Université de Montréal

Caractérisation de deux familles de pharmacogènes, les gènes CYP3A et CYP4F

par

Alex Richard-St-Hilaire

Département de biochimie et médecine moléculaire

Faculté des arts et des sciences

Mémoire présenté en vue de l'obtention du grade de

Maître ès sciences (M.Sc.)

en Bio-informatique

Avril 2021

Université de Montréal

Faculté des études supérieures et postdoctorales

Ce mémoire intitulé

Caractérisation de deux familles de pharmacogènes, les gènes CYP3A et CYP4F

présenté par

Alex Richard-St-Hilaire

a été évalué par un jury composé des personnes suivantes :

Martin Smith
(président-rapporteur)

Julie Hussin
(directeur de recherche)

Martine Tétreault
(membre du jury)

Résumé

Les cytochromes P450 (CYP450) sont des hémoprotéines intervenant généralement dans la détoxication de l'organisme sous forme de biodégradation de molécules xénobiotiques et participent à la décomposition de certains médicaments. Cependant, les gènes codant pour les protéines CYP450 sont souvent sous-analysés dans les études génomiques à grande échelle en raison de leur difficulté d'analyse due à un haut taux de polymorphisme. Deux sous-familles seront étudiées plus en profondeur: les sous-familles CYP3A et CYP4F. La sous-famille CYP3A métabolise environ 50% des médicaments alors que les enzymes CYP4F, quant à eux, sont impliquées dans le métabolisme de composés endogènes, de nutriments et de médicaments. Les gènes de ces sous-familles sont fortement polymorphes et ce, à travers les populations humaines. Ainsi, la variabilité entre les différentes populations peut affecter la réponse aux médicaments et autres fonctions métaboliques. Dans ce projet, deux grands jeux de données, l'un en génétique des populations (le Projet des 1000 Génomes) et l'autre en transcriptomique (GTEx) seront utilisés afin d'identifier des signatures de sélection naturelle dans les gènes CYP3A et CYP4F, ainsi que leur impact sur l'expression génique de ces gènes. Nous avons détecté différentes forces de sélection (positive et balancée) dans les deux sous-familles. Certains polymorphismes identifiés comme étant sous pression sélective sont associés à une expression différentielle des gènes des deux sous-familles. Ce projet permet de mieux comprendre l'impact des mutations sous pression sélective se situant dans les gènes des sous-familles CYP3A et CYP4F. Cette caractérisation génétique permettra d'obtenir des prédictions plus fiables en pharmacogénomique et en génomique humaine, en raison de l'influence de ces gènes sur la réponse aux médicaments.

Mots clés : Cytochromes P450, CYP3A, CYP4F, Génétique des populations, Bioinformatique, Génomique, Transcriptomique

Abstract

Cytochromes P450 (CYP450) are hemoproteins generally involved in the detoxification of the body of xenobiotic molecules and participate in the metabolism of certain drugs. However, genes encoding CYP450 proteins are often under-analyzed in large-scale genomic studies due to their difficulty of analysis because of their high rate of polymorphism. Two subfamilies will be studied thoroughly: the CYP3A and CYP4F subfamilies. The CYP3A subfamily metabolizes approximately 50% of drugs while CYP4F enzymes are involved in the metabolism of endogenous compounds, nutrients and drugs. The genes of these subfamilies are highly polymorphic across populations. Thus, variability between different populations can affect drug response and other metabolic functions. In this project, two large datasets, one in population genetics (the 1000 Genomes Project) and the other in transcriptomics (GTEx) will be used to assess sites under selective pressure found in the CYP3A and CYP4F genes, as well as their impact on the gene expression of these genes. Different natural selection forces (positive and balancing) were detected in the two subfamilies. Certain polymorphisms that have been identified as being under selective pressure are associated with differential expression of genes from the two subfamilies. This project will lead to a better understanding of the impact of mutations under selective pressure in the genes of the CYP3A and CYP4F families. This genetic characterization will provide more reliable predictions in pharmacogenomics and human genomics, due to the influence of these genes on drug response.

Keywords: Cytochromes P450, CYP3A, CYP4F, Population genetics, Bioinformatics, Genomics, Transcriptomics

Table des matières

Liste des tableaux

Liste des figures

Liste des sigles et des abréviations

1000G	Projet des 1000 Génomes (*The 1000 Genomes Project*)
ADN	Acide désoxyribonucléique
ADNc	Acide désoxyribonucléique complémentaire
ARN	Acide ribonucléique
ARNm	Acide ribonucléique messager
ARNr	Acide ribonucléique ribosomique
ATRA	Acide tout-trans-rétinoïque
CYP450	Cytochromes P450
EHH	*Extended haplotype homozygosity*
eQTL	*Expression quantitative trait loci*
FIN	Population de la Finlande du projet des 1000 Génomes
GTEx	*The Genotype-Tissue Expression project*

GWD	Population de la province occidentale de la Gambie du projet des 1000 Génomes
iHS	*Integrated Haplotype Score*
Kb	Kilobase
LD	Déséquilibre de liaison
LWK	Population Luhya de Webuye au Kenya du projet des 1000 Génomes
MAF	Fréquence de l'allèle mineur
Mb	Mégabase
Pb	Paire de bases
SDHEA	Sulfate dedéhydroépiandrostérone
SNP	*Single-Nucleotide Polymorphism*
TSI	Population Toscane d'Italie du projet des 1000 Génomes
UCSC	*University of California Santa Cruz*
uLD	*Unusual linkage disequilibrium*
YRI	Population Yoruba d'Ibadan du Nigeria du projet des 1000 Génomes

Remerciements

Je tiens d'abord à remercier ma directrice de recherche Julie Hussin de m'avoir accueillie dans son laboratoire alors que j'étais étudiante au baccalauréat et de m'avoir guidé tout au long de mes études à la maîtrise. Son support et son savoir m'ont grandement aidé dans mon cheminement. Je me sens privilégiée d'avoir eu la chance d'effectuer mes études sous sa supervision et d'avoir vu grandir le laboratoire pour en arriver où il est rendu aujourd'hui.

Je remercie Marie-Pierre Dubé d'avoir parrainé mon projet et pour ses conseils durant nos rencontres.

Je remercie également Martin Smith et Martine Tétreault de votre intérêt pour mon projet en acceptant d'être sur le jury de mon mémoire.

Merci à tous les membres du laboratoire pour les discussions et les nombreux conseils durant mes 3 ans et demi au laboratoire. Un merci particulier à Jean-Christophe Grenier d'être toujours là pour nous aider avec nos questionnements et pour nous aider à résoudre nos problèmes techniques. Également, un merci particulier à Raphaël Poujol pour les nombreuses discussions constructives par rapport à mon article et à mon mémoire. Je tiens à remercier mes collègues pour les nombreux midi mots croisés et pour les dîners en virtuel. Les dîners en virtuel ont grandement aidé à rendre cette fin de maîtrise en télétravail plus agréable.

Je remercie la faculté des études supérieures de l'Université de Montréal pour les bourses qui m'ont été octroyées au cours de ma maîtrise.

Merci à Élaine Meunier, pour son efficacité et ses nombreux rappels de date limite, mais également d'avoir été si patiente dans mes nombreux changements de choix de cours.

Une pensée pour mes amies qui m'ont permis de décrocher le temps d'un dîner ou d'un souper. Merci de m'avoir écouté parler de mon projet même si ce n'était pas nécessairement votre domaine.

Merci à mes parents de m'avoir encouragée à poursuivre mes études et pour leur support durant toutes ces années.

Et finalement, un sincère remerciement à Martin de m'avoir supportée durant toutes mes études universitaires et de m'avoir endurée pendant toutes ces fins de session où j'étais stressée et à cran. P.s C'est enfin fini !

Introduction

La réponse aux médicaments varie d'un individu à un autre en fonction de son bagage génétique et de son ancestralité. Le métabolisme des médicaments à l'origine de cette réponse est réalisé par les cytochromes P450, une famille de d'hémoprotéines[1].

En effet, les cytochromes P450 sont des hémoprotéines qui interviennent généralement dans les chaînes de transfert d'électrons sous le rôle d'oxydases. Ils sont impliqués dans la détoxication de l'organisme des molécules xénobiotiques[2] et participent à la décomposition d'un bon nombre de médicaments. Ils sont également impliqués dans le métabolisme des composés endogènes[3] et des nutriments. Cependant, les gènes codant pour les protéines CYP450 sont fortement polymorphes et comprennent donc de nombreux variants génétiques[4]. Ces variants sont également appelés «polymorphisme d'un seul nucléotide (SNP)» lorsque la variation n'affecte qu'une paire de base (pb). Le caractère hautement polymorphe de ces gènes complique leur analyse et ils sont souvent sous-analysés dans les études génomiques à grande échelle malgré leur importance. Cependant, la présence de ces variants dans ces gènes peut ainsi influencer la réponse aux médicaments et les fonctions métaboliques. Il est donc important de bien comprendre l'impact de ceux-ci.

L'approche utilisée ici pour caractériser ces variants se fera à l'aide de la génétique des populations. La génétique des populations étudie les changements des fréquences alléliques des variants à travers les différentes populations. La fluctuation des fréquences alléliques

[1]Protéine comportant un groupement hème
[2]Molécule qui est étrangère à l'organisme et toxique
[3]Composé synthétisé par l'organisme en question
[4]Un variant génétique est une position de l'ADN humain, ou locus, qui diffère du génome de référence dû au processus de mutation génétique

peut être causée, entre autres, par la sélection naturelle en fonction de l'impact du variant sur le succès reproducteur. La détection des différentes pressions sélectives, grâce à des méthodes statistiques établies en génétique des populations, permet donc l'identification de variants d'intérêts. L'association de ces variants d'intérêts à des phénotypes[5] humains peut permettre une meilleure compréhension de l'impact fonctionnel de ces variants dans la réponse aux médicaments dans les différentes populations.

Ce mémoire comporte 3 parties. La première partie est une revue de littérature des articles antérieurement publiés portant sur les notions théoriques, statistiques et bio-informatiques nécessaires à la compréhension du projet. Les hypothèses et les objectifs sont ensuite présentés. Par la suite, le chapitre 2 présente les résultats obtenus au cours de ce projet sous forme d'article scientifique. Finalement, les résultats seront discutés et une brève conclusion sera énoncée au chapitre 3.

[5]Traits observables d'un organisme

Chapitre 1

1.1. Revue de littérature

Dans ce chapitre, j'expliquerai en premier lieu ce qu'est la génétique des populations ainsi que les notions statistiques qui seront utilisées dans les sections suivantes. Par la suite, les gènes des cytochromes P450 seront décrits. Les sous-familles CYP3A et CYP4F seront décrites plus en détails, comme elles seront étudiées tout au long du mémoire.

1.1.1. Génétique des populations

La génétique des populations a pour but d'évaluer les processus affectant la diversité génétique des populations. La diversité génétique est étudiée à deux niveaux populationnels, soit au niveau intra-populationnel[1], soit au niveau inter-populationnel[2] (Relethford 2012).

