–236, 2014.

[101] A. V. Khera et al. Genetic risk, adherence to a healthy lifestyle, and coronary disease. *New England Journal of Medicine*, 375, 2016.

[102] R.A. Fisher, J.H. Bennett, and H. Bennett. *The Genetical Theory of Natural Selection: A Complete Variorum Edition*. OUP Oxford, 1999.

[103] R. M. Nelson, M. E. Pettersson, and Ö. Carlborg. A century after fisher: time for a new paradigm in quantitative genetics. *Trends in Genetics*, 29(12):669–676, 2013.

[104] A. Collins. *Linkage Disequilibrium and Association Mapping: Analysis and Applications*. Methods in molecular biology. Humana Press, 2007.

[105] William S. Bush and Jason H. Moore. Chapter 11: Genome-wide association studies. *PLOS Computational Biology*, 8(12):1–11, 2012.

[106] G. R. Norman and D. L. Streiner. *Biostatistics: The Bare Essentials*. Pmph USA Ltd Series. B.C. Decker, 2008.

[107] I. Jolliffe. *Principal component analysis*. Springer, 2011.

[108] S.J. Hebbring. The challenges, advantages, and future of phenome-wide association studies. *Immunology*, 141, 2013.

[109] J.C. Denny, L. Bastarache, and D.M. Roden. Phenome-wide association studies as a tool to advance precision medicine. *Annual review of genomics and human genetics*, 17:353–373, 2016.

[110] C. A. Emdin, A. V. Khera, and S. Kathiresan. Mendelian Randomization. *JAMA*, 318(19):1925–1926, 2017.

[111] M.V. Holmes et al. Mendelian randomization of blood lipids for coronary heart disease. *European Heart Journal*, 36(9):539–550, 2014.

[112] T.G. Richardson et al. Evaluating the relationship between circulating lipoprotein lipids and apolipoproteins with risk of coronary heart disease: A multivariable mendelian randomisation analysis. *PLOS Medicine*, 17:1–22, 2020.

[113] S.L. Larsson, S. Burgess, A.M. Mason, and K. Michaëlsson. Alcohol consumption and cardiovascular disease. *Circulation: Genomic and Precision Medicine*, 13(3):e002814, 2020.

[114] S. Hägg et al. Adiposity as a cause of cardiovascular disease: a Mendelian randomization study. *International Journal of Epidemiology*, 44(2):578–586, 2015.

[115] C. Sudlow and al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine*, 12(3), 2015.

[116] GTEx Consortium. The genotype-tissue expression (gtex) project. *Nature Genetics*, 45(6):580–585, 2013.

[117] B.L. Pierce and S. Burgess. Efficient design for mendelian randomization studies: Subsample and 2-sample instrumental variable estimators. *American journal of epidemiology*, 178(7):1177–1184, 2013.

[118] A.R. Carter et al. Mendelian randomisation for mediation analysis: current methods and challenges for implementation. *European Journal of Epidemiology*, 36(5):465–478, 2021.

[119] M. Schwab. *Encyclopedia of Cancer*. Encyclopedia of Cancer. Springer Berlin Heidelberg, 2008.

[120] G. Davey Smith and G. Hemani. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Human Molecular Genetics*, 23(R1):R89–R98, 2014.

[121] S. Burgess, A. Butterworth, and S.G. Thompson. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic Epidemiology*, 37(7):658–665, 2013.

[122] Stephen Burgess and Simon G. Thompson. Interpreting findings from mendelian randomization using the MR-egger method. *European Journal of Epidemiology*, 32(5):377–389, 2017.

[123] J. Bowden, G. Davey Smith, P. C. Haycock, and S. Burgess. Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic epidemiology*, 40(4):304–314, 2016.

[124] S. Burgess and S. G. Thompson. Multivariable mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. *American journal of epidemiology*, 181(4):251–260, 2015.

[125] M. Hamilton. *Population Genetics*. Wiley, 2011.

[126] B. Hall. *Evolution: Principles and Processes*. Jones and Bartlett Topics in Biology. Jones & Bartlett Learning, 2011.

[127] J.L. Serre. *Génétique des populations*. Sciences de la vie. Dunod, 2006.

[128] C.A. Andrews. Natural selection, genetic drift, and gene flow do not act in isolation in natural population. *Nature Education Knowledge*, 3(10):5, 2010.

[129] K. Choi and I. Henderson. Meiotic recombination hotspots – a comparative view. *The Plant journal : for cell and molecular biology*, 83, 2015.

[130] M.M. Zdanowicz and American Society of Health-System Pharmacists. *Concepts in Pharmacogenomics*. EBSCO ebook academic collection. American Society of Health-System Pharmacists, 2010.

[131] C. Darwin, C.A. Roger, and V. Masson. *De l'origine des espèces ou Des lois du progrés chez les étres organisés*. Guillaumin et Cie., libraires-éditeurs, 1862.

[132] T. Lefevre, M. Raymond, and F. Thomas. *Biologie évolutive*. Biologie. De Boeck supérieur, 2016.

[133] J.C. Knight. *Human Genetic Diversity: Functional Consequences for Health and Disease*. OUP Oxford, 2009.

[134] C. Olito and C. de Vries. The demographic costs of sexually antagonistic selection in partially selfing populations. *The American Naturalist*, 200(3):401–418, 2022.

[135] B. Vicoso and B. Charlesworth. Evolution on the x chromosome: unusual patterns and processes. *Nature Reviews Genetics*, 7(8):645–653, 2006.

[136] W. R. Rice and A. K. Chippindale. Intersexual ontogenetic conflict. *Journal of Evolutionary Biology*, 14(5):685–693, 2001.

[137] M. M. Patten. The X chromosome favors males under sexually antagonistic selection. *Evolution*, 73(1):84–91, 2019.

[138] R. Bonduriansky and S. F. Chenoweth. Intralocus sexual conflict. *Trends in Ecology & Evolution*, 24(5):280–288, 2009.

[139] L. M. McIntyre and al. Sex-specific expression of alternative transcripts in drosophila. *Genome Biology*, 7(8), 2006.

[140] T. Day and R. Bonduriansky. Intralocus sexual conflict can drive the evolution of genomic imprinting. *Genetics*, 167(4):1537–1546, 2004.

[141] J. Vitti, S. Grossman, and P. Sabeti. Detecting natural selection in genomic data. *Annual Reviews Genetics*, 47:97–120, 2013.

[142] A. Bigham et al. Identifying signatures of natural selection in tibetan and andean populations using dense genome scan data. *PLOS Genetics*, 6(9):1–14, 2010.

[143] K.E. Holsinger and B.S. Weir. Genetics in geographically structured populations: defining, estimating and interpreting f(st). *Nature reviews Genetics*, 10(9):639–650, 2009.

[144] G. Bhatia, N. Patterson, S. Sankararaman, and A. L. Price. Estimating and interpreting fst: the impact of rare variants. *Genome Research*, 23(9):1514–1521, 2013.

[145] G. Chen, A. Yuan, D. Shriner, F. Tekola-Ayele, J. Zhou, A.R. Bentley, X. Zhou, C. Wang, M.J. Newport, A. Adeyemo, and C.N. Rotimi. An improved fst estimator. *PLOS ONE*, 10(8):1–15, 2015.

[146] C.C. Cockerham. Variance of gene frequencies. *Evolution*, 23(1):72–84, 1969.

[147] R.R. Hudson, M. Slatkin, and W.P. Maddison. Estimation of levels of gene flow from dna sequence data. *Genetics*, 132(2):583–589, 1992.

[148] R. Janha, A. Worwui, K. Linton, S. O Shaheen, F. Sisay-Joof, and R. Walton. Inactive alleles of cytochrome p450 2c19 may be positively selected in human evolution. *BMC evolutionary biology*, 14:71, 2014.

[149] M. Nelis et al. Genetic structure of europeans: a view from the north-east. *PLoS One*, 4(5):e5472, 2009.

[150] X. Yi, Y. Liang, E. Huerta-Sanchez, X. Jin, Z. Xi Ping Cuo, J. E Pool, X. Xu, H. Jiang, N. Vinckenbosch, T. Korneliussen, H. Zheng, T. Liu, W. He, K. Li, R. Luo, X. Nie, H. Wu, M. Zhao, H. Cao, and J. Wang. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science (New York, N.Y.)*, 329:75–8, 2010.

[151] J. Li et al. Global patterns of genetic diversity and signals of natural selection for human adme genes. *Human molecular genetics*, 20:528–540, 2010.

[152] P. C. Sabeti and al. Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419(6909):832–837, 2002.

[153] B. F. Voight, S. Kudaravalli, X. Wen, and J. K. Pritchard. A map of recent positive selection in the human genome. *PLoS biology*, 4(3), 2006.

[154] E. Koch, M. Ristroph, and M. Kirkpatrick. Long range linkage disequilibrium across the human genome. *PLOS ONE*, 8(12):1–10, 2013.

[155] M. Nei and W.-H. Li. Linkage disequilibrium in subdivided populations. *Genetics*, 75(1):213–219, 1973.

[156] P.C. Phillips. Epistasis - the essential role of gene interactions in the structure and evolution of genetic systems. *Nature reviews. Genetics*, 9:855–867, 2008.