1.1.1.1. *Hardy-Weinberg*

Un principe important en génétique des populations est la modèle d'Hardy-Weinberg. Le modèle d'Hardy-Weinberg permet de vérifier si un état d'équilibre est présent ou s'il y a des facteurs, comme la sélection naturelle, qui agissent sur le locus en question. L'équilibre est présent si les fréquences génotypiques restent les mêmes d'une génération à une autre et que la population est idéale. Sous ce modèle, une population est définie comme étant idéale si (Hamilton 2011):

- Les organismes sont diploïdes ;

[1] À l'intérieur d'une population
[2] Entre populations

– La reproduction est aléatoire ;
– Il n'y a pas de migration;
– Il n'y a pas de sélection;
– Il n'y a pas de mutation;
– La taille de population est infinie (ou très grande);
– Les fréquences alléliques sont les mêmes chez les mâles et les femelles;
– Les générations ne se chevauchent pas.

Afin de déterminer si l'équilibre est présent, la fréquence des deux allèles[3] est nécessaire. La fréquence allélique (p) est calculée en divisant le nombre de copie de l'allèle (x) par la taille de l'échantillon (n) (équation 1.1.1).

$$p = \frac{x}{n} \tag{1.1.1}$$

La fréquence des deux allèles (définie par p et q) est ensuite utilisée dans l'équation de l'équilibre d'Hardy-Weinberg (équation 1.1.2).

$$p^2 + 2pq + q^2 = 1 \tag{1.1.2}$$

Lorsque la sommation n'est pas égale à 1, cela signifie que le locus dévie significativement de l'équilibre d'Hardy-Weinberg, et donc qu'une des conditions présentées ci-haut n'est pas respectée. Une déviation d'Hardy-Weinberg élevée, c'est-à-dire lorsque le rejet de l'équilibre est associée à une valeur p $< 10^{-5}$, peut indiquer des erreurs systématiques lors du génotypage ou du séquençage, des relations consanguines ou un biais causé par la structure de population (B. Chen et al. 2017; Hosking et al. 2004).

1.1.1.2. *Structure populationnelle*

La reproduction non-aléatoire est une cause possible de la déviation de l'équilibre d'Hardy-Weinberg. Celle-ci cause une différence au niveaux des fréquences alléliques entre les populations et donc une structure populationnelle, également appelée stratification de la

[3]Un allèle est une des différentes formes que peut prendre un locus

population. Cette structure populationnelle peut être causée par une isolation géographique, par exemple (Hellwege et al. 2017).

La structure populationnelle est une cause importante des résultats d'associations[4] non-répliqués et peut donc fausser les résultats (Cardon et al. 2003; "Population Stratification" 2006). Une méthode utilisée afin de tenir compte de la structure populationnelle est l'analyse en composantes principales (Menozzi et al. 1978; Novembre et al. 2008). Cette méthode permet la réduction de la dimension des données et permet d'obtenir les composantes principales qui contiennent la variation génétique présente. La première composante comprend alors la combinaison mathématique des mesures qui explique la plus grande variabilité des données (Reich et al. 2008). Le principe est le même pour les composantes suivantes. Ces composantes peuvent capter la variation génétique causée par la structure et la stratification des populations et ainsi capter la variabilité causée par l'ancestralité et l'ethnicité des individus par exemple. Les composantes principales seront alors incluses dans les analyses afin de corriger ces différents effets.

1.1.1.3. *Dérive génétique*

Un processus affectant la fréquence des allèles dans une population est la dérive génétique. Celle-ci se caractérise par les différentes fluctuations des fréquences alléliques en fonction de la reproduction (Lefevre et al. 2016). Ceci est expliqué par le fait que la reproduction est un phénomène stochastique. En effet, le choix des gamètes qui seront transmises à la génération suivante est aléatoire, en fonction des individus qui vont se reproduire et transmettre leur bagage génétique. Ainsi, selon ce concept, certains allèles ne seront pas transmis à la génération suivante.

D'ailleurs, la fluctuation des allèles va tendre à la fixation ou à la perte des allèles. Cette fluctuation est non-directionnelle, c'est-à-dire qu'il n'est pas déterminé à l'avance si l'allèle en question va être fixé ou perdu. À long terme, ceci diminue la variabilité à l'intérieur

[4]Un résultat d'association se définit comme étant un variant associé avec un trait ou une maladie. Une étude d'association vérifie la corrélation entre les fréquences alléliques dans deux groupes (cas et témoins) et le trait étudié.

des populations (Relethford 2012). Une population avec une taille effective[5] petite est plus influencée par la dérive génétique qu'une grande population et a donc tendance à s'homogénéiser plus rapidement. Ce processus de fluctuation des fréquences alléliques est indépendant entre les populations et mène à la différenciation des populations (Hartl 1988).

1.1.1.4. *Sélection naturelle*

La sélection naturelle fait varier les fréquences alléliques des mutations, soit en augmentant la probabilité de transmission, donc la fréquence, en la diminuant ou en la maintenant à un niveau intermédiaire en fonction de l'influence de la mutation sur le succès reproducteur (Lefevre et al. 2016). Ces trois différentes pressions sélectives sont la sélection positive, la sélection négative et la sélection balancée. Ainsi, contrairement à la dérive génétique qui est non-directionnelle et qui ne tient donc pas en compte l'impact de la mutation, la sélection naturelle est une force directionnelle.

1.1.1.4.1. *Sélection positive*

La sélection positive est présente lorsque la mutation est avantageuse. Comme la mutation apporte une meilleure adaptabilité de l'organisme et une meilleure survie à son environnement, le succès reproducteur est ainsi meilleur, ce qui affecte positivement la valeur adaptative, aussi appelé le *fitness* (Orr 2009). Puisque les traits qui sont favorables pour un organisme ont plus de chance d'être transmis à leurs descendants (Darwin 1870), la fréquence de cet allèle tend donc à augmenter et à venir se fixer dans la population (Figure 1.1a) (Choudhuri 2014). Un phénomène, le balayage sélectif, se produit lorsqu'une mutation gagne en fréquence et que la diversité autour de la mutation avantageuse tend à diminuer (Kim et al. 2002).

1.1.1.4.2. *Sélection négative*

La sélection négative a comme effet de diminuer la fréquence allélique des variants qui sont délétères et qui ne sont donc pas avantageux dans leur environnement (Figure 1.1b).

[5]La taille effective d'une population représente le nombre d'individus qui participent à la formation de la génération suivante. Cette taille est généralement plus petite que la taille totale réelle de la population.

Comme ces variants diminuent la valeur adaptative, il est primordial que leurs fréquences restent basses afin de ne pas nuire aux fonctions biologiques (Cvijovic et al. 2018). La sélection purificatrice diminue ainsi la diversité génétique.

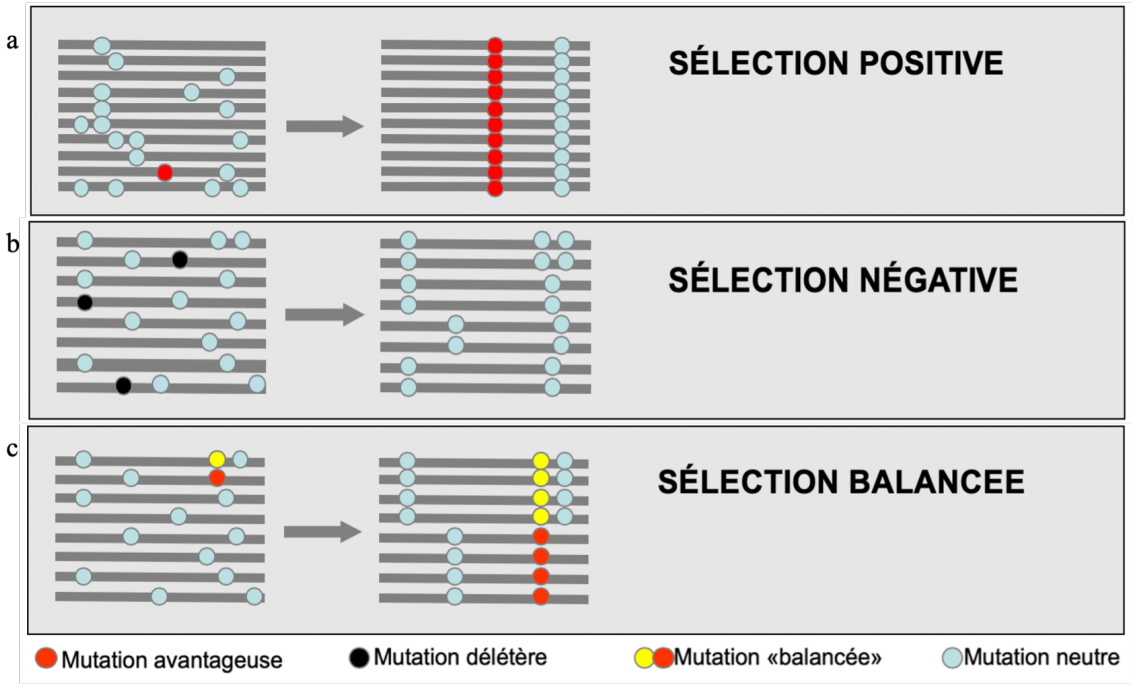


Fig. 1.1. Sélection naturelle - Effets des trois différentes forces de sélection naturelle sur les haplotypes présents. La sélection positive fait augmenter la fréquence d'un allèle avantageux (a), alors que la sélection négative fait diminuer la fréquence d'un allèle délétère (b) et que la sélection balancée maintient des allèles à une fréquence intermédiaire (c). Tirée de (Barreiro 2017)

1.1.1.4.3. *Sélection balancée*

La sélection balancée maintient les fréquences alléliques à un niveau intermédiaire, ce qui conserve la diversité génétique présente (Figure 1.1c)(Choudhuri 2014). Il y a deux formes de sélection balancée : la super-dominance et la sélection fréquence-dépendante négative. Le mécanisme de super-dominance est présent lorsque les individus hétérozygotes ont une valeur adaptative (*fitness*) plus élevée que les individus homozygotes. Ils ont donc un avantage comparativement aux individus homozygotes. Les fréquences alléliques sont donc maintenues afin de préserver cet avantage. Le deuxième mécanisme de sélection balancée, la sélection fréquence-dépendante négative, est présent lorsqu'un allèle rare est plus favorable qu'un allèle

à fréquence plus élevée. L'allèle avantageux étant sous sélection, sa fréquence va augmenter. En devenant commun, il ne sera plus favorisé et de nouveaux allèles rares seront avantageux. Ce mécanisme provoque donc une grande diversité et l'accumulation de variants à faible fréquence (Andrés 2011; Llaurens et al. 2017; Richman 2000).

La théorie neutraliste de l'évolution stipule que la plupart des mutations seraient neutres et donc que leur *fitness* le serait tout autant (M. Kimura 1991). Par conséquent, comme la sélection est près de zéro, la force de la dérive génétique est plus forte que la sélection naturelle, ce qui peut engendrer la fixation d'une mutation délétère dans une petite population. D'ailleurs, la sélection naturelle peut, dans une population avec une grande taille effective, contrer les effets de la dérive génétique (Hall 2011; Henry et al. 2008). Il y a donc un phénomène d'interaction, ou d'équilibre, entre la sélection naturelle et la dérive génétique.

1.1.1.5. *Démographie*

Comme énoncé au début de la revue de littérature, une condition nécessaire au modèle d'Hardy-Weinberg est d'avoir une population idéale, et donc que la taille de la population soit infinie ou très grande. Or, la taille d'une population varie au fil du temps, ce qui implique que la population est plus sujette aux changements et à une différenciation des fréquences alléliques. Deux situations affectant la taille des populations bien connues en génétique des populations sont l'expansion populationnelle et le goulot d'étranglement (Lefevre et al. 2016).

Tout d'abord, l'expansion populationnelle, ou la croissance de la population, est un phénomène démographique ayant un impact sur les fréquences alléliques. En effet, lors d'une croissance populationnelle rapide, un excès d'allèles rares est présent (Keinan et al. 2012). D'ailleurs, la croissance d'une population a comme conséquence d'augmenter le nombre de sites polymorphes et d'augmenter le nombre moyen de mutation par individu (Gazave et al. 2013).

Ensuite, le goulot d'étranglement se définit par une diminution importante de la taille de la population. Par exemple, une famine, une inondation ou un feu peut causer un goulot

d'étranglement. La taille effective de la population et la diversité génétique diminue alors de façon drastique. Ainsi, certains allèles peuvent être perdus alors que d'autres peuvent être sur-représentés. Comme la taille de la population est petite, les effets de la dérive génétique seront importants (Krishnamurthy 2003; Lefevre et al. 2016; Reece et al. 2012).

D'ailleurs, il peut arriver que ces deux situations se produisent une à la suite de l'autre. C'est le cas de l'effet fondateur, où un goulot d'étranglement est suivi d'une expansion populationnelle. L'effet fondateur se produit lorsqu'une nouvelle population est créée à partir d'un petit sous-groupe d'une grande population. Le patrimoine génétique de la nouvelle population est un sous-ensemble du patrimoine génétique de la population d'origine et la diversité génétique est moindre. Un cas connu d'effet fondateur est celui du Québec, où plusieurs maladies génétiques rares ont une prévalence plus élevée qu'en France, la France étant le pays d'origine des individus colonisateurs (Krishnamurthy 2003; Lefevre et al. 2016; Reece et al. 2012; Scriver 2001). Par exemple, la fibrose kystique est un cas connu de maladie rare ayant une fréquence plus élevée au Saguenay-Lac-St-Jean en raison d'un effet fondateur (Daigneault et al. 1991).

1.1.1.6. *Migration*

La migration survient quand des individus d'une population migrent vers une autre population. Le matériel génétique d'une population est donc transféré à une autre population (Henry et al. 2008). Dans la situation où les fréquences alléliques ne sont pas les mêmes entre les populations, ce transfert cause un changement au niveau des fréquences alléliques au sein de la population hôte de cette migration. La migration permet de contrer, en partie, la dérive génétique et ainsi, de limiter la différenciation populationnelle causée par la dérive génétique (Hartl 1988).

Plusieurs modèles de migrations existent, dont le modèle de migration unidirectionnelle, le modèle en îles et le modèle d'isolement par la distance. La migration unidirectionnelle est caractérisée par un flux de migration constant provenant d'un continent à une île. À long terme, les fréquences alléliques présentes sur l'île vont tendre vers celles du continent.

Le deuxième modèle est le modèle en îles. Ce modèle comprend plusieurs populations indépendantes se trouvant sur diverses îles (Sewall Wright 1943). La migration est possible et semblable entre toutes les îles. Ainsi, les fréquences alléliques vont tendre vers la moyenne des fréquences alléliques initiales de l'ensemble des îles (Hartl 1988). Le troisième modèle de migration est le modèle d'isolement par la distance. Ce modèle stipule que deux populations géographiquement proches seront plus semblables que deux populations géographiquement éloignées (Yoichi Ishida 2009).

1.1.1.7. *Recombinaison génétique*

Un autre phénomène affectant la diversité génétique est la recombinaison génétique. Celle-ci se produit durant la méiose, où un enjambement entre deux chromatides crée des chromosomes recombinés. Ces chromosomes sont les croisements entre les deux chromosomes parentaux, et ils sont donc un mélange de l'ADN maternelle et paternelle. Ainsi, ces nouvelles combinaisons contribuent à la variation et à la diversité génétique. Les nouveaux haplotypes issus de la recombinaison seront sujets à la dérive et aux pressions sélectives. Ces nouveaux haplotypes augmenteront en fréquence s'ils sont bénéfiques ou diminueront en fréquence s'ils sont délétères (Reece et al. 2012). La figure 1.2 démontre ce phénomène d'enjambement.

Il est possible de quantifier la recombinaison avec le taux de recombinaison. Le taux de recombinaison, r, a comme unité de mesure cM/Mb [6] et se définit comme étant la distance génétique entre deux loci. Ce taux varie le long du génome, alternant entre des *hotspots* et des *coldspots* de recombinaison. Un *hotspot* de recombinaison se définit par étant une région où le taux de recombinaison est nettement supérieur à la moyenne observée (Lichten et al. 1995) alors qu'à l'inverse, un *coldspot* est une région où la fréquence de recombinaison est faible (Figure 1.3).

[6]Le centimorgan (cM) est une unité de mesure de distance génétique. 1 cM signifie que 1 % des méioses sont recombinantes et donc, 1 enjambement pour 100 méioses.

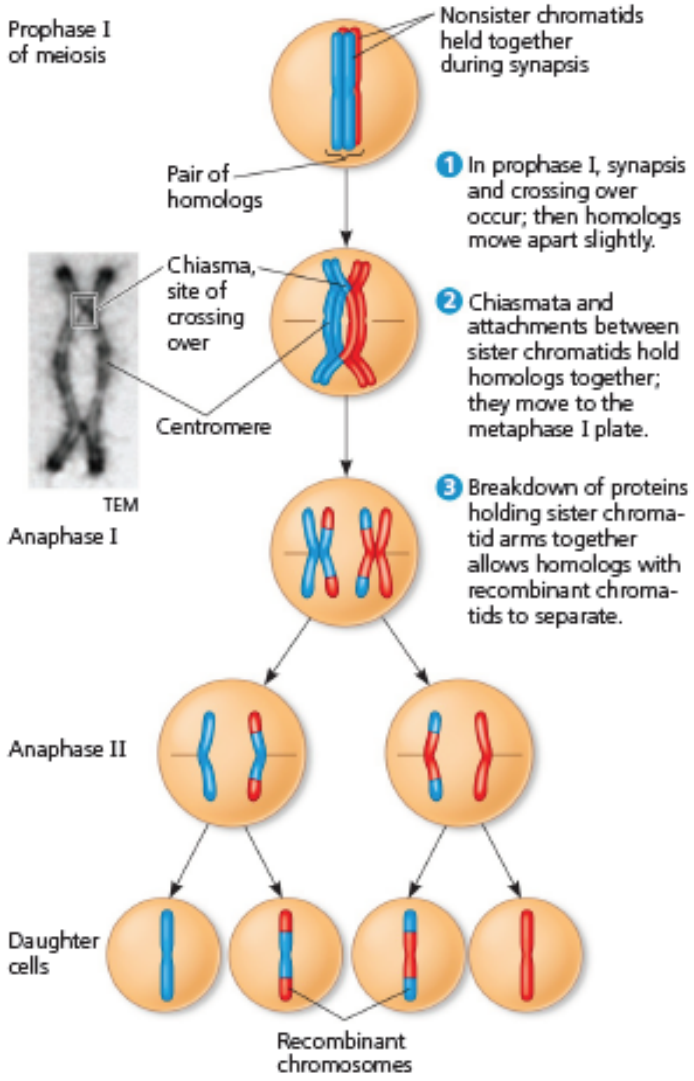


Fig. 1.2. Résultats de l'enjambement durant la méiose. Tirée de (Reece et al. 2012)

1.1.1.8. *Déséquilibre de liaison*

Le déséquilibre de liaison est présent entre les deux allèles à des loci distincts lorsque ces derniers ne sont pas indépendants. En d'autres termes, lorsque deux allèles (A et B) sont transmis ensemble plus souvent qu'attendu, c'est qu'il y a un déséquilibre de liaison (Barnes 2007). Le déséquilibre de liaison peut être mesuré à l'aide de trois variables : D, D' et r^2.

Tout d'abord, le déséquilibre de liaison peut être mesuré à l'aide de la variable D. Cette valeur mesure la différence entre la fréquence observée des allèles (pAB) aux fréquences attendues sous équilibre de liaison (pA et pB) (équation 1.1.3). Lorsque la valeur D est

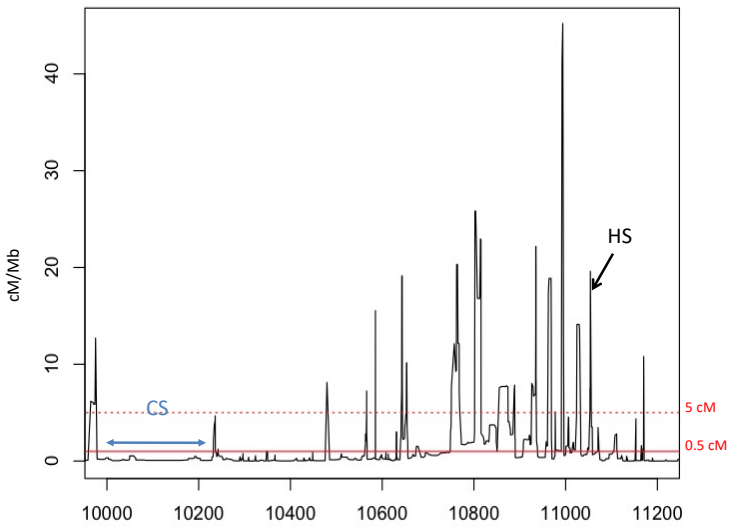


Fig. 1.3. *Coldspots* (CS) et *hotspots* (HS) de recombinaison - Plusieurs *coldspots* de recombinaison se trouvent au début de la région ($< 0.05cM/Mb$) alors qu'un grand nombre de *hotspots* ($> 5cM/Mb$) sont observés à la fin de la région illustrée. Modifiée de (Hussin et al. 2015)

différente de zéro, c'est donc qu'il y a un déséquilibre de liaison, alors qu'un équilibre de liaison donne une valeur D de zéro (Christiansen 2000). Ainsi, la valeur D représente la magnitude du déséquilibre de liaison.

$$D = p_{AB} - p_A + p_B \tag{1.1.3}$$

Ensuite, une autre mesure du déséquilibre de liaison est le D'. Cette mesure est une variante de la valeur D puisqu'elle est normalisée (équation 1.1.4). Lors de cette normalisation, la fréquence des allèles alternatifs (p_a et p_b) des deux loci est prise en compte. Une valeur égale à un indique qu'aucun événement de recombinaison n'a eu lieu et suggère un fort déséquilibre de liaison alors qu'une valeur inférieure à un indique qu'il y a eu des événements de recombinaison entre les loci et suggère donc un faible déséquilibre de liaison. La valeur D' peut cependant être influencée par une petite taille d'échantillon. Par conséquent, les valeurs intermédiaires de D' sont plus difficile à interpréter (Barnes 2007; Foulkes 2009; Lewontin 1964; Zdanowicz et al. 2010).

$$D' = \frac{|D|}{D_{max}} \quad où \quad D_{max} = \begin{cases} min(p_A p_b) & \text{D}> 0 \\ min(p_a p_B) & \text{D}< 0 \end{cases} \tag{1.1.4}$$

Puis, le coefficient de corrélation (r^2) est également une mesure utilisée pour mesurer le déséquilibre de liaison (équation 1.1.5). Une valeur r^2 varie entre zéro et un. Une valeur élevée indique un degré de corrélation élevé et donc, un fort déséquilibre de liaison. Contrairement au D', le coefficient de corrélation ne donne aucune information sur les motifs de recombinaison (Foulkes 2009; Pritchard et al. 2001; Zdanowicz et al. 2010).

$$r^2 = \frac{D^2}{p_A p_B p_a p_b} \tag{1.1.5}$$

Ainsi, les mesures les plus communément utilisées pour mesurer l'étendue du déséquilibre de liaison sont la valeur D' et le coefficient de corrélation r^2. Cependant, la valeur D' n'est pas totalement indépendante des valeurs des fréquences alléliques des deux loci, le coefficient de corrélation est donc plus couramment utilisé (Pritchard et al. 2001).