[157] B. K. Maples, S. Gravel, E. E. Kenny, and C. D. Bustamante. Rfmix: a discriminative modeling approach for rapid and robust local-ancestry inference. *American journal of human genetics*, 93(2):278–288, 2013.

[158] R. C. Lewontin and Ken ichi Kojima. The evolutionary dynamics of complex polymorphisms. *Evolution*, 14(4):458–472, 1960.

[159] R. V. Rohlfs, W. J. Swanson, and B. S. Weir. Detecting coevolution through allelic association between physically unlinked loci. *The American Journal of Human Genetics*, 86(5):674–685, 2010.

[160] S. Nurk et al. The complete sequence of a human genome. *Science*, 376(6588):44–53, 2022.

[161] C. Buccitelli and M. Selbach. mrnas, proteins and the emerging principles of gene expression control. *Nature Reviews Genetics*, 21(10):630–644, 2020.

[162] R. Lowe, N. Shirley, M. Bleackley, S. Dolan, and T. Shafee. Transcriptomics technologies. *PLOS Computational Biology*, 13(5):1–23, 2017.

[163] Z. Wang, M. Gerstein, and M. Snyder. Rna-seq: A revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10:57–63, 2008.

[164] J. Jiang et al. Whole transcriptome analysis with sequencing: Methods, challenges and potential solutions. *Cellular and molecular life sciences : CMLS*, 72, 2015.

[165] A. Sarkar, K. Panati, and V.R. Narala. Code inside the codon: The role of synonymous mutations in regulating splicing machinery and its impact on disease. *Mutation Research/Reviews in Mutation Research*, 790:108444, 2022.

[166] D. L. Bentley. Rules of engagement: co-transcriptional recruitment of pre-mrna processing factors. *Current Opinion in Cell Biology*, 16(3):251–256, 2005.

[167] H. Khatter, M. K. Vorländer, and C. W. Müller. Rna polymerase i and iii: similar yet unique. *Current Opinion in Structural Biology*, 47:88–94, 2017.

[168] J. R. Warner. The economics of ribosome biosynthesis in yeast. *Trends in biochemical sciences*, 24(11):437–440, 1999.

[169] S. Hahn. Structure and mechanism of the rna polymerase ii transcription machinery. *Nature Structural & Molecular Biology*, 11(5):394–403, 2004.

[170] T. K. Kim and R. Shiekhattar. Architectural and functional commonalities between enhancers and promoters. *Cell*, 162(5):948–959, 2015.

[171] E. Brasset and C. Vaury. Insulators are fundamental components of the eukaryotic genomes. *Heredity*, 94(6):571–576, 2005.

[172] B. Pang, J. H. van Weerd, F. L. Hamoen, and M. P. Snyder. Identification of non-coding silencer elements and their regulation of gene expression. *Nature Reviews Molecular Cell Biology*, 2022.

[173] Y. Wang, J. Liu, B. O. Huang, and al. Mechanism of alternative splicing and its regulation. *Biomedical reports*, 3(2):152–158, 2015.

[174] O. Kelemen, P. Convertini, Z. Zhang, and al. Function of alternative splicing. *Gene*, 514(1):1–13, 2013.

[175] G.S. Wang and T.A. Cooper. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nature Reviews Genetics*, 8(10):749–761, 2007.

[176] J. Southby, C. Gooding, and C. W. J. Smith. Polypyrimidine tract binding protein functions as a repressor to regulate alternative splicing of α-actinin mutally exclusive exons. *Molecular and Cellular Biology*, 19(4):2699–2711, 1999.

[177] C. Gooding, F. Clark, M. C. Wollerton, S.-N. Grellscheid, H. Groom, and C. W. J. Smith. A class of human exons with predicted distant branch points revealed by analysis of ag dinucleotide exclusion zones. *Genome Biology*, 7(1), 2006.

[178] M. C. Wahl, C. L. Will, and R. Lührmann. The spliceosome: Design principles of a dynamic rnp machine. *Cell*, 136(4):701–718, 2009.

[179] U. Braunschweig, S. Gueroussov, A. M. Plocik, B. R. Graveley, and B. J. Blencowe. Dynamic integration of splicing within gene regulatory pathways. *Cell*, 152(6):1252–1269, 2013.

[180] A. R. Kornblihtt. Coupling transcription and alternative splicing. *Advances in experimental medicine and biology*, 623:175–189, 2007.

[181] H. Keren, G. Lev-Maor, and G. Ast. Alternative splicing and evolution: diversification, exon definition and function. *Nature Reviews Genetics*, 11(5):345–355, 2010.

[182] L. F. Lareau, A. N. Brooks, D. A. Soergel, Q. Meng, and S. E. Brenner. The coupling of alternative splicing and nonsense-mediated mrna decay. *Advances in experimental medicine and biology*, 623:190–211, 2007.

[183] R. F. Luco and T. Misteli. More than a splicing code: integrating the role of rna, chromatin and non-coding rna in alternative splicing regulation. *Current opinion in genetics & development*, 21(4):366–372, 2011.

[184] H. Sun and L. A. Chasin. Multiple splicing defects in an intronic false exon. *Molecular and cellular biology*, 20(17):6414–6425, 2000.

[185] Z. Wang and C. B. Burge. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA*, 14(5):802–813, 2008.

[186] S. Datta and D. Nettleton. *Statistical Analysis of Next Generation Sequencing Data*. Frontiers in Probability and the Statistical Sciences. Springer International Publishing, 2014.

[187] J. C. Long and J. F. Caceres. The sr protein family of splicing factors: master regulators of gene expression. *Biochemical journal*, 417(1):15–27, 2009.

[188] R. Martinez-Contreras, P. Cloutier, L. Shkreta, and al. hnrnp proteins and splicing control. *Advances in experimental medicine and biology*, 623:123–147, 2007.

[189] S. Sharma, A. M. Falick, and D. L. Black. Polypyrimidine tract binding protein blocks the 5' splice site-dependent assembly of u2af and the prespliceosomal e complex. *Molecular cell*, 19(4):485–496, 2006.

[190] Y. Barash, J. A. Calarco, W. Gao, and al. Deciphering the splicing code. *Nature*, 465(7294):53–59, 2010.

[191] E. Park et al. The expanding landscape of alternative splicing variation in human populations. *The American Journal of Human Genetics*, 102(1):11–26, 2018.

[192] C.L. Will and R. Lührmann. Spliceosome structure and function. *Cold Spring Harbor Perspectives in Biology*, 3(7):a003707, 2011.

[193] A.K. Aksaas et al. Protein kinase a–dependent phosphorylation of serine 119 in the proto-oncogenic serine/arginine-rich splicing factor 1 modulates its activity as a splicing enhancer protein. *Genes & Cancer*, 2(8):841–851, 2011.

[194] K. Colwill et al. The clk/sty protein kinase phosphorylates sr splicing factors and regulates their intranuclear distribution. *The EMBO journal*, 15(2):265–275, 1996.

[195] J. Shi et al. Cyclic amp-dependent protein kinase regulates the alternative splicing of tau exon 10. *The Journal of Biological Chemistry*, 286(16):14639–14648, 2011.

[196] Y. Zhang, J. Qian, C. Gu, and Y. Yang. Alternative splicing and cancer: a systematic review. *Signal Transduction and Targeted Therapy*, 6(1):78, 2021.

[197] B.L. Robberson, G.J. Cote, and S.M. Berget. Exon definition may facilitate splice site selection in rnas with multiple exons. *Molecular and Cellular Biology*, 10(1):84–94, 1990.

[198] M.V. Kotlajich, T.L. Crabb, and K.J. Hertel. Spliceosome assembly pathways for different types of alternative splicing converge during commitment to splice site pairing in the a complex. *Molecular and Cellular Biology*, 29(4):1072–1082, 2009.

[199] E. T. Wang, R. Sandberg, S. Luo, and al. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):47–476, 2008.

[200] L. Cartegni, S. L. Chew, and A. R. Krainer. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nature Reviews Genetics*, 3(4):285–298, 2002.

[201] Y. Marquez et al. Unmasking alternative splicing inside protein-coding exons defines exitrons and their role in proteome plasticity. *Genome Research*, 25(7):995–1007, 2015.

[202] C.J. Wright, C.W.J. Smith, and C.D. Jiggins. Alternative splicing as a source of phenotypic diversity. *Nature Reviews Genetics*, 23(22):697–710, 2022.

[203] L. Martinez-Gomez et al. Few sines of life: Alu elements have little evidence for biological relevance despite elevated translation. *NAR Genomics and Bioinformatics*, 2(1):lqz023, 2020.

[204] L. Lin et al. The contribution of alu exons to the human proteome. *Genome Biology*, 17(1):15, 2016.

[205] M.C. King and A. C. Wilson. Evolution at two levels in humans and chimpanzees. *Science*, 188(4184):107–116, 1975.

[206] R. H. Waterson, E. S. Lander, R. K. Wilson, Sequencing The Chimpanzee, and Consortium Analysis. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437(7055):69–87, 2005.

[207] G. Martín et al. Alternative splicing landscapes in arabidopsis thaliana across tissues and stress conditions highlight major functional differences with animals. *Genome Biology*, 22(1):35, 2021.

[208] B. Modrek and C.J. Lee. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nature Genetics*, 34(2):177–180, 2003.

[209] E.O. Gracheva et al. Ganglion-specific splicing of trpv1 underlies infrared sensation in vampire bats. *Nature*, 476(7358):88–91, 2011.