1.1.2. Statistiques pour la détection de la sélection naturelle

Plusieurs approches statistiques permettent de détecter les signatures de pression sélective. Les approches sont divisées en trois catégories : l'approche par fréquence, l'approche par déséquilibre de liaison et l'approche par différenciation populationnelle. Les approches par fréquence et par différenciation populationnelle détectent les signatures sélectives plus anciennes alors que l'approche par déséquilibre de liaison détecte des signatures plus récentes (Cadzow et al. 2014).

1.1.2.1. *D de Tajima*

La statistique D de Tajima est un test de neutralité permettant de détecter si des signatures sélectives ou des signatures démographiques sont présentes. Cette approche est basée sur la fréquence allélique et compare le nombre moyen de différences entre chaque paire de séquences, représenté par la variable θ_π, au nombre de sites polymorphes, représenté par la variable θ_s (équation 1.1.6) (Crawford 2007).

$$D = \frac{\theta_\pi - \theta_s}{\sqrt{Var(\theta_\pi - \theta_s)}} \tag{1.1.6}$$

Une valeur D de Tajima inférieure à zéro indique un excès d'allèles à faible fréquence. Ce phénomène indique soit une sélection positive, une sélection négative ou une expansion populationnelle. Lorsqu'au contraire, la valeur D de Tajima est supérieure à zéro, c'est qu'un excès d'allèles à fréquence intermédiaire est présent. Ce phénomène suggère une sélection balancée ou un goulot d'étranglement populationnel (Tajima 1989). Ainsi, cette statistique permet de tester l'hypothèse de la neutralité.

1.1.2.2. *Score* β

La détection des régions maintenues à une fréquence intermédiaire peut permettre de découvrir des régions associées avec des phénotypes d'intérêts. Ainsi, une statistique publiée récemment permet de détecter les signatures de sélection balancée dans le génome. Cette statistique est le score β (Siewert et al. 2017). L'approche permet de détecter les *clusters* de variants ayant des fréquences alléliques corrélées.

Tout d'abord, la corrélation des fréquences alléliques doit être mesurée. Pour ce faire, la fréquence des deux variants doit être connue, représentée par f_i et f_0 dans l'équation 1.1.8 et 1.1.9. La variable n représente le nombre de chromosome échantillonné et la variable p, dans l'équation 1.1.9, est une constante. L'équation 1.1.7 permet d'obtenir la fréquence de l'allèle dérivé alors que l'équation 1.1.8 permet de déterminer la différence maximale entre la fréquence allélique dérivée des deux variants. L'équation 1.1.9, quant à elle, permet de mesurer la similarité en fréquences, représentée par la variable d_i.

$$g(f) = min\,(f, n - f) \tag{1.1.7}$$

$$m = max\left(g(f_0), \frac{n}{2} - g(f_0)\right) \tag{1.1.8}$$

$$d_i = \left(\frac{m - |g(f_0) - g(f_i)|}{m} \right)^p \tag{1.1.9}$$

Par la suite, les prochaines équations permettent de calculer le score β. L'estimateur $\hat{\theta_\beta}$ (équation 1.1.11) est calculé en multipliant S_i, soit le nombre de variant dérivé dans une fenêtre de n chromosome, par i qui est le nombre de fois que les allèles dérivés ont été trouvés et il est pondéré en fonction de la similarité des fréquences allélliques (d_i). La variable $\hat{\theta_w}$ (équation 1.1.12) est l'estimateur Watterson (Watterson 1975). Les deux estimateurs sont ensuite comparés (équation 1.1.10). Une valeur β près de zéro indique la neutralité alors qu'une valeur supérieure à zéro indique une sélection balancée.

$$\beta = \hat{\theta_\beta} - \hat{\theta_w} \tag{1.1.10}$$

$$\hat{\theta_\beta} = \frac{\sum_{i=1}^{n-1} i d_i S_i}{\sum_{i=1}^{n-1} d_i} \tag{1.1.11}$$

$$\hat{\theta_w} = \frac{\sum_{i=1}^{n-1} S_i}{\sum_{i=1}^{n-1} \frac{1}{i}} \tag{1.1.12}$$

Le score β peut être complémentaire à la statistique D de Tajima. En effet, comme décrit précédemment, une valeur D de Tajima peut suggérer une sélection balancée ou un goulot d'étranglement populationnel. Le score β permet donc de valider ou d'infirmer la présence de sélection balancée lorsque les valeurs D de Tajima obtenues sont supérieures à zéro.

1.1.2.3. *Integrated haplotype score (iHS)*

La statistique iHS est une approche par déséquilibre de liaison qui permet de détecter les signatures de sélection positive récentes pour les variants n'ayant pas encore été fixés. Cette statistique est une variante de l'approche *Extended Haplotype Homozygosity* (EHH). La statistique EHH détecte la transmission d'un haplotype étendu n'ayant pas recombiné. La particularité de la signature positive détectée est que la fréquence allélique a augmentée rapidement sans permettre à l'haplotype d'avoir eu le temps d'accumuler des événements de recombinaisons. Le déséquilibre de liaison est ainsi élevé pour une région étendue (Sabeti et al. 2002), mais uniquement pour l'allèle sous sélection. Pour la statistique iHS, les variables

$iHHA$ et $iHHD$ provient de la statistique EHH pour l'allèle ancestral et l'allèle dérivé (équation 1.1.13).

$$iHS(nonstandardisé) = \ln\left(\frac{iHH_A}{iHH_D}\right) \tag{1.1.13}$$

Lorsque la valeur iHS est grande et positive, ceci indique de longs haplotypes portant l'allèle ancestral alors qu'une valeur iHS grande, mais négative, indique de longs haplotypes portant l'allèle dérivé. Comme une petite fréquence allélique est généralement associé avec un haplotype plus étendu, la fréquence allélique des variants va avoir un influence sur la valeur iHS. Ainsi, la valeur iHS est normalisée (équation 1.1.14) afin que la moyenne soit zéro et que la variance soit un, indépendamment de la fréquence allélique. La valeur normalisée iHS permet de mesurer si l'haplotype autour d'un variant est inhabituel (Voight et al. 2006).

$$iHS = \frac{\ln\left(\frac{iHH_A}{iHH_D}\right) - E_p\Big[\ln\left(\frac{iHH_A}{iHH_D}\right)\Big]}{SD_p\Big[\ln\left(\frac{iHH_A}{iHH_D}\right)\Big]} \tag{1.1.14}$$

D'ailleurs, la détection des signaux de balayage sélectif en cours est importante puisqu'elle indique la présence de variants ayant des effets significatifs sur les phénotypes humains (Voight et al. 2006).

1.1.2.4. *Indice de fixation (F_{ST})*

L'index de fixation, aussi appelé F_{ST}, permet de mesurer la structure populationnelle ainsi que la différenciation entre différentes populations (S. Wright 1984). Afin de mesurer la valeur F_{ST}, l'hétérozygotie calculée à partir des fréquences alléliques moyennes des populations (Ht) est comparée à l'hétérozygotie de chaque population (HS) (équation 1.1.15). Par conséquent, la statistique quantifie la réduction de l'hétérozygotie associée à la structure des populations en comparant avec l'hétérozygotie attendue pour une population idéale.

$$F_{ST} = \frac{H_T - H_S}{H_S} \tag{1.1.15}$$

Si la valeur F_{ST} est petite cela indique que les fréquences alléliques sont semblables entre les populations, et donc que les populations sont similaires. Si, au contraire, la valeur F_{ST} est grande, cela indique que les fréquences alléliques des populations sont différentes (Holsinger et al. 2009). De surcroît, la sélection influence la valeur de F_{ST}. En effet, si la sélection positive est présente à un locus d'une population, mais pas dans les autres, la fréquence allélique va différer entre les diverses populations et la valeur F_{ST} va augmenter.

L'indice de fixation peut également être utilisé pour identifier des sites qui sont potentiellement sous sélection balancée. Effectivement, lorsqu'il y a une sélection balancée dans plus d'une population, les valeurs F_{ST} de plusieurs loci adjacents seront très faibles, soient près de zéro. Ceci s'explique par le fait que la sélection balancée maintient les fréquences alléliques de différents loci à une fréquence similaire à travers les populations.

Lorsqu'il n'y a pas de sélection, les valeurs F_{ST} seront plus élevées comme la dérive génétique va avoir un impact sur les fréquences alléliques. Par contre, afin d'avoir un pouvoir de détection suffisant, un grand nombre d'échantillon est nécessaire dans les populations comparées (Weedall et al. 2010).

Dans cette section, nous avons vu que plusieurs phénomènes vont affecter la diversité génétique présente dans les populations. Dans ce mémoire, la force qui nous intéresse le plus est la sélection naturelle et ainsi que la manière dont elle peut façonner les génomes. Dans la section suivante, l'expression génique, qui joue un rôle important dans les processus d'adaptation, sera abordée.

1.1.3. Transcriptomique

La transcriptomique consiste en l'étude du transcriptome, soit l'ensemble des transcrits exprimés par un organisme ou par une cellule. Elle a pour but de répertorier l'ensemble des produits transcriptionnels du génome tels les acides ribonucléiques messagers (ARNm) et les ARNs non-codant, de déterminer les structures transcriptionnelles comme le site de départ, les extrémités 5' et 3', les sites d'épissage et les chaînes d'exons et de quantifier les niveaux d'expression (Z. Wang et al. 2009). Bien comprendre le transcriptome est primordial afin d'interpréter les éléments fonctionnels du génome et de comprendre le développement et les maladies. (Z. Wang et al. 2009)

1.1.3.1. *Approches expérimentales*

Deux approches sont utilisées en transcriptomique: les puces (*microarrays*) et le séquençage de l'ARN (*RNA-Seq*) (Lowe et al. 2017). Le séquençage de l'ARN est plus communément utilisé dorénavant comparativement aux *microarrays*. L'analyse des données de ces deux approches requiert de bonnes ressources informatiques ainsi que des connaissances bioinformatiques.