[210] R. Bumgarner. Overview of DNA microarrays: types, applications, and their future. *Current Protocols in Molecular Biology*, Chapter 22:Unit 22.1., 2013.

[211] M. Meyerson, S. Gabriel, and G. Getz. Advances in understanding cancer genomes through second-generation sequencing. *Nature reviews*, 11(10):685–696, 2010.

[212] N. Anaparthy, Y. J. Ho, L. Martelotto, M. Hammell, and J. Hicks. Single-cell applications of next-generation sequencing. *Cold Spring Harbor perspectives in medicine*, 9(10):a026898, 2019.

[213] S.L. Amarasinghe et al. Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*, 21(1):30, 2020.

[214] G. Gibson. The environmental contribution to gene expression profiles. *Nature reviews. Genetics*, 9(8):575–581, 2008.

[215] A. Dobin and al. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[216] L. Wan, X. Yan, T. Chen, and F. Sun. Modeling rna degradation for rna-seq with applications. *Biostatistics (Oxford, England)*, 13(4):734–747, 2012.

[217] A. Conesa et al. A survey of best practices for rna-seq data analysis. *Genome Biology*, 17(1):13, 2016.

[218] B. Li and C. N. Dewey. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC Bioinformatics*, 12(1):323, 2011.

[219] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter. Near-optimal probabilistic rna-seq quantification. *Nature Biotechnology*, 34(5):525–527, 2016.

[220] M. J. Oshlack, A. ADN Wakefield. Transcript length bias in rna-seq data confounds systems biology. *Biology direct*, 4:14, 2009.

[221] F. Rapaport et al. Comprehensive evaluation of differential gene expression analysis methods for rna-seq data. *Genome Biology*, 14(9):3158, 2013.

[222] M. D. Robinson and A. Oshlack. A scaling normalization method for differential expression analysis of rna-seq data. *Genome biology*, 11(3), 2010.

[223] C. W. Law, Y. Chen, W. Shi, and G. K. Smyth. voom: precision weights unlock linear model analysis tools for rna-seq read counts. *Genome Biology*, 15(2):R29, 2014.

[224] Y. Katz, E. T. Wang, E. M. Airoldi, and C. B. Burge. Analysis and design of rna sequencing experiments for identifying isoform regulation. *Nature Methods*, 7(12):1009–1015, 2010.

[225] C. Zhang, B. Zhang, L.-L. Lin, and S. Zhao. Evaluation and comparison of computational tools for rna-seq isoform quantification. *BMC Genomics*, 18(1):583, 2017.

[226] S. Anders, A. Reyes, and W. Huber. Detecting differential usage of exons from rna-seq data. *Genome Res*, 22(10):2008–2017, 2012.

[227] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[228] A. Mehmood et al. Systematic evaluation of differential splicing tools for RNA-seq studies. *Briefings in Bioinformatics*, 21(6):2052–2065, 2019.

[229] J. Vaquero-Garcia et al. A new view of transcriptome complexity and regulation through the lens of local splicing variations. *eLife*, 5:e11752, 2016.

[230] Y. I. Li et al. Annotation-free quantification of rna splicing using leafcutter. *Nature Genetic*, 50(1):151–158, 2018.

[231] M. Estefania, R. Andres, I. Javier, Y. Marcelo, and C. Ariel. ASpli: Integrative analysis of splicing landscapes through RNA-seq assays. *Bioinformatics (Oxford, England)*, page btab141, 2021.

[232] J. T. Leek and J. D. Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLOS Genetics*, 3(9):1–12, 2007.

[233] O. Stegle, L. Parts, R. Durbin, and J. Winn. A bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eqtl studies. *PLoS computational biology*, 6(5), 2010.

[234] E. Zeggini and A. Morris. *Analysis of Complex Disease Association Studies: A Practical Guide*. Elsevier Science, 2010.

[235] G. Karp, J. Isawa, and W. Marshall. *Biologie cellulaire et moléculaire*. HORS COLLECTION SCIENCES. De Boeck supérieur, 2018.

[236] G.S. Miglani. *Developmental Genetics*. I.K. International Publishing House Pvt. Limited, 2013.

[237] P.C. Phillips. The language of gene interaction. *Genetics*, 149(3):1167–1171, 1998.

[238] H.J. Cordell. Detecting gene-gene interactions that underlie human diseases. *Nature reviews*, 10(6):3920494, 2010.

[239] Y. Huang, S. Wuchty, and T. Przytycka. eqtl epistasis – challenges and computational approaches. *Frontiers in Genetics*, 4:51, 2013.

[240] S. Musani et al. Detection of gene-gene interactions in genome-wide association studies of human population data. *Human heredity*, 63:67–84, 2007.

[241] D. M. Weinreich, R. Watson, and L. Chao. Perspective: Sign epistasis and genetic constraint on evolutionary trajectories. *Evolution*, 59:1165 – 1174, 2007.

[242] B. Lehner. Molecular mechanisms of epistasis within and between genes. *Trends in Genetics*, 27(9):323–331, 2011.

[243] Sewall Wright. The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proceedings of the Sixth International Congress on Genetics*, 1:356–366, 1932.

[244] N. Johnson. Sewall wright and the development of shifting balance theory. *Nature Education*, 1(1):52, 2008.

[245] D. Futuyma and J. Antonovics. *Oxford Surveys in Evolutionary Biology*. Oxford University Press, 1992.

[246] L. Azevedo et al. Epistatic interactions: how strong in disease and evolution? *Trends in Genetics*, 22(11):581 – 585, 2006.

[247] J. Becker et al. A systematic eqtl study of cis–trans epistasis in 210 hapmap individuals. *European Journal Of Human Genetics*, 20:97, 2011.

[248] J. Shang et al. Performance analysis of novel methods for detecting epistasis. *BMC Bioinformatics*, 12(1):475, 2011.

[249] C. Niel, C. Sinoquet, C. Dina, and G. Rocheleau. A survey about methods dedicated to epistasis detection. *Frontiers in genetics*, 6:285, 2015.

[250] A. Strange et al. A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between hla-c and erap1. *Nat Genet*, 42(11):985–990, 2010.

[251] W Sadee. The relevance of "missing heritability " in pharmacogenomics. *Clinical Pharmacology & Therapeutics*, 92(4):428–430, 2012.

[252] 1000 Genomes Project Consortium and al. A global reference for human genetic variation. *Nature*, 526:68, 2015.

[253] T. Lappalainen et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501:506, 2013.

[254] P. Awadalla and al. Cohort profile of the cartagene study: Quebec's population-based biobank for public health and personalized genomics. *International Journal of Epidemiology*, 42(5):1285–1299, 2013.

[255] J.C. Tardif and al. Genotype-dependent effects of dalcetrapib on cholesterol efflux and inflammation. *Circulation. Cardiovascular Genetics*, 9(4):340–348, 2016.

[256] M. Poulter and al. The causal element for the lactase persistence/non-persistence polymorphism is located in a 1 mb region of linkage disequilibrium in europeans. *Annals of Human Genetics*, 67(4):298–311, 2003.

[257] T. Bersaglieri and al. Genetic signatures of strong recent positive selection at the lactase gene. *American Journal of Human Genetics*, 74(6):1111–1120, 2004.

[258] N.S. Enattah and al. Evidence of still-ongoing convergence evolution of the lactase persistence t-13910 alleles in humans. *American Journal of Human Genetics*, 81(3):615–625, 2007.

[259] Y. Itan, A. Powell, M.A. Beaumont, J. Burger, and M.G. Thomas. The origins of lactase persistence in europe. *PLOS Computational Biology*, 5(8):e1000491, 2009.

[260] C. Gamba, E. Jones, M. Teasdale, and al. Genome flux and stasis in a five millennium transect of european prehistory. *Nature Communications*, 5(1):5257, 2014.

[261] V. Labrie and al. Lactase non-persistence is directed by dna variation-dependent epigenetic aging. *Nature structural & molecular biology*, 23(6):566–573, 2016.

[262] M. Fumagalli and al. Greenlandic inuit show genetic signatures of diet and climate adaptation. *Science*, 349(6254):1343, 2015.

[263] J.A. Hollenbach and al. Hla diversity, differentiation, and haplotype evolution in mesoamerican natives. *Human Immunology*, 62(4):378–390, 2001.

[264] M.E Blais and al. High frequency of hiv mutations associated with hla-c suggests enhanced hla-c-restricted ctl selective pressure associated with an aids-protective polymorphism. *Journal of immunology (Baltimore, Md. : 1950)*, 188(9):4663–4670, 2012.

[265] T. Tashi and al. Gain-of-function egln1 prolyl hydroxylase (phd2 d4e:c127s) in combination with epas1 (hif-2α) polymorphism lowers hemoglobin concentration in tibetan highlanders. *Journal of Molecular Medicine*, 95(6):665–670, 2017.

[266] P. Li and al. A regulatory insertion-deletion polymorphism in the fads gene cluster influences pufa and lipid profiles among chinese adults: a population-based study. *The American Journal of Clinical Nutrition*, 107(6):867–875, 2018.

[267] D. Meyer and al. A genomic perspective on hla evolution. *Immunogenetics*, 70(1):5–27, 2018.