1.1.3.1.1. *Micropuce à ARN*

Une micropuce à ARN, ou *microarray*, consiste à un ensemble de sondes, soit des oligonucléotides, fixées à un substrat ou à une surface. Les données obtenues à l'aide des *microarrays* sont sous forme d'image de haute-résolution où l'intensité de la fluorescence visualisée détermine l'abondance des transcrits. En effet, l'abondance des transcrits est déterminée par hybridation des transcrits marqués par fluorescence à des sondes. Lors de l'analyse de l'image, il est primordial d'identifier et d'éliminer les d'artefacts (Lowe et al. 2017). Un artefact possible est la contamination. La contamination peut être causée, entre autres, par la présence de poussière, d'impureté ou de gouttelette étrangère (Petrov et al. 2004). Un autre type d'artefact possible est lié à la position sur la micropuce. En effet, les

profils d'expression des gènes étant situés près les uns des autres sur la micropuce seront davantage corrélés (Yu et al. 2007).

Une limite de cette approche est qu'une connaissance à priori du transcriptome de l'organisme étudié est nécessaire. En effet, une séquence de référence ou une librairie de transcrits est nécessaire afin de générer les sondes de la micropuce (Lowe et al. 2017).

1.1.3.1.2. *Séquençage de l'ARN de 2^e génération*

Le séquençage de l'ARN, ou *RNA-Seq*, est une approche utilisée en trancriptomique qui permet d'obtenir le profil transcriptomique, en alignant à un génome de référence ou en assemblant *de novo* les *reads* et en quantifiant les transcrits.

Plusieurs protocoles existent afin de cibler les différents types d'ARN. La figure 1.4 illustre trois types de séquençage: queues poly-A, déplétion des ARN ribosomiques et sélection de taille. (Kukurba et al. 2015). Pour la méthode poly-A, la queue poly-A de l'extrémité 3' des molécules d'ARNm est ciblée, puis attachée au substrat. Par la suite, il est important de savoir que la présence de transcrits d'ARNr lors de la construction de la librairie va nuire au séquençage puisque l'abondance d'ARNr va réduire la couverture gobale et rendre plus difficile la détection des ARN moins abondants. Ainsi, le protocole de déplétion des ARN ribosomiques se fait à l'aide d'un *kit* commercial. L'avantage de ce protocole est qu'il permet la détection d'ARN non-codant non-détecté avec le protocole poly-A . Finalement, le protocole de sélection de taille permet de cibler les ARN de petites tailles à l'aide d'une électrophorèse sur gel ou à l'aide d'une colonne (Kukurba et al. 2015).

Par la suite, un ensemble d'ARN est converti en une librairie de fragments d'ADNc avec un adaptateur à une extrémité ou aux 2 extrémités. Ces molécules peuvent être amplifiées si nécessaire (Z. Wang et al. 2009). Elles sont ensuite séquencées à l'aide du séquençage à haut débit. Cette étape est prise en charge par les instruments de séquençage, où les signaux de l'image sont convertis en séquence. Les séquences obtenues ont généralement environ 100 pb, mais elles peuvent varier entre 30 et 10000 pb (Lowe et al. 2017).

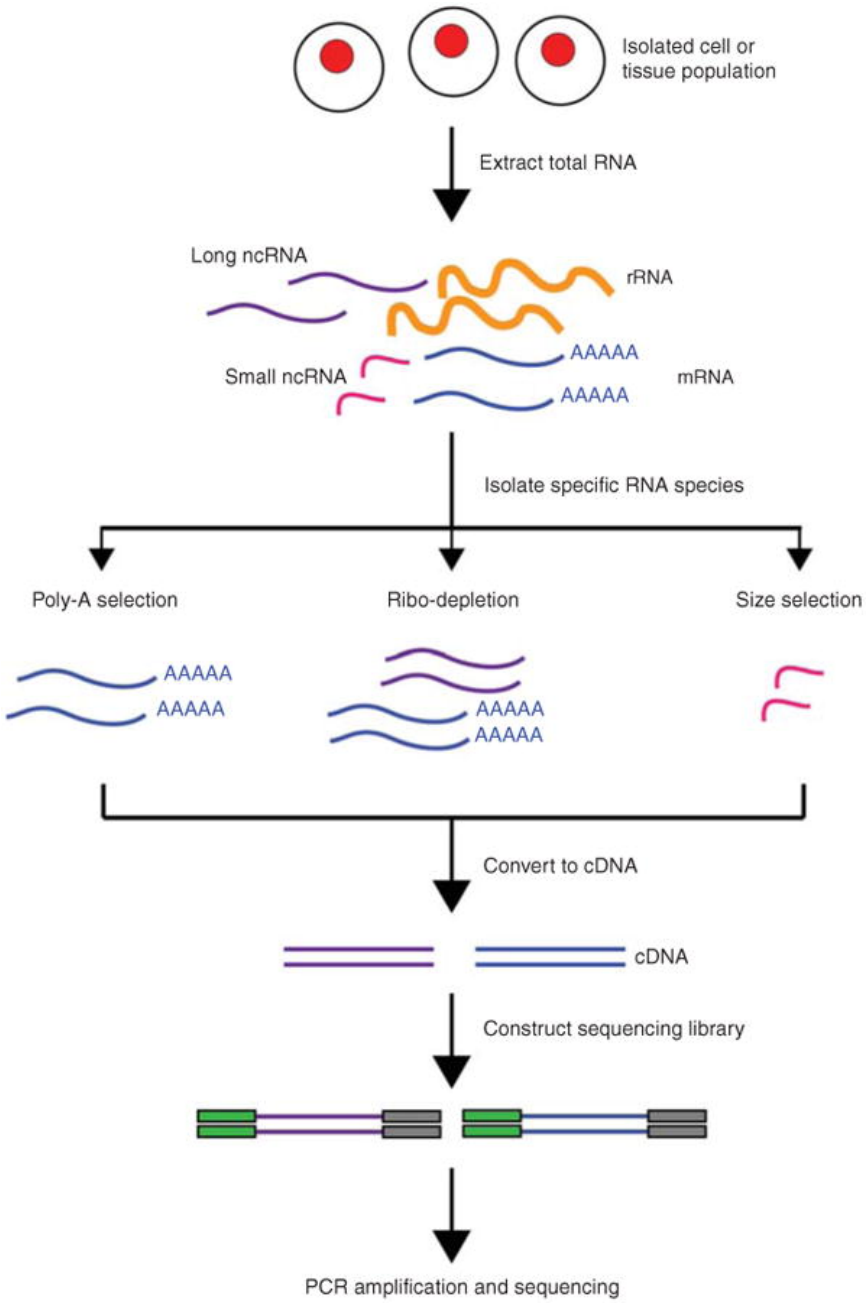


Fig. 1.4. Protocole du séquençage de l'ARN (*RNA-Seq*) - Tirée de (Kukurba et al. 2015).

Suite au séquençage, l'analyse des données obtenues requiert normalement trois étapes: le contrôle de qualité, l'alignement et la quantification. En premier lieu, la qualité des séquences doit être analysée afin d'enlever les erreurs pouvant s'y être insérées. La qualité des séquences est déterminées à l'aide du score de qualité de la base, du taux GC et d'un taux élevé de séquences dupliquées. Le logiciel FastQC (Institute 2021) permet ce contrôle

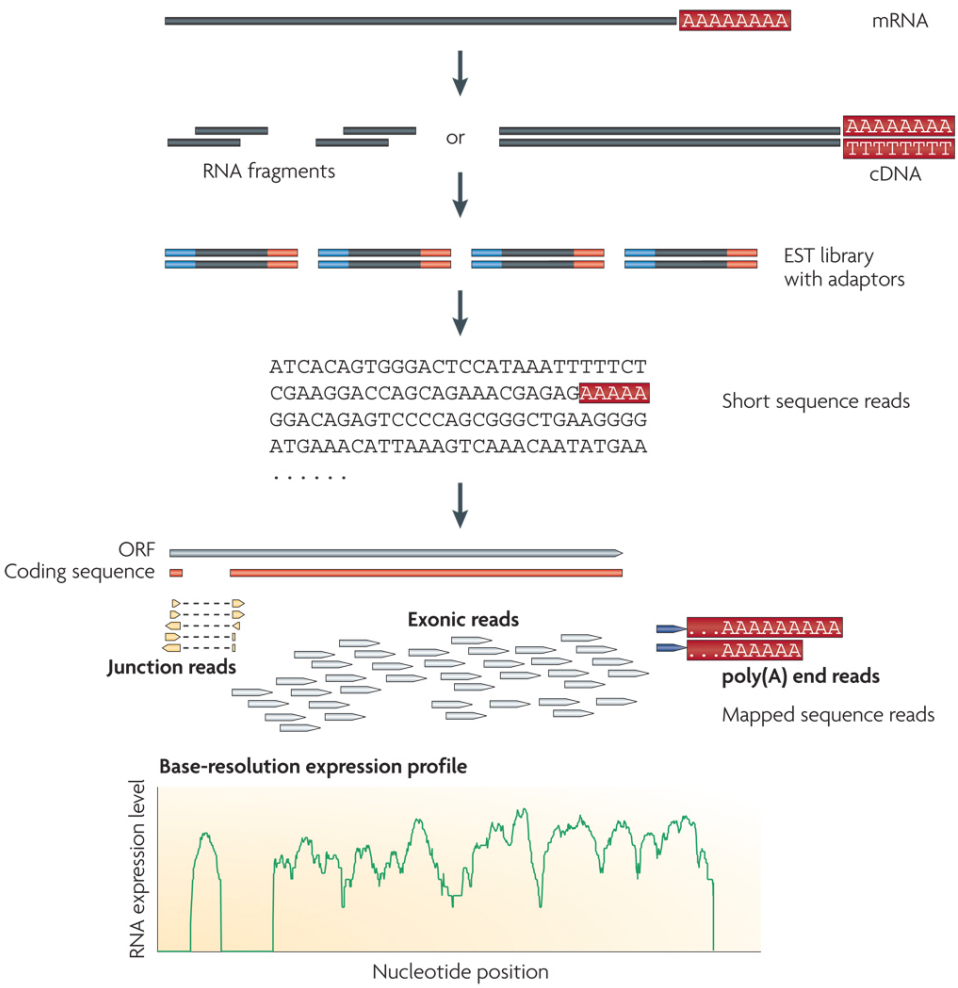


Fig. 1.5. Séquençage de l'ARN (*RNA-Seq*) - Les longs ARN sont d'abord convertis en une banque de fragments d'ADNc par fragmentation. Des adaptateurs de séquençage, illustrés en bleus sur la figure, sont ensuite ajoutés à chaque fragment d'ADNc et une courte séquence est obtenue en utilisant une technologie de séquençage à haut débit à partir de chaque ADNc. Les *reads* sont alignés sur la séquence de référence et classés en trois types: *reads* exoniques, *reads* de jonction et *reads* de queue poly-A. Ces trois types sont ensuite utilisés pour générer un profil d'expression. Tirée de (Z. Wang et al. 2009).

qualité et le logiciel Trimmomatic (Bolger, A. M. et al. 2014) permet de couper les séquences problématiques ou de faible qualité qui ont été identifiées durant le contrôle qualité. Par exemple, si les adaptateur sont présents, ceux-ci peuvent être enlevés à l'aide de Trimmomatic. Ceci va diminuer le nombre de séquences non-alignées et minimiser les erreurs d'alignement. Ensuite, les séquences sont soit alignées sur un génome de référence ou *de novo*, c'est-à-dire sans génome de référence. Il existe des avantages à ces deux approches. Notamment, lorsque le génome de référence est utilisé lors de l'alignement, les erreurs de

séquençage peuvent être détectées et résolues. De plus, comparativement à l'approche *de novo*, l'alignement avec un génome de référence est plus sensible aux transcrits peu exprimé (Lu et al. 2013). Un avantage de l'approche *de novo* est que l'assemblage peut identifier de nouveaux transcrits puisque les transcrits ne sont pas limités à ceux du génome de référence (Lu et al. 2013). Habituellement, l'approche *de novo* va nécessiter plus de ressources informatiques que l'approche avec génome de référence. Ainsi, comme les deux approches ont des avantages distincts, il peut être avantageux de combiner les deux approches lorsque le génome de référence est disponible (Lu et al. 2013). D'ailleurs, lors de cette étape d'alignement, il est possible d'ajouter une étape supplémentaire afin d'identifier les jonctions d'épissage, ce qui aide à prévenir le mauvais alignement des séquences.