[268] L.M. Reynolds and al. Fads genetic and metabolomic analyses identify the δ5 desaturase (fads1) step as a critical control point in the formation of biologically important lipids. *Scientific Reports*, 10(1):15873, 2020.

[269] J.E. Crawford and al. Natural selection on genes related to cardiovascular health in high-altitude adapted andeans. *American Journal of Human Genetics*, 101(5):752–767, 2017.

[270] S. Asgari and al. A positively selected fbn1 missense variant reduces height in peruvian individuals. *Nature*, 582(7811):234–239, 2020.

[271] Y. Luo and al. Early progression to active tuberculosis is a highly heritable trait driven by 3q23 in peruvians. *Nature Communications*, 10(1), 2019.

[272] D. Reich and al. Reconstructing native american population history. *Nature*, 488(7411):370–374, 2012.

[273] O. Stegle, L. Parts, M. Piipari, J. Winn, and R. Durbin. Using probabilistic estimation of expression residuals (peer) to obtain increased power and interpretability of gene expression analyses. *Nature protocols*, 7(3):500–507, 2012.

[274] L.P. Lemieux Perreault, S. Provost, M.A. Legault, A. Barhdadi, and Dubé M.P. pygenclean: efficient tool for genetic data clean up before association testing. *Bioinformatics (Oxford, England)*, 29(13):1704–1705, 2013.

[275] D. López-Terrada, S. W. Cheung, M. J. Finegold, and B. B. Knowles. Hep g2 is a hepatoblastoma-derived cell line. *Human Pathology*, 40(10):1512–1515, 2009.

[276] M. Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10–12, 2011.

[277] J. D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002.

[278] M. Muffato and al. Ensembl comparative genomics resources. *Database: The Journal of Biological Databases and Curation*, 2016, 2016.

[279] H. Li and al. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[280] C. A. Maclean and J. G. D. Chue Hong. hapbin: an efficient program for performing haplotype-based scans for positive selection in large genomic datasets. *Molecular Biology and Evolution*, 32(11):3027–3029, 2015.

[281] P. Danecek et al. The variant call format and vcftools. *Bioinformatics (Oxford, England)*, 27(15):2156–2158, 2011.

[282] D.N. Harris and al. Evolutionary genomic dynamics of peruvians before, during, and after the inca empire. *Proceedings of the National Academy of Sciences of the United States of America*, 115(28), 2018.

[283] W.H. Li and M. Nei. Stable linkage disequilibrium without epistasis in subdivided populations. *Theoretical Population Biology*, 6(2):173–183, 1974.

[284] L. Park. Population-specific long-range linkage disequilibrium in the human genome and its influence on identifying common disease variants. *Scientific Reports*, 9(1), 2019.

[285] M. Slatkin. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nature reviews. Genetics*, 9(6):477–485, 2008.

[286] B. Brynedal, J. Choj, R. Bjornson, B.E. Stranger, B.M. Neale, B.F. Voight, and C. Cotsapas. Large-scale trans-eqtls affect hundreds of transcripts and mediate patterns of transcriptional co-regulation. *American Journal of Human Genetics*, 100(4):581–591, 2017.

[287] C.M. Beall. Two routes to functional adaptation: Tibetan and andean high-altitude natives. *Proceedings of the National Academy of Sciences of the United States of America*, 104(Suppl 1):8655–8660, 2007.

[288] T.D Brutsaert, E.J. Parra, M.D. Shriber, A. Gamboa, M. Rivera-Ch, and F. León-Velarde. Ancestry explains the blunted ventilatory response to sustained hypoxia and lower exercise ventilation of quechua altitude natives. *American Journal of Physiology. Regulatory, Integrative and Comparative Physiology*, 289(1):R225–234, 2005.

[289] C.G. Julian and L.G. Moore. Human genetic adaptation to high altitude: evidence from the andes. *Genes*, 10(2):150, 2019.

[290] L.G. Moore. Human genetic adaptation to high altitudes: current status and future prospects. *Quaternary International: The Journal of the International Union for Quaternary Research*, 461:4–13, 2017.

[291] J.S. Milledge, J.B. West, and R.B. Schoene. *High altitude medicine and physiology*, volume 4. CRC Press, 2007.

[292] A.J. Murray. Energy metabolism and the high-altitude environment. *Experimental Physiology*, 101(1):23–27, 2016.

[293] T. Yamada, M. Kawata, H. Arai, M. Fukasawa, K. Inoue, and T. Sato. Astroglial localization of cholesteryl ester transfer protein in normal and alzheimer's disease brain tissues. *Acta Neuropathologica*, 90(6):633–636, 1995.

[294] E.A. Murphy et al. Cetp polymorphisms associate with brain structure, atrophy rate, and alzheimer's disease risk in an apoe-dependent manner. *Brain imaging and behavior*,

6(1):16–26, 2012.

[295] P. Lieberman, A. Morey, J. Hochstadt, M. Larson, and S. Mather. Mount everest: a space analogue for speech monitoring of cognitive deficits and stress. *Aviation, Space, and Environmental Medicine*, 76(6 Suppl):B198–207, 2005.

[296] B. Shukitt-Hale, M.J. Stillman, D.I. Welch, A. Levy, J.A. Devine, and H.R. Lieberman. Hypobaric hypoxia impairs spatial memory in an elevation-dependent fashion. *Behavioral and Neural Biology*, 62(3):244–252, 1994.

[297] J. Horiuchi, L.M. McDowall, and R.A.L. Dampney. Vasomotor and respiratory responses evoked from the dorsolateral periaqueductal grey are mediated by the dorsomedial hypothalamus. *The Journal of Physiology*, 587(21):5149–5162, 2009.

[298] K. Rahmouni. Cardiovascular regulation by the arcuate nucleus of the hypothalamus: neurocircuitry and signaling systems. *Hypertension*, 67(6):1064–1071, 2016.

[299] B. Peter and G.J. Simon R. Effect of altitude on the heart and the lungs. *Circulation*, 116(19):2191–2202, 2007.

[300] R. Hainsworth, M.J. Drinkhill, and M. Rivera-Chira. The autonomic nervous system at high altitude. *Clinical Autonomic Research: Official Journal of the Clinical Autonomic Research Society*, 17(1):13–19, 2007.

[301] A.W. Bigham and L.S. Lee. Human high-altitude adaptation: forward genetics meets the hif pathway. *Genes & Development*, 28(20):2189–2204, 2014.

[302] L.G. Moore. Measuring high-altitude adaptation. *Journal of Applied Physiology*, 123(5):1371–1385, 2017.

[303] H. Abe, H. Semba, and N. Takeda. The roles of hypoxia signaling in the pathogenesis of cardiovascular diseases. *Journal of Atherosclerosis and Thrombosis*, 24(9):884–894, 2017.

[304] J.W. Lee, J. Ko, C. Ju, and H.K. Eltzschig. Hypoxia signaling in human diseases and therapeutic targets. *Experimental & Molecular Medicine*, 51(6):1–13, 2019.

[305] D. Faeh, F. Gutzwiller, F. Bopp, and Swiss National Cohort Study Group. Lower mortality from coronary heart disease and stroke at higher altitudes in switzerland. *Circulation*, 120(6):495–501, 2009.

[306] R. Naeije. Physiological adaptation of the cardiovascular system to high altitude. *Progress in Cardiovascular Diseases*, 52(6):456–466, 2010.

[307] B. Ostadal and F. Kolar. Cardiac adaptation to chronic high-altitude hypoxia: Beneficial and adverse effects. *Respiratory Physiology & Neurobiology*, 158(2):224–236, 2007.

[308] C. J. Riley and M. Gavin. Physiological changes to the cardiovascular system at high altitude and its effects on cardiovascular disease. *High Altitude Medicine & Biology*, 18(2):102–113, 2017.

[309] J.J. Savla, B.D. Levine, and H.A. Sadek. The effect of hypoxia on cardiovascular disease: friend or foe? *High Altitude Medicine & Biology*, 19(2):124–130, 2018.

[310] H.A. Abhishekh, P. Nisarga, R. Kisan, A. Meghana, S. Chandran, and T. Raju. Influence of age and gender on autonomic regulation of heart. *Journal of Clinical Monitoring and Computing*, 27(3):259–264, 2013.

[311] A.M. Cart, X.J. Du, and B.A. Kingwell. Gender, sex hormones and autonomic nervous control of the cardiovascular system. *Cardiovascular Research*, 53(3):678–687, 2002.

[312] A.C. Nugent, E.E. Bain, J.F. Thayer, J.J. Sollers, and W.C. Drevets. Sex differences in the neural correlates of autonomic arousal: a pilot pet study. *International journal of psychophysiology : official journal of the International Organization of Psychophysiology*, 80(3):182–191, 2011.

[313] E.H. Morrow. Implications of sex-specific selection for the genetic basis of disease. *Evolutionary Applications*, 6(8):1208–1217, 2013.

[314] T. Connallon and A.G. Clark. Sex linkage, sex-specific selection, and the role of recombination in the evolution of sexually dimorphic gene expression. *Evolution; international journal of organic evolution*, 64(12):3417–3442, 2010.

[315] J. Parsch and H. Ellegren. The evolutionary causes and consequences of sex-biased gene expression. *Nature Reviews Genetics*, 14(2):83–87, 2013.

[316] T.M. Williams and S.B. Carroll. Genetic and molecular insights into the development and evolution of sexual dimorphism. *Nature Reviews Genetics*, 10(11):797–804, 2009.

[317] R.M. Cox and al. Hormonally mediated increases in sex-biased gene expression accompany the breakdown of between-sex genetic correlations in a sexually dimorphic lizard. *The American Naturalist*, 189(3):315–332, 2017.

[318] J.W. McGlothlin, R.M. Cox, and E.D.III Brodie. Sex-specific selection and the evolution of between-sex genetic covariance. *Journal of Heredity*, 110(4):422–432, 2019.

[319] G.L. Wojcik and al. Genetic analyses of diverse populations improves discovery for complex traits. *Nature*, 570(7762):514–518, 2019.

[320] F. Aguet and al. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213, 2017.

[321] M.P. Metzinger et al. Effect of anacetrapib on cholesterol efflux capacity: a substudy of the define trial. *Journal of the American Heart Association*, 9(24):e018136, 2020.

[322] P. Deelen and al. Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration. *BMC Research Notes*, 7(1):901, 2014.

[323] S. McCarthy and al. A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*, 48(/0):1279–1283, 2016.

[324] O. Delaneau, J.F. Zagury, and J. Marchini. Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods*, 10(1):5–6, 2013.

[325] R. Durbin. Efficient haplotype matching and storage using the positional burrows–wheeler transform (pbwt). *Bioinformatics*, 30(9):1266–1272, 2014.

[326] S. Purcell et al. Plink: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics*, 81, 2007.

[327] J. McInnes, J. Healy, and J. Melville. Umap: uniform manifold approximation and projection for dimension reduction. *arXiv:1802.03426 [cs, stat]*, 2020.

[328] J.G. Hussin and al. Recombination affects accumulation of damaging and disease-associated mutations in human populations. *Nature Genetics*, 47(4):400–404, 2015.

[329] M.J. Favé and al. Gene-by-environment interactions in urban populations modulate risk phenotypes. *Nature Communications*, 9, 2018.

[330] K.R. Kasimatis, P.L. Ralph, and P.C. Phillips. Limits to genomic divergence under sexually antagonistic selection. *G3: Genes, Genomes, Genetics*, 9(11):3813–3824, 2019.

[331] G. Abraham and M. Inouye. Fast principal component analysis of large-scale genome-wide data. *PLOS ONE*, 9(4):e93766, 2014.

[332] M. Ritchie et al. Limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43, 2015.

[333] J. T. Haas and B. Staels. Cholesteryl-ester transfer protein (CETP): A kupffer cell marker linking hepatic inflammation with atherogenic dyslipidemia? *Hepatology*, 62(6):1659–1661, 2015.

[334] Y. Luo and A. R. Tall. Sterol upregulation of human CETP expression in vitro and in transgenic mice by an LXR element. *The Journal of Clinical Investigation*, 105(4):513–520, 2000.

[335] A. Skoczyńska et al. Serum lipid transfer proteins in hypothyreotic patients are inversely correlated with thyroid-stimulating hormone (TSH) levels. *Medical Science Monitor : International Medical Journal of Experimental and Clinical Research*, 22:4661–4669, 2016.

[336] J. A. others Berti. Cholesteryl ester transfer protein expression is down-regulated in hyperinsulinemic transgenic mice. *Journal of Lipid Research*, 44(10):1870–1876, 2003.

[337] H. Hayashibe et al. Increased plasma cholesteryl ester transfer activity in obese children. *Atherosclerosis*, 129(1):53–58, 1997.

[338] F. Apostolou, I. F. Gazi, K. Lagos, C. C. Tellis, A. D. Tselepis, E. N. Liberopoulos, and M. Elisaf. Acute infection with epstein–barr virus is associated with atherogenic lipid changes. *Atherosclerosis*, 212(2):607–613, 2010.

[339] D. J. Greene, L. Izem, and R. E. Morton. Defective triglyceride biosynthesis in CETP-deficient SW872 cells. *Journal of Lipid Research*, 56(9):1669–1678, 2015.

[340] L. Izem and R. E. Morton. Cholesteryl ester transfer protein biosynthesis and cellular cholesterol homeostasis are tightly interconnected. *The Journal of Biological Chemistry*, 276(28):26534–26541, 2001.

[341] D. Lucero et al. Does non-alcoholic fatty liver impair alterations of plasma lipoproteins and associated factors in metabolic syndrome? *Clinica Chimica Acta; International Journal of Clinical Chemistry*, 412(7):587–592, 2011.

[342] T Radeau, P Lau, M Robb, M McDonnell, G Ailhaud, and R McPherson. Cholesteryl ester transfer protein (CETP) mRNA abundance in human adipose tissue: relationship to cell size and membrane cholesterol content. *Journal of Lipid Research*, 36(12):2552–2561, 1995.

[343] T. Gautier, D. Masson, and L. Lagrost. The potential of cholesteryl ester transfer protein as a therapeutic target. *Expert Opinion on Therapeutic Targets*, 20(1):47–59, 2016.

[344] J. Kettunen et al. Lipoprotein signatures of cholesteryl ester transfer protein and HMG-CoA reductase inhibition. *PLOS Biology*, 17(12):e3000572, 2019.

[345] T. R. Webb et al. Systematic evaluation of pleiotropy identifies 6 further loci associated with coronary artery disease. *Journal of the American College of Cardiology*, 69(7):823–836, 2017.

[346] G. Hartmann et al. Disposition into adipose tissue determines accumulation and elimination kinetics of the cholesteryl ester transfer protein inhibitor anacetrapib in mice. *Drug Metabolism and Disposition: The Biological Fate of Chemicals*, 44(3):428–434, 2016.

[347] D. G. Johns et al. Impact of drug distribution into adipose on tissue function: The cholesteryl ester transfer protein (CETP) inhibitor anacetrapib as a test case. *Pharmacology Research & Perspectives*, 7(6):e00543, 2019.

[348] S. Liu et al. Crystal structures of cholesteryl ester transfer protein in complex with inhibitors. *The Journal of Biological Chemistry*, 287(44):37321–37329, 2012.

[349] D. andothers Rhainds. Role of adenylate cyclase 9 in the pharmacogenomic response to dalcetrapib. *Circulation. Genomic and Precision Medicine*, 14(2):e003219, 2021.

[350] F. J. Rios et al. Cholesteryl ester-transfer protein inhibitors stimulate aldosterone biosynthesis in adipocytes through nox-dependent processes. *Journal of Pharmacology and Experimental Therapeutics*, 353(1):27–34, 2015.

[351] A. F. Schmidt et al. Cholesteryl ester transfer protein (cetp) as a drug target for cardiovascular disease. *Nature Communications*, 12(1):5640, 2021.

[352] P Roy et al. Structure-function relationships of human cholesteryl ester transfer protein: analysis using monoclonal antibodies. *Journal of Lipid Research*, 37(1):22–34, 1996.

[353] S Wang, L Deng, R. W. Milne, and A. R. Tall. Identification of a sequence within the c-terminal 26 amino acids of cholesteryl ester transfer protein responsible for binding a neutralizing monoclonal antibody and necessary for neutral lipid transfer activity. *Journal of Biological Chemistry*, 267(25):17487–17490, 1992.

[354] S. J. Bush, L. Chen, J. M. Tovar-Corona, and A. O. Urrutia. Alternative splicing and the evolution of phenotypic novelty. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1713):20150474, 2017.

[355] J. Tazi, N. Bakkour, and S. Stamm. Alternative splicing and disease. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1792(1):14–26, 2009.

[356] A. J. Ward and T. A. Cooper. The pathobiology of splicing. *The Journal of pathology*, 220(2):152–163, 2010.

[357] A. C. Papp et al. Cholesteryl ester transfer protein (cetp) polymorphisms affect mrna splicing, hdl levels, and sex-dependent cardiovascular risk. *PLOS ONE*, 7(3):1–9, 2012.

[358] Adam Suhy, Katherine Hartmann, Leslie Newman, Audrey Papp, Thomas Toneff, Vivian Hook, and Wolfgang Sadee. Genetic variants affecting alternative splicing of human cholesteryl ester transfer protein. *Biochemical and biophysical research communications*, 443(4):1270–1274, 2014.

[359] Adam Suhy, Katherine Hartmann, Audrey Papp, Danxin Wang, and Wolfgang Sadee. Regulation of CETP expression by upstream polymorphisms: Reduced expression associated with rs247616. *Pharmacogenetics and genomics*, 25(8):394–401, 2015.

[360] I. Gamache et al. A sex-specific evolutionary interaction between ADCY9 and CETP. *eLife*, 10:e69198, 2021.

[361] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.

[362] G. Abraham, Y. Qiu, and M. Inouye. Flashpca2: principal component analysis of Biobank-scale genotype datasets. *Bioinformatics*, 33(17):2776–2778, 2017.

[363] G. Schwarzer, J. R. Carpenter, and G. Rücker. *Meta-Analysis with R*. Use R! Springer International Publishing, 2015.

[364] S. A. Kim et al. gpart: human genome partitioning and visualization of high-density SNP data by identifying haplotype blocks. *Bioinformatics (Oxford, England)*, 35(21):4419–4421, 2019.

[365] G. Hemani et al. The MR-base platform supports systematic causal inference across the human phenome. *eLife*, 7:e34408, 2018.