Par la suite, les séquences sont quantifiées. Lors de cette étape, les séquences qui sont alignées à plusieurs endroits sont identifiées afin d'être enlevées ou d'être alignées à l'endroit le plus probable. Finalement, l'expression génique est mesurée et normalisée (Lowe et al. 2017).

Un des avantages de cette approche, comparativement à l'approche avec micropuce, est que la détection des transcrits est possible même sans séquence de référence. De surcroît, le bruit est très faible comparativement aux *microarrays* (Lowe et al. 2017; Z. Wang et al. 2009). Toutefois, il existe également des défis avec cette approche. Notamment, la construction de la librairie nécessite parfois la fragmentation des ARNs s'ils sont trop long, c'est-à-dire si leurs tailles excèdent celle de la technologie de séquençage, soit généralement entre 200 et 500 pb. Cette fragmentation peut causer un biais positionnel, où la fragmentation se produit plus souvent au centre de la séquence plutôt qu'aux extrémités 5' et 3' (Z. Wang et al. 2009). De plus, l'amplification en chaîne par polymérase (PCR) peut causer des artefacts. Ces artefacts peuvent être identifés lorsqu'il y a plusieurs mêmes séquences courtes. En effet, l'amplification en chaîne par polymérase (PCR) peut générer des séquences dupliquées puisqu'elles proviennent de la même molécule. Il y a également des défis au niveau bioinformatique. Par exemple, un transcriptome peut contenir un grand nombre de séquences qui s'aligne au même endroit, ce qui complique l'étape de la quantification (Z. Wang et al.

2009). En somme, bien que cette approche comporte plusieurs avantages comparativement aux micropuces, ces analyses sont complexes et sujettes à plusieurs biais.

1.1.3.2. *Effet des variants génétiques sur le transcriptome*

Lors de l'analyse des données de transcriptomique, une approche statistique utilisée consiste à identifier la présence de locus de traits quantitatifs d'expression, communément appelé *eQTL*. Un *eQTL* est un locus de caractères quantitatifs (QTL) où le caractère quantifié est l'expression génique. En effet, un *eQTL* se défini comme étant un locus génétique qui explique la variance de l'expression génique. Ainsi, l'analyse cherche à déterminer si des variants génétiques sont associés à la variation de l'expression génique en faisant un test d'association (Nica et al. 2013).

Cette approche permet donc de déterminer si des régions régulent l'expression génique (Gilad et al. 2008). Cette régulation, ainsi que les *eQTLs*, peuvent être proximal (cis-*eQTL*) ou distal (trans-*eQTL*). Les études indiquent que la majorité de la régulation génique se fait localement, soit près du gène régulé (Stranger et al. 2007). Il est d'ailleurs plus difficile de détecter les trans-*eQTLs* et il requiert davantage de ressources informatiques afin de tester les associations qui seraient distales. De plus, il est attendu que l'ampleur de l'effet (*effect size*) est moins élevée pour un trans-*eQTL* comparativement à un cis-*eQTL* (Dixon et al. 2007). Ainsi, bien qu'il semble y avoir plus de cis-*eQTL* que de trans-*eQTL*, la proportion cis-*eQTL*/trans-*eQTL* reste à déterminer.

En premier lieu, les données de transcriptomique doivent être normalisées afin de corriger pour de possibles biais (L. Li et al. 2012). Ces biais peuvent provenir d'erreurs systématiques. Un exemple d'erreur systématique est l'effet de lot (*batch effect*)[7]. Suite à la normalisation des données, l'association entre les variants et les niveaux d'expression sera testée à l'aide d'une régression linéaire (L. Li et al. 2012). Il est cependant important de prendre en compte la présence d'effets confondants et de corriger pour ceux-ci. Un moyen de corriger pour les facteurs confondants est à l'aide des facteurs PEER (Stegle et al. 2012). Ces derniers

[7]Un effet de lot se produit lorsque des facteurs techniques ou expérimentaux (ex: instruments de séquençage utilisés, conditions du laboratoire) provoquent des changements dans les données des échantillons, ce qui peut mener à des fausses conclusions.

déterminent les facteurs cachés qui expliquent la variance observée. Ces facteurs sont ensuite ajoutés comme covariable lors de l'analyse et permettent d'augmenter le pouvoir de détection.

La correction pour tests multiples doit être appliquée lors de cette analyse. En effet, dans le cas où une analyse effectue des tests multiples, plus le nombre de tests est élevé, plus il y a de risque d'avoir une erreur de type I, autrement appelée un faux-positif (Herzog et al. 2019). Dans le cas où la correction n'est pas effectuée, le résultat obtenu pourrait être dû qu'au hasard et non être un «vrai» résultat. Une méthode utilisée est la correction de Bonferroni (Bland et al. 1995) qui corrige en fonction du nombre de tests effectués. Cette correction diminue le seuil de signification afin de prendre en compte l'augmentation des erreurs de type I causée par les tests multiples et se fait en divisant le niveau alpha par le nombre de tests effectués. Le niveau alpha est généralement de 0.01 ou 0.05. Une autre méthode pouvant être utilisée pour la correction des tests multiples, est de déterminer la valeur p empirique à l'aide des permutations (Churchill et al. 1994). Cependant, cette méthode peut être coûteuse en temps.

En somme, dans cette section, nous avons vu que certains variants peuvent avoir une influence sur l'expression génique et donc, réguler l'expression. Ces *eQTLs* peuvent d'ailleurs être sous pression sélective en fonction de l'impact engendré par la variation d'expression (Quiver et al. 2018). La section suivante décrira les cytochromes P450, ainsi que deux sous-familles en particulier, les sous-familles CYP3A et CYP4F. Puisqu'elles seront le sujet de nos analyses du chapitre 2, l'état des connaissances actuelles sera présenté.

1.1.4. Cytochromes P450

Les cytochromes P450 seraient apparus il y a plus de 3.5 milliard d'années. Ils sont présents chez les champignons, les plantes, les bactéries, les animaux et les humains (G. W. M. Chang et al. 1999). La famille de gènes des cytochromes P450 démontre diverses fonctions, comme la détoxication de l'organisme des médicaments et des molécules xénobiotiques. Chez l'humain, cette famille comporte 57 gènes et 58 pseudogènes (David R Nelson et al. 2004). Une nomenclature a été mise en place afin de bien nommer et classifier les gènes CYP450. Les gènes sont regroupés en famille et en sous-famille en fonction de la similarité de leur séquence. Afin que les gènes soient regroupés en une famille, la similarité de leur séquence protéique doit être supérieure à 40 % et pour être regroupés en sous-famille, la similarité de leur séquence protéique doit être supérieure à 55 % (D. R. Nelson et al. 1996). Au total, il y a 18 familles. Les gènes sont numérotés en ordre chronologique de découverte peu importe dans quelle espèce s'il se trouve. C'est pourquoi, chez l'humain par exemple, les gènes d'une même sous-famille n'ont pas toujours des numéros consécutifs. Ces gènes ont la particularité d'être fortement polymorphes et de contenir plusieurs types de polymorphismes, dont l'insertion et la délétion de variants et la variation en nombre de copies (CNV) (M. Ingelman-Sundberg et al. 2007). Certains gènes ont été étudiés en profondeur, comme CYP2C9, CYP2C19 et CYP2D6 (Eichelbaum et al. 2006), mais ce n'est pas le cas pour tous les gènes des cytochromes P450.

Une hypothèse ayant été émise serait qu'une co-évolution existerait entre les gènes des cytochromes P450 et la consommation de plantes (Gonzalez et al. 1990). Les animaux ayant commencé à manger des plantes, les plantes ont commencé à produire la phytoalexine [8]. Cependant, les animaux se sont adaptés en produisant de nouvelles enzymes pour détoxiquer leur organisme, menant ainsi à une co-évolution. À une échelle plus courte, dans le cas des humains par exemple, l'évolution des gènes des cytochromes P450 a comme conséquence la présence de polymorphismes. Cependant, ces polymorphismes peuvent provoquer, entre autres, des réponses indésirables aux médicaments (Gonzalez et al. 1990).

[8]Substance produite par les plantes comme moyen de défense suite à une exposition à des pathogènes ou à un stress

L'expansion de cette famille de gènes est causée par la duplication génique. Il y aurait eu 3 événements majeurs de duplication qui se seraient produits. Le premier événement aurait donné naissance aux familles CYP11 et CYP4, impliquées dans le métabolisme des acides gras, du cholestérol et de ses dérivés. Le deuxième événement aurait donné naissance aux familles CYP19, CYP21 et CYP27, impliquées dans la synthèse endogène des stéroïdes. Ensuite, le dernier événement aurait causé l'expansion des familles CYP2, CYP3 CYP4 et CYP6, impliquées dans le métabolisme des xénobiotiques (Danielson 2002). D'ailleurs, ce phénomène de duplication génique mène également à l'inactivation de certains gènes, et donc à la conversion de ces gènes en pseudogènes (Pan et al. 2016). Lors de l'alignement, les lectures de séquençage de ces séquences dupliquées vont s'aligner à plusieurs endroits différents. Elles pourront être identifiées à l'aide du score d'alignement puisque celui-ci va diminuer et ainsi, être faible ou nul.

De plus, les pseudogènes ne sont pas contraints par la sélection naturelle (David R Nelson et al. 2004). Par conséquent, l'accumulation de mutations se fait plus rapidement que dans les gènes, ce qui mène à une espérance de vie limitée. Ils sont cependant importants pour les analyses de sélection en raison de leur horloge moléculaire[9] particulière (David R Nelson et al. 2004). Cependant, il a été reporté qu'un type de pseudogène, soit les pseudogènes traités[10] pourraient être sous sélection négative (Xu et al. 2016). Notamment, certains pseudogènes pourraient agir en tant qu'élément régulateur et seraient donc fonctionnels (Pink et al. 2011). Ainsi, comme ces pseudogènes sont fonctionnels, ils seraient contraints par la sélection naturelle.