[366] B. Elsworth et al. The mrc ieu opengwas data infrastructure. *bioRxiv*, 2020.

[367] O. O. Yavorska and S. Burgess. Mendelianrandomization: an r package for performing mendelian randomization analyses using summarized data. *International Journal of Epidemiology*, 46(6):1734–1739, 2017.

[368] A. Mizuno and Y. Okada. Biological characterization of expression quantitative trait loci (eQTLs) showing tissue-specific opposite directional effects. *European Journal of Human Genetics*, 27(11):1745–1756, 2019.

[369] L. H. Duntas. Thyroid disease and lipids. *Thyroid®*, 12(4):287–293, 2002.

[370] K. C. B. Tan, S. W. M. Shiu, and A. W. C. Kung. Plasma cholesteryl ester transfer protein activity in hyper- and hypothyroidism1. *The Journal of Clinical Endocrinology & Metabolism*, 83(1):140–143, 1998.

[371] T. Ittermann et al. Low serum TSH levels are associated with low values of fat-free mass and body cell mass in the elderly. *Scientific Reports*, 11:10547, 2021.

[372] K. M. Shekhda. The association of hyperthyroidism and immune thrombocytopenia: Are we still missing something? *Tzu-Chi Medical Journal*, 30(3):188–190, 2018.

[373] Ö. Tarım. Thyroid hormones and growth in health and disease. *Journal of Clinical Research in Pediatric Endocrinology*, 3(2):51–55, 2011.

[374] K. Watanabe et al. A global overview of pleiotropy and genetic architecture in complex traits. *Nature Genetics*, 51(9):1339–1348, 2019.

[375] H. Roh and D. Lee. Respiratory function of university students living at high altitude. *Journal of Physical Therapy Science*, 26(9):1489–1492, 2014.

[376] C. A. Weitz, R. M. Garruto, and C.-T. Chin. Larger FVC and FEV1 among tibetans compared to han born and raised at high altitude. *American Journal of Physical Anthropology*, 159(2):244–255, 2016.

[377] Francesca Gaccioli, Susanne Lager, Ulla Sovio, D. Stephen Charnock-Jones, and Gordon C.S. Smith. The pregnancy outcome prediction (POP) study: Investigating the relationship between serial prenatal ultrasonography, biomarkers, placental phenotype and adverse pregnancy outcomes. *Placenta*, 59:S17–S25, 2017.

[378] A. Iglesias, A. Montelongo, E. Herrera, and M. A. Lasunción. Changes in cholesteryl ester transfer protein activity during normal gestation and postpartum. *Clinical Biochemistry*, 27(1):63–68, 1994.

[379] C. Zhang et al. Changes in cholesteryl ester transfer protein concentration during normal gestation. *European Journal of Lipid Science and Technology*, 108(9):730–734, 2006.

[380] M. C. P. Roland, K. Godang, P. Aukrust, T. Henriksen, and T. Lekva. Low CETP activity and unique composition of large VLDL and small HDL in women giving birth to small-for-gestational age infants. *Scientific Reports*, 11(1):6213, 2021.

[381] M. Fleckenstein et al. Age-related macular degeneration. *Nature Reviews Disease Primers*, 7(1):31, 2021.

[382] Wenchao Zheng, Rachel E. Reem, Saida Omarova, Suber Huang, Pier Luigi DiPatre, Casey D. Charvet, Christine A. Curcio, and Irina A. Pikuleva. Spatial distribution of the pathways of cholesterol homeostasis in human retina. *PLoS ONE*, 7(5):e37926, 2012.

[383] Q. Gu et al. Cyclic amp-dependent protein kinase a regulates the alternative splicing of camkiiδ. *PLOS ONE*, 6(11):1–8, 2011.

[384] A. R. Frisancho. Human growth and pulmonary function of a high altitude peruvian quechua population. *Human Biology*, 41(3):365–379, 1969.

[385] M. Kiyamu et al. Developmental and genetic components explain enhanced pulmonary volumes of female peruvian quechua. *American Journal of Physical Anthropology*, 148(4):534–542, 2012.

[386] L. G. Moore, S. M. Charles, and C. G. Julian. Humans at high altitude: Hypoxia and fetal growth. *Respiratory Physiology & Neurobiology*, 178(1):181–190, 2011.

[387] M. A. Nieves-Colón et al. Clotting factor genes are associated with preeclampsia in high-altitude pregnant women in the peruvian andes. *The American Journal of Human Genetics*, 109(6):1117–1139, 2022.

[388] J. Pacheco-Romero et al. Genetic markers for preeclampsia in peruvian women. *Colombia Médica : CM*, 52(1):e2014437, 2023.

[389] J. T. Robinson et al. Integrative genomics viewer. *Nature biotechnology*, 29(1):24–26, 2011.

[390] C. Alfieri, M. Birkenbach, and E. Kieff. Early events in epstein-barr virus infection of human b lymphocytes. *Virology*, 181(2):595–608, 1991.

[391] S. Jiang et al. The epstein-barr virus regulome in lymphoblastoid cells. *Cell host & microbe*, 22(4):561–573.e4, 2017.

[392] L. W. Wang et al. Epstein-barr virus subverts mevalonate and fatty acid pathways to promote infected b-cell proliferation and survival. *PLOS Pathogens*, 15(9):e1008030, 2019.

[393] K. K. Anagnostopoulou et al. Sex-associated effect of CETP and LPL polymorphisms on postprandial lipids in familial hypercholesterolaemia. *Lipids in Health and Disease*,

8:24, 2009.

[394] P. B. Duell and E. L. Bierman. The relationship between sex hormones and high-density lipoprotein cholesterol levels in healthy adult men. *Archives of Internal Medicine*, 150(11):2317–2320, 1990.

[395] G. B. Lim. Role of sex hormones in cardiovascular diseases. *Nature Reviews Cardiology*, 18(6):385–385, 2021.

[396] G. B. Phillips, B. H. Pinkernell, and T.-Y. Jing. Relationship between serum sex hormones and coronary artery disease in postmenopausal women. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 17(4):695–701, 1997.

[397] S. P. Poulin, R. Dautoff, J. C. Morris, L. F. Barrett, and B. C. Dickerson. Amygdala atrophy is prominent in early alzheimer's disease and relates to symptom severity. *Psychiatry research*, 194(1):7–13, 2011.

[398] A. Arias-Vásquez et al. The cholesteryl ester transfer protein (CETP) gene and the risk of alzheimer's disease. *Neurogenetics*, 8(3):189–193, 2007.

[399] A. Tawakol et al. Relation between resting amygdalar activity and cardiovascular events: a longitudinal and cohort study. *The Lancet*, 389(10071):834–845, 2017.

[400] P. J. Gianaros et al. Preclinical atherosclerosis covaries with individual differences in reactivity and functional connectivity of the amygdala. *Biological psychiatry*, 65(11):943–950, 2009.

[401] S. Guo et al. Association between eight functional polymorphisms and haplotypes in the cholesterol ester transfer protein (CETP) gene and dyslipidemia in national minority adults in the far west region of china. *International Journal of Environmental Research and Public Health*, 12(12):15979–15992, 2015.

[402] H. Hou et al. Association between six CETP polymorphisms and metabolic syndrome in uyghur adults from xinjiang, china. *International Journal of Environmental Research and Public Health*, 14(6):653, 2017.

[403] P. H. Quanjer, G. L. Hall, S. Stanojevic, T. J. Cole, and J. Stocks. Age- and height-based prediction bias in spirometry reference equations. *European Respiratory Journal*, 40(1):190–197, 2012.

[404] Z. A. Broere-Brown et al. Sex-specific differences in fetal and infant growth patterns: a prospective population-based cohort study. *Biology of Sex Differences*, 7(1):65, 2016.

[405] D. W. Nebert and T. P. Dalton. The role of cytochrome p450 enzymes in endogenous signalling pathways and environmental carcinogenesis. *Nature Reviews Cancer*, 6(12):947–960, 2006.

[406] P. B. Danielson. The cytochrome P450 superfamily: biochemistry, evolution and drug metabolism in humans. *Current Drug Metabolism*, 3:561–597, 2002.

[407] W. C. Wright, J. Chenge, and T. Chen. Structural perspectives of the cyp3a family and their small molecule modulators in drug metabolism. *Liver Research*, 3(3):132–142, 2019.

[408] D. R. Nelson et al. P450 superfamily: update on new sequences, gene mapping, accession numbers and nomenclature. *Pharmacogenetics*, 6(1):1–42, 1996.

[409] D. R. Nelson et al. Comparison of cytochrome P450 (CYP) genes from the mouse and human genomes, including nomenclature recommendations for genes, pseudogenes and alternative-splice variants. *Pharmacogenetics and Genomics*, 14(1):1–18, 2004.

[410] D. W. Nebert, K. Wikvall, and W. L. Miller. Human cytochromes P450 in health and disease. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1612):20120431, 2013.

[411] A. Gaedigk. Complexities of CYP2D6 gene analysis and interpretation. *International Review of Psychiatry*, 25(5):534–553, 2013.

[412] A. Gaedigk et al. Prediction of CYP2D6 phenotype from genotype across world populations. *Genetics in Medicine*, 19(1):69–76, 2017.

[413] L. Elens et al. CYP3A4*22: promising newly identified CYP3A4 variant allele for personalizing pharmacotherapy. *Pharmacogenomics*, 14(1):47–62, 2012.

[414] L. Rojas et al. Effect of CYP3A5*3 on kidney transplant recipients treated with tacrolimus: a systematic review and meta-analysis of observational studies. *The Pharmacogenomics Journal*, 15(1):38–48, 2015.

[415] B. Tavira et al. A search for new CYP3A4 variants as determinants of tacrolimus dose requirements in renal-transplanted patients. *Pharmacogenetics and Genomics*, 23(8):445–448, 2013.

[416] D. Wang et al. Intronic polymorphism in CYP3A4 affects hepatic expression and response to statin drugs. *The Pharmacogenomics Journal*, 11(4):274–286, 2011.

[417] J. Lamba et al. PharmGKB summary: very important pharmacogene information for CYP3A5. *Pharmacogenetics and genomics*, 22(7):555–558, 2012.

[418] C. S. Carlson et al. Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Research*, 15(11):1553–1565, 2005.

[419] R. K. Bains et al. Molecular diversity and population structure at the Cytochrome P450 3A5 gene in africa. *BMC genetics*, 14:34–34, 2013.

[420] Y. Yasukochi and Y. Satta. Molecular evolution of the CYP2D subfamily in primates: purifying selection on substrate recognition sites without the frequent or long-tract gene conversion. *Genome Biology and Evolution*, 7(4):1053–1067, 2015.

[421] S.-A. Brown and N. Pereira. Pharmacogenomic Impact of CYP2C19 Variation on Clopidogrel Therapy in Precision Cardiovascular Medicine. *Journal of Personalized Medicine*, 8(1):8, 2018.

[422] S. A. Scott et al. Clinical Pharmacogenetics Implementation Consortium Guidelines for CYP2C19 Genotype and Clopidogrel Therapy: 2013 Update. *Clinical Pharmacology & Therapeutics*, 94(3):317–323, 2013.

[423] S. Kudaravalli et al. Gene expression levels are a target of recent natural selection in the human genome. *Molecular Biology and Evolution*, 26(3):649–658, 2009.

[424] A. C. Nica and E. T. Dermitzakis. Expression quantitative trait loci: present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1620), 2013.

[425] Y. Huang, S. Wuchty, and T. M. Przytycka. eQTL Epistasis – Challenges and Computational Approaches. *Frontiers in Genetics*, 4, 2013.

[426] R. V. Rohlfs, W. J. Swanson, and B. S. Weir. Detecting Coevolution through Allelic Association between Physically Unlinked Loci. *American Journal of Human Genetics*, 86(5):674–685, 2010.

[427] M. Byrska-Bishop et al. High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *bioRxiv*, page 2021.02.06.430068, 2021.

[428] M. Kircher and others. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46(3):310–315, 2014.

[429] B. S. Weir and C. Clark Cockerham. Estimating F-Statistics for the Analysis of Population structure. *Evolution*, 38(6):1358–1370, 1984.

[430] K. M. Siewert and B. F. Voight. Detecting long-term balancing selection using allele frequency correlation. *Mol Biol Evol*, 34(11):2996–3005, 2017.

[431] B. F. Voight et al. A Map of Recent Positive Selection in the Human genome. *PLOS Biology*, 4(3):e72, 2006.

[432] H. Li. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987–2993, 2011.

[433] C. A. Maclean, N. P. Chue Hong, and J. G.D. Prendergast. hapbin: An Efficient Program for Performing Haplotype-Based Scans for Positive Selection in Large Genomic datasets. *Molecular Biology and Evolution*, 32(11):3027–3029, 2015.

[434] W. James Kent, Charles W. Sugnet, Terrence S. Furey, Krishna M. Roskin, Tom H. Pringle, Alan M. Zahler, and David Haussler. The human genome browser at ucsc. *Genome Research*, 12(6):996–1006, 2002.

[435] C. C. Chang et al. Second-generation plink: rising to the challenge of larger and richer datasets. *Gigascience*, 4:7, 2015.

[436] J. Lonsdale et al. The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, 45(6):580–585, 2013.

[437] M. Lawrence, R. Gentleman, and V. Carey. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics*, 25(14):1841–1842, 2009.

[438] F. Hammal et al. Remap 2022: a database of human, mouse, drosophila and arabidopsis regulatory regions from an integrative analysis of dna-binding sequencing experiments. *Nucleic acids research*, 50(D1):D316–d325, 2022.

[439] J. C. Denny et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nature Biotechnology*, 31(12):1102–1111, 2013.

[440] M. S Lyon and Others. The variant call format provides efficient and robust storage of gwas summary statistics. *Genome Biology*, 22(1):32, 2021.

[441] Y. Kim and W. Stephan. Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics*, 160(2):765–777, 2002.

[442] N. L. Kirischian and J. Y. Wilson. Phylogenetic and functional analyses of the cytochrome P450 family 4. *Mol Phylogenet Evol*, 62(1):458–71, 2012.

[443] E. E. Thompson, H. Kuttab-Boulos, D. Witonsky, L. Yang, B. A. Roe, and A. Di Rienzo. Cyp3a variation and the evolution of salt-sensitivity variants. *American journal of human genetics*, 75(6):1059–1069, 2004.

[444] A. Aqil, L. Speidel, P. Pavlidis, and O. Gokcumen. Balancing selection on genomic deletion polymorphisms in humans. *eLife*, 12:e79111, 2023.

[445] V. Llaurens, A. Whibley, and M. Joron. Genetic architecture and balancing selection: the life and death of differentiated variants. *Mol Ecol*, 26(9):2430–2448, 2017.

[446] J. M. Akey et al. The effect that genotyping errors have on the robustness of common linkage-disequilibrium measures. *The American Journal of Human Genetics*, 68(6):1447–1456, 2001.

[447] X. Chen et al. Molecular Population Genetics of Human CYP3A Locus: Signatures of Positive Selection and Implications for Evolutionary Environmental medicine. *Environmental Health Perspectives*, 117(10):1541–1548, 2009.

[448] E. E. Thompson et al. Sequence diversity and haplotype structure at the human CYP3A cluster. *Pharmacogenomics J*, 6(2):105–14, 2006.

[449] J. E. Zhang et al. Effect of genetic variability in the CYP4F2, CYP4F11, and CYP4F12 Genes on Liver mRNA Levels and Warfarin response. *Front Pharmacol*, 8:323, 2017.

[450] R. Liang et al. Influence of CYP4F2 genotype on warfarin dose requirement–a systematic review and meta-analysis. *Thrombosis Research*, 130(1):38–44, 2012.

[451] O. Singh et al. Influence of CYP4F rs2108622 (V433M) on Warfarin Dose Requirement in Asian patients. *Drug Metabolism and Pharmacokinetics*, 26(2):130–136, 2011.

[452] Y. Guttman, A. Nudel, and Z. Kerem. Polymorphism in Cytochrome P450 3A4 Is Ethnicity related. *Frontiers in Genetics*, 10, 2019.

[453] M. A. Clark et al. Plasmodium vivax infection compromises reticulocyte stability. *Nature Communications*, 12(1):1629, 2021.

[454] M. Yi et al. Functional characterization of a common CYP4F11 genetic variant and identification of functionally defective CYP4F11 variants in erythromycin metabolism and 20-HETE synthesis. *Archives of Biochemistry and Biophysics*, 620:43–51, 2017.

[455] K. Stark, B. Wongsud, R. Burman, and E. H. Oliw. Oxygenation of polyunsaturated long chain fatty acids by recombinant CYP4F8 and CYP4F12 and catalytic importance of Tyr-125 and Gly-328 of cyp4f8. *Archives of Biochemistry and Biophysics*,

441(2):174–181, 2005.

[456] K. L. Korunes and K. Samuk. pixy: Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Molecular Ecology Resources*, 21(4):1359–1368, 2021.

[457] M. E. Ritchie et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47, 2015.

[458] C. W. Law, Y. Chen, W. Shi, and G. K. Smyth. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15(2):1–17, 2014.

[459] The GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015.

[460] F. Mostefai et al. Population genomics approaches for genetic characterization of sars-cov-2 lineages. *Frontiers in Medicine*, 9, 2022.

[461] E. Paradis and K. Schliep. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35:526–528, 2019.

[462] E. Paradis. pegas: an R package for population genetics with an integrated–modular approach. *Bioinformatics*, 26:419–420, 2010.

[463] S. Khare et al. Gisaid's role in pandemic response. *China CDC Weekly*, 3:1049, 2021.

[464] M.-A. Legault et al. Study of effect modifiers of genetically predicted cetp reduction. *Genetic Epidemiology*, 47(2):198–212, 2023.

[465] D. G. Augusto and M. L. Petzl-Erler. KIR and HLA under pressure: evidences of coevolution across worldwide populations. *Human genetics*, 134(9):929–940, 2015.

[466] P. H. Krijger and W. de Laat. Regulation of disease-associated gene expression in the 3d genome. *Nature reviews. Molecular cell biology*, 17(12):771–782, 2016.

[467] G. S. Getz and C. A. Reardon. Animal models of atherosclerosis. *Arteriosclerosis, thrombosis, and vascular biology*, 32(5):1104–1115, 2012.

[468] C. Yao et al. Dynamic Role of trans Regulation of Gene Expression in Relation to Complex Traits. *American journal of human genetics*, 100(4):571–580, 2017.