1.1.4.1. *Sous-famille 3A (CYP3A)*

La sous-famille CYP3A comprend 4 gènes : CYP3A4, CYP3A5, CYP3A7 et CYP3A43. Les gènes de la sous-famille CYP3A se situent sur le chromosome 7, les uns près des autres,

[9]L'hypothèse de l'horloge moléculaire stipule que les séquences d'ADN et de gènes évoluent à une vitesse relativement constante et ce, à travers les différents organismes (Motoo Kimura 1987)

[10]Un pseudogène traité est issu d'un événement de rétrotransposition alors qu'un pseudogène non-traité est issu d'un événement de duplication

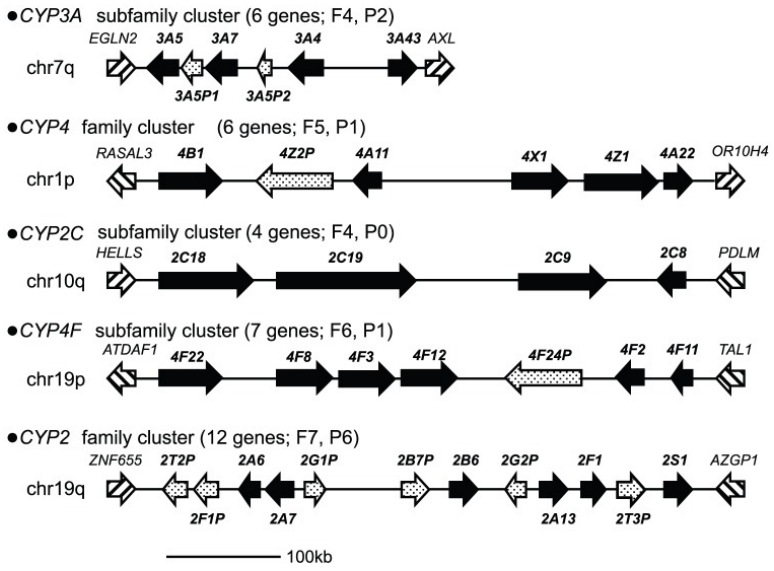


Fig. 1.6. *Cluster* de gènes des cytochromes P450 - Les gènes sont représentés par les flèches pleines alors que les pseudogènes sont représentés par les flèches à pois. La longueur du *cluster* CYP3A est d'environ 250kb alors que les autres sont d'environ 500kb. Tirée de (Kawashima et al. 2014)

comme illustré sur la figure 1.6. Selon l'arrangement des gènes dans la région, ces gènes proviendraient de multiples évènements de duplication (Gellner et al. 2001).

Tout d'abord, CYP3A4 est le gène les plus étudié de cette famille comme il métabolise une grande partie des médicaments sur le marché (Guengerich 1999). Ainsi, il détoxifie l'organisme des xénobiotiques. Il permet à un large spectre de substrat de se fixer à son site actif de telle sorte qu'il métabolise un grand nombre de substrats. Or, cette caractéristique peut causer des interactions médicamenteuses lors de l'administration concomitante de médicaments étant métabolisés par CYP3A4. Il en résulte donc d'une perte d'efficacité du médicament non-métabolisé ainsi qu'une possible toxicité (Ogu et al. 2000; Sevrioukova et al. 2013). Cette toxicité s'explique par le fait que le médicament non-métabolisé s'accumule dans l'organisme.

D'ailleurs, de nombreux inhibiteurs de CYP3A4 existent. Un exemple d'inhibiteur est le jus de pamplemousse (Bailey et al. 1998). Ce dernier augmente la biodisponibilité du médicament causant soit une réaction bénéfique amplifiée, soit une réaction indésirable amplifiée.

Ainsi, l'inhibition de CYP3A4 peut changer la disponibilité de la molécule, causant soit une perte d'efficacité, soit une toxicité.

CYP3A4 est principalement exprimé dans le foie et dans le tractus gastro-intestinal. Par contre, l'activité métabolique de CYP3A4 comporte une variabilité inter-individuelle et elle pourrait être en partie causée par des variants se situant dans CYP3A4. Un exemple est le variant CYP3A4*22[11], où celui-ci cause une diminution de l'expression (Werk et al. 2014). Notamment, le variant provoque une diminution de l'élimination du tacrolimus et de la cyclosporine, deux immunosuppresseurs utilisés lors de transplantation d'organes.

Ensuite, CYP3A5 est également impliqué dans le métabolisme des médicaments et il est le principal gène de la sous-famille CYP3A à être exprimé hors du foie et des intestins(Kuehl et al. 2001). CYP3A5 est également exprimé dans les glandes surrénales, le colon, l'oesophage, le pancréas, les poumons, la prostate, la peau, l'estomac et le vagin[12]. Il est également exprimé dans le foie et dans l'intestin grêle chez les porteurs de CYP3A5*1 (Kuehl et al. 2001). La particularité de ce variant est que sa fréquence varie fortement entre les diverses ethnicités, dont entre les caucasiens et les afro-américains. En effet, ce variant est plus fréquent chez les afro-américains que chez les caucasiens.

L'expression de CYP3A5 varie donc en fonction des allèles, où CYP3A5*1 est considéré comme un allèle permettant l'expression alors que CYP3A5*3, CYP3A5*6 et CYP3A5*7 sont considérés comme des allèles causant une faible expression ou même aucune expression (Bains et al. 2013).

Par la suite, CYP3A7 a initialement été identifié au niveau fœtal puisqu'il est majoritairement exprimé à ce moment du développement (Kitada et al. 1985; Komori et al. 1989). Comme CYP3A4 et CYP3A5, il est impliqué dans le métabolisme des médicaments et des xénobiotiques. Ainsi, il a comme rôle potentiel de protéger le fœtus à ces expositions (H. Li et al. 2019). Par ailleurs, il métabolise deux composés importants dans le développement et

[11]La nomenclature utilisée en pharmacogénétique (*star-allele nomenclature*) identifie les variants à l'aide d'un symbole étoile ainsi qu'un chiffre. Une lettre peut également suivre ce chiffre (ex: CYP3A5*3B). *1 désigne l'allèle de référence.

[12]GTEx portal: `https://gtexportal.org/home/gene/CYP3A5`

la croissance du foetus et du nouveau-né, soit le sulfate de déhydroépiandrostérone (SDHEA) et l'acide tout-trans-rétinoïque (ATRA).

Une particularité au niveau de l'expression génique se trouve entre le gène CYP3A7 et CYP3A4, où CYP3A7 est actif dans le foie fœtal alors que CYP3A4 est actif dans le foie adulte. L'activité de CYP3A7 est maximale durant le développement fœtal, puis diminue suite à la naissance. Contrairement à CYP3A7, CYP3A4 a une faible activité à la naissance et celle-ci augmente après la naissance. Cela suggère que la forme fœtal (CYP3A7) se fait remplacer par la forme adulte (CYP3A4) durant la période néonatale (Lacroix et al. 1997). Toutefois, l'expression de CYP3A7 n'est pas exclusive au fœtus et l'expression de CYP3A4 n'est pas exclusive à l'adulte (He et al. 2016). En effet, de faibles niveaux d'expression de CYP3A7 dans le foie adulte ont été détectés chez les porteurs du variant CYP3A7*1C (Sim et al. 2005). Une hypothèse stipule que le variant causerait un changement dans le promoteur de CYP3A7 pour une région de promoteur de CYP3A4. Néanmoins, l'expression de CYP3A7 dans le foie adulte n'est pas seulement expliquée par CYP3A7*1C et reste à être investiguée davantage (Kuehl et al. 2001). D'ailleurs, CYP3A7 est principalement exprimé dans le foie. Il est également exprimé dans les tissus mammaire[13].

CYP3A7 est associé à plusieurs maladies dont la leucémie lymphoïde chronique, le cancer du sein et du poumon et le syndrome des ovaires polykystiques (H. Li et al. 2019). Ces associations seraient en partie causées par la diminution des niveaux de SDHEA qu'entraîne le génotype CYP3A7*1C (Smit et al. 2005).

Finalement, le gène CYP3A43 est le dernier membre de la sous-famille CYP3A à avoir été découvert et est moins connu (Domanski et al. 2001). Bien que son rôle n'est pas établi, une fonction connue de CYP3A43 est la conversion de la testostérone vers sa forme moins active. Il aurait donc un rôle dans le métabolisme de la testostérone (Thompson, Kuttab-Boulos, Yang, et al. 2006). D'ailleurs, un SNP se trouvant dans CYP3A43 a été identifié comme étant associé au risque de cancer de la prostate chez les africain-américains (Stone et al. 2005). La fonction de ce polymorphisme est cependant inconnue.

[13]GTEx portal: `https://gtexportal.org/home/gene/CYP3A7#geneExpression`

Comparativement aux autres gènes de la sous-famille CYP3A, CYP3A43 est peu exprimé. De faibles niveaux d'expression ont été détectés dans le foie, la prostate et les testicules et les niveaux d'expression de CYP3A43 les plus élevés ont été détectés dans la prostate (Gellner et al. 2001). Cependant, des données plus récentes, soit les données du projet GTEx (section 1.3.2), indiqueraient plutôt que CYP3A43 serait davantage exprimé dans le foie et le pancréas que dans la prostate[14]. Comme ces données proviennent du séquençage de l'ARN de 2^e génération, elles sont plus robustes que les données de micropuce de l'étude ayant initialement découvert CYP3A43 en 2001.

Comme les niveaux d'expression sont faibles, une hypothèse émise serait qu'il ne semblerait pas avoir de fonction importante dans le métabolisme des médicaments (Thompson, Kuttab-Boulos, Yang, et al. 2006). Cependant, bien qu'il semble moins impliqué dans le métabolisme des médicaments, un variant a été identifié comme ayant un impact sur le métabolisme du médicament olanzapine, confirmant ainsi l'implication de CYP3A43 dans le métabolisme des médicaments. En effet, le variant rs472660 cause des différences dans le métabolisme de l'olanzapine. Chez les individus africain-américains, l'allèle A est plus fréquent, ainsi que le génotype AA. Ceci fait donc en sorte que plus d'africain-américains ont une clairance relativement élevée et donc qu'ils sont sous le seuil thérapeutique (Bigos et al. 2011).

Ensuite, il existe de nombreux transcrits de CYP3A43 qui sont causés soit par un épissage alternatif ou par un épissage aberrant (Gellner et al. 2001). L'épissage aberrant a été identifié chez l'ensemble des échantillons d'une étude, ce qui indiquerait que c'est une caractéristique de ce gène et non l'effet d'une différence inter-individuelle (Burk et al. 2004). Les transcrits produits par l'épissage peuvent coder une protéine, avoir de la rétention d'introns ou être non-codant.

Des signaux de sélection naturelle ont été observés chez certains gènes de cette sous-famille. En effet, le gène CYP3A5 a un signal de sélection positive (Bains et al. 2013)(Thompson, Kuttab-Boulos, Witonsky, et al. 2004), tout comme CYP3A4 (Qiu et al. 2008). En

[14]GTEx Portal :`https://gtexportal.org/home/gene/CYP3A43#geneExpression`

somme, cette sous-famille est donc fortement impliquée dans le métabolisme des molécules xénobiotiques, dont les médicaments. Cependant, le spectre des substrats est large, causant des interactions médicamenteuses fréquentes et de nombreux polymorphismes ont un impact sur l'expression de ces gènes. De plus, on ignore encore si les pressions de sélection agissent sur les changements au niveau de l'expression de ces gènes.