[469] J. Chen, Q. Ding, L. An, and H. Wang. Ca2+-stimulated adenylyl cyclases as therapeutic targets for psychiatric and neurodevelopmental disorders. *Frontiers in Pharmacology*, 13:949384, 2022.

[470] P. Nghe and S. J. Kogenaru, M. Tans. Sign epistasis caused by hierarchy within signalling cascades. *Nature Communications*, 9(1):1451, 2018.

[471] F. Baier, F. Gauye, R. Perez-Carrasco, J. L. Payne, and Y. Schaerli. Environment-dependent epistasis increases phenotypic diversity in gene regulatory networks. *Science Advances*, 9(21):eadf1773, 2023.

[472] K. Liu et al. Epistatic evidence for gender-dependant slow neurotransmission signalling in substance use disorders: PPP1R12B versus PPP1R1B. *EBioMedicine*, 61:103066, 2020.

[473] Y. Li et al. Statistical and Functional Studies Identify Epistasis of Cardiovascular Risk Genomic Variants From Genome-Wide Association Studies. *Journal of the American Heart Association*, 9(7):e014146, 2020.

[474] J. A. Jefferson et al. Increased oxidative stress following acute and chronic high altitude exposure. *High Altitude Medicine & Biology*, 5(1):61–69, 2004.

[475] K. Duhig, L. C. Chappell, and A. H. Shennan. Oxidative stress in pregnancy and reproduction. *Obstetric Medicine*, 9(3):113–116, 2016.

[476] H. Ogita and J. K. Liao. Endothelial function and oxidative stress. *Endothelium*, 11(2):123–132, 2004.

[477] N. Engedal and Others. From oxidative stress damage to pathways, networks, and autophagy via micrornas. *Oxidative Medicine and Cellular Longevity*, 2018:4968321, 2018.

[478] A. C. B. A. Wanschel et al. The presence of cholesteryl ester transfer protein (cetp) in endothelial cells generates vascular oxidative stress and endothelial dysfunction. *Biomolecules*, 11(1), 2021.

[479] H. Soran, J. D. Schofield, and P. N. Durrington. Antioxidant properties of hdl. *Frontiers in Pharmacology*, 6, 2015.

[480] D. I. Chiarello et al. Oxidative stress: Normal pregnancy versus preeclampsia. *Biochimica et biophysica acta - Molecular Basis Disease*, 1866(2):165354, 2020.

[481] D. Mondal, T. S. Galloway, T. C. Bailey, and F. Mathews. Elevated risk of stillbirth in males: systematic review and meta-analysis of more than 30 million births. *BMC Medicine*, 12(1):220, 2014.

[482] E. Elsmén, K. Källén, K. Maršál, and L. Hellström-Westas. Fetal gender and gestational-age-related incidence of pre-eclampsia. *Acta Obstetricia et Gynecologica Scandinavica*, 85(11):1285–1291, 2006.

[483] D. Mankuta, M. Elami-Suzin, A. Elhayani, and S. Vinker. Lipid profile in consecutive pregnancies. *Lipids in Health and Disease*, 9(1):58, 2010.

[484] J. Challis, J. Newnham, F. Petraglia, M. Yeganegi, and A. Bocking. Fetal sex and preterm birth. *Placenta*, 34(2):95–99, 2013.

[485] R. Schacht, D. Tharp, and K. R. Smith. Sex ratios at birth vary with environmental harshness but not maternal condition. *Scientific Reports*, 9(1):9066, 2019.

[486] I. Sreckovic et al. Distinct composition of human fetal HDL attenuates its anti-oxidative capacity. *Biochimica et biophysica acta*, 1831(4):737–746, 2013.

[487] A. Zeljkovic et al. Changes in LDL and HDL Subclasses in Normal Pregnancy and Associations with Birth Weight, Birth Length and Head Circumference. *Maternal and child health journal*, 17(3):556–565, 2013.

[488] I. Žitňanová et al. Gender differences in LDL- and HDL-cholesterol subfractions in patients after the acute ischemic stroke and their association with oxidative stress markers. *Journal of clinical biochemistry and nutrition*, 63(2):144–148, 2018.

[489] E. A. Brown, M. Ruvolo, and P. C. Sabeti. Many ways to die, one way to arrive: how selection acts through pregnancy. *Trends in Genetics*, 29(10):585–592, 2013.

[490] M. Kockx, L. Roberts, J. Wang, C. Tran, M. A. Brown, and L. Kritharides. Effects of pre-eclampsia on HDL-mediated cholesterol efflux capacity after pregnancy. *Atherosclerosis plus*, 48:12–19, Apr 2022.

[491] B. D. Taylor et al. The impact of female fetal sex on preeclampsia and the maternal immune milieu. *Pregnancy Hypertens*, 12:53–57, Apr 2018.

[492] E. A. Townsend, V. M. Miller, and Y. S. Prakash. Sex Differences and Sex Steroids in Lung Health and Disease. *Endocrine Reviews*, 33(1):1–47, 2012.

[493] N. S. Dhalla et al. *Sex-Specific Differences in β-Adrenoceptor Signal Transduction in Heart Failure Due to Volume-Overload*, pages 147–158. Springer International Publishing, Cham, 2020.

[494] E. J. Filardo, J. A. Quinn, Jr. Frackelton, A. R., and J. I. Bland. Estrogen Action Via the G Protein-Coupled Receptor, GPR30: Stimulation of Adenylyl Cyclase and cAMP-Mediated Attenuation of the Epidermal Growth Factor Receptor-to-MAPK Signaling Axis. *Molecular Endocrinology*, 16(1):70–84, 2002.

[495] M. Horiuchi, Y. Kirihara, Y. Fukuoka, and H. Pontzer. Sex differences in respiratory and circulatory cost during hypoxic walking: potential impact on oxygen saturation. *Scientific Reports*, 9(1):9550, 2019.

[496] D. J. Hunter. Gene-environment interactions in human diseases. *Nature reviews. Genetics*, 6(4):287–298, 2005.

[497] N. Mulder et al. H3africa: current perspectives. *Pharmacogenomics and personalized medicine*, 11:59–66, 2018.

[498] P.N. Durrington. Cholesteryl ester transfer protein (cetp) inhibitors. *British Journal of Cardiology*, 19(3):126, 2012.

[499] M. M. A. Willemars, M. Nabben, J. A. J. Verdonschot, and M. F. Hoes. Evaluation of the Interaction of Sex Hormones and Cardiovascular Function and Health. *Current heart failure reports*, 19(4):200–212, 2022.

[500] J. Domingo, P. Baeza-Centurion, and B. Lehner. The causes and consequences of genetic interactions (epistasis). *Annual Review of Genomics and Human Genetics*, 20(1):433–460, 2019.

[501] H. Kemble et al. Flux, toxicity, and expression costs generate complex genetic interactions in a metabolic pathway. *Science Advances*, 6(23):eabb2236, 2020.

[502] L. Wang et al. Epstein-Barr Virus Episome Physically Interacts with Active Regions of the Host Genome in Lymphoblastoid Cells. *J Virol*, 94(24), Nov 2020.

[503] A. Regev et al. Science forum: The human cell atlas. *eLife*, 6:e27041, 2017.

[504] S. T. Pan et al. Computational Identification of the Paralogs and Orthologs of Human Cytochrome P450 Superfamily and the Implication in Drug Discovery. *International Journal of Molecular Sciences*, 17(7), 2016.

[505] K. Gellner et al. Genomic organization of the human CYP3A locus: identification of a new, inducible CYP3A gene. *Pharmacogenetics and Genomics*, 11(2):111–121, 2001.

[506] W. Liu et al. African origin of the malaria parasite Plasmodium vivax. *Nature Communication*, 5:3346, 2014.

Annexe A

Méthodes de la discussion

A.1. Épissage alternatif dans les cellules ADCY9-Knock Down

Afin d'évaluer l'impact d'une diminution des niveaux d'ADCY9 sur l'épissage alternatif de l'exon 9, nous avons utilisé les données des cellules hépatocytes HepG2 qui ont été utilisées dans notre article du chapitre 2. Les fragments de lecture traités ont été alignés sur le génome de référence GRCh38 à l'aide de STAR (v.2.6.1a)[215]. Ensuite, nous avons utilisé ASpli [231] pour quantifier les valeurs PSI de chaque échantillon pour l'épissage alternatif de l'exon 9.

Il y avait cinq échantillons appariés, dont une condition était de type sauvage (WT, *Wild-Type* en anglais) et l'autre condition était ADCY9-KD. Un échantillon et son apparié ont été exclus en raison d'une faible couverture (moins de cinq fragments de lecture) sur la région d'épissage.

Les valeurs PSI des deux conditions ont été comparées statistiquement en utilisant un test de Wilcoxon-T apparié avec le logiciel de base de R [361]. Dans la figure 6.2b, les comptages des fragments de lecture représentant l'inclusion de l'exon 9 ("$\rightarrow Ex9 \rightarrow$" dans la figure 6.2) correspondent à la somme des fragments de lecture de l'exon 8 à l'exon 9, et de l'exon 9 à l'exon 10. Les comptages des fragments de lecture représentant l'exclusion de l'exon 9 ($Ex8 \rightarrow Ex10$ dans la figure 6.2) correspondent au nombre de fragments de lecture de l'exon 8 à l'exon 10 multiplié par deux.