1.1.4.2. *Sous-famille 4F (CYP4F)*

La sous-famille CYP4F est composée de 6 gènes : CYP4F2, CYP4F3, CYP4F8, CYP4F11, CYP4F12 et CYP4F22. Les gènes CYP4F se trouvent sur le chromosome 19, près les uns des autres. La figure 1.6 permet d'illustrer l'arrangement de cette sous-famille sur le chromosome 19 (Kawashima et al. 2014). Comme ils sont en amas, ceci suggère un phénomène de duplication génique (A. Kalsotra et al. 2006). Cette sous-famille a été étudiée davantage au niveau biochimique, ce qui est au-delà du cadre du mémoire. Ainsi, cet aspect n'est pas abordé en profondeur et les fonctions des gènes sont donc brièvement abordées.

Le gène CYP4F2 est impliqué dans le métabolisme de la w-hydroxylation de l'acide arachidonique et de la vitamine E (Powell et al. 1998; Sontag et al. 2002) et il est exprimé dans le foie, en plus d'être exprimé dans les intestins et les reins[15]. De plus, CYP4F2 est un facteur influençant la dose requise de warfarine, un anti-coagulant communément utilisé. Ce médicament a un index thérapeutique étroit et donc la dose de warfarine doit être méticuleusement ajustée afin d'éviter un risque de saignement. Notamment, des polymorphismes se trouvant dans divers gènes, dont CYP2C9 et VKORC1, influencent la dose requise de warfarine. Cependant, l'effet de CYP4F2 sur la dose requise est plus modéré que les gènes CYP2C9 et VKORC1 (Liang et al. 2012). D'ailleurs, un polymorphisme de CYP4F2, rs2108622, a été identifié comme ayant un impact sur la dose requise de warfarine puisque les porteurs de l'allèle T requièrent une plus grande dose que ceux avec un génotype CC (Singh et al. 2011).

[15]GTEx portal: `https://gtexportal.org/home/gene/CYP4F2#geneExpression`

Ce génotype est plus présent chez les européens que chez les africain-américains (Shendre et al. 2016).

Le gène CYP4F3 est impliqué dans l'inactivation de la leucotriène B4. La leucotriène B4 est impliquée dans le processus de l'inflammation ce qui indique que CYP4F3 aurait potentiellement un rôle dans le contrôle inflammatoire (Christmas, Ursino, et al. 1999). CYP4F3 est également le principal catalyseur de l'oxydation des acides gras époxydés (*fatty acid epoxides*) (Le Quéré et al. 2004). Il y a d'ailleurs des évidences que les acides gras pourraient être des biomarqueurs dans le développement et la progression du cancer des poumons (Liu et al. 2014; Y. Zhang et al. 2014). Suite à ces évidences, un SNP présent dans CYP4F3, rs4646904, a effectivement été identifié comme étant un facteur étiologique du cancer des poumons chez les fumeurs (Yin et al. 2017).

Au niveau de l'expression génique de CYP4F3, il y a deux isoformes connus : CYP4F3A et CYP4F3B. CYP4F3A est exprimé au niveau des neutrophiles et des cellules myéloïdes alors que CYP4F3B est exprimé au niveau du foie et des reins (Christmas, Jones, et al. 2001; Corcos et al. 2012). CYP4F3 serait également exprimé dans le cerveau ainsi que dans les intestins (Kirischian et al. 2012), l'oesophage, le sang et le vagin[16].

Le gène CYP4F8 a été découvert lors d'une étude portant sur les cytochromes P450 dans les vésicules séminales (Johan Bylund, Finnström, et al. 1999). En plus de l'expression détectée dans les vésicules séminales, CYP4F8 serait également exprimé dans la prostate et dans le foie, mais à un plus faible niveau (Johan Bylund, Finnström, et al. 1999). Selon le portail de GTEx, CYP4F8 n'a pas été détecté dans le foie, mais plutôt dans la prostate et dans la peau[17]. D'ailleurs, il y aurait des évidences que CYP4F8 serait impliqué dans le métabolisme des prostaglandines (Johan Bylund, Hidestrand, et al. 2000).

Le gène CYP4F11 est principalement exprimé dans le foie et les reins. Il est également exprimé à plus faible niveau dans le coeur, le muscle squelettique et le placenta (Cui et al.

[16]GTEx Portal: `https://www.gtexportal.org/home/gene/CYP4F3#geneExpression`
[17]GTEx portal: `https://gtexportal.org/home/gene/CYP4F8#geneExpression`

2000). Selon des données plus récentes, CYP4F11 serait également exprimé dans l'oesophage et la prostate[18]. Une étude suggère que des SNPs présents dans CYP4F11 pourraient être associés avec un risque de saignement de l'intestin grêle induit par l'aspirine chez des patients japonnais (Shiotani et al. 2013). En comparant les activités métaboliques de CYP4F11 et CYP4F3, CYP4F3 serait davantage impliqué dans le métabolisme des eicosanoïdes[19] alors que CYP4F11 serait davatange impliqué dans le métabolisme des médicaments (Auinash Kalsotra et al. 2004). Les médicaments métabolisés par CYP4F11 sont l'érythromycine, benzphetamine, l'éthylmorphine, la chlorpromazine et l'imipramine.

Le gène CYP4F12 est impliqué dans le métabolisme de l'acide arachidonique et des prostaglandines et il est exprimé dans le foie, les reins, le colon, l'intestin grêle et le coeur (J. Bylund et al. 2001). CYP4F12 est d'ailleurs exprimé dans divers autres tissus, dont la peau, le vagin, l'oesophage, le colon, la vessie et la prostate[20]. De surcroît, il serait impliqué dans l'hydroxylation de l'ébastine (Hashizume, Imaoka, Hiroi, et al. 2001; Hashizume, Imaoka, Mise, et al. 2002).

Le gène CYP4F22 est impliqué dans la formation de la barrière de perméabilité de la peau, cette dernière étant importante afin de protéger l'organisme. En effet, la barrière de perméabilité protège l'organisme de possibles infections et empêche le corps de perdre une trop grande quantité d'eau. Un lipide important de cette barrière est l'acylceramide. Une hydroxylation est nécessaire lors de la synthèse d'acylceramide et CYP4F22 a été identifié comme étant l'hydroxylase permettant cette l'hydroxylation (Ohno et al. 2015). D'ailleurs, des mutations dans CYP4F22 ont été associées à une diminution de la synthèse d'alcylceramide et à une maladie de peau, l'ichtyose congénitale autosomique récessive (Hotz et al.

[18]GTEx portal: `https://gtexportal.org/home/gene/CYP4F11#geneExpression`
[19]Substance dérivée d'acides gras poly-insaturés à 20 atomes de carbone.
[20]GTEx portal: `https://gtexportal.org/home/gene/CYP4F12#geneExpression`

2018). L'expression de CYP4F22 est donc été détectée dans la peau, en plus d'être détectée dans le vagin et l'oesophage[21].

Un fort déséquilibre de liaison est présent dans cette sous-famille de gènes. En effet, les SNPs présents dans CYP4F pourraient affecter les niveaux d'expression d'un autre gène de la sous-famille CYP4F. Par exemple, le variant rs2108622 se trouvant dans CYP4F2 aurait un impact sur l'expression hépatique de CYP4F11. Dans le même ordre d'idée, le variant rs1060467 se trouvant dans CYP4F11 est associé avec une baisse d'expression de CYP4F2 (J. E. Zhang et al. 2017).

Au niveau phylogénétique, le gène CYP4F22 serait l'ancêtre de la sous-famille CYP4F (Kirischian et al. 2012) et la sous-famille CYP4F proviendrait de la sous-famille CYP3A (Pan et al. 2016).

En somme, cette sous-famille est impliquée dans le métabolisme de certains médicaments, mais plus particulièrement dans la modulation des concentrations des eicosanoïdes, soit la leucotriène B4 et les prostaglandines (Hardwick 2008). Par conséquent, cette sous-famille aurait un rôle dans la réponse aux médicaments et inflammatoire.

[21]GTEx portal: `https://gtexportal.org/home/gene/CYP4F22#geneExpression`

1.2. Hypothèses et objectifs

Mes analyses préliminaires[22] suggèrent des pressions sélectives dans les deux sous-familles et que ces pressions ne seraient pas les mêmes pour les gènes CYP3A que pour les gènes CYP4F. Ainsi, notre hypothèse est que la diversité génétique des gènes a été façonnée par la sélection naturelle et a évolué en fonction de niveaux élevés d'épistasie[23]. Donc, nous supposons que les variants identifiés lors des analyses de sélection naturelle seront associés à des *eQTLs* d'intérêts.

Afin de valider les hypothèses émises, des analyses de génétique des populations et de transcriptomique seront effectuées sur les deux sous-familles sélectionnées.

Le projet comporte donc deux objectifs. Tout d'abord, le premier objectif du projet de recherche est d'évaluer la diversité génétique et les sites sous pression sélective se trouvant dans les gènes de la sous-famille 3A et de la sous-famille 4F des cytochromes P450. Ensuite, le deuxième objectif du projet est d'évaluer l'impact des polymorphismes des gènes CYP3A et CYP4F sur l'expression génique des ensembles de gènes respectifs.

1.3. Jeux de données

1.3.1. Le projet des 1000 Génomes

Le jeu de données public du projet des 1000 Génomes (*The 1000 Genomes Project*) a été utilisé lors de mes analyses. Ce projet a pour but de permettre une meilleure compréhension des variants génétiques présents dans le génome humain. Pour ce faire, 2504 individus de 26 différentes populations ont été séquencés pour l'ensemble de leur génome (Genomes Project et al. 2015). Ces 26 populations sont regroupées en cing grandes populations : Africaine (AFR), Américaine Centrale et du Sud (AMR), Asiatique de l'Est (EAS), Européenne (EUR) et Asiatique du Sud (SAS), comme illustré sur la figure 1.7. Les données proviennent de la troisième phase du projet et ont été alignées sur la version du génome de référence GRCh37 (Sudmant et al. 2015).

[22]Ces analyses réalisées au cours de mon baccalauréat seront abordées dans l'article scientifique qui se trouve au chapitre 2.
[23]Interaction existant entre deux ou plusieurs gènes

Un jeu de données plus récent du projet des 1000 Génomes a également été utilisé (Byrska-Bishop et al. 2021). Il contient de nouvelles données de séquençage à couverture élevée (30x) alignées sur GRCh38. Ce jeu de données contient les individus de la phase 3 du projet, en plus de 698 individus apparentés à ces derniers, ce qui permet d'avoir un total de 602 trios parents-enfant complets.

Fig. 1.7. Populations incluses dans le projet des 1000 génomes. Les 26 populations sont regroupées en cinq grandes populations (super-populations): Africaine (jaune), Américaine Centrale et du Sud (rouge), Asiatique de l'Est (vert), Européenne (bleu) et Asiatique du Sud (mauve). Les sous-populations de chaque super-population sont définies ici: `https://www.internationalgenome.org/faq/which-populations-are-part-your-study/`. Tirée de `https://www.internationalgenome.org/home`

1.3.2. Projet *Genotype-Tissue Expression* (GTEx)

Le projet *Genotype-Tissue Expression* (GTEx) a été mis sur pied afin d'avoir un jeu de données permettant d'étudier la relation entre les variants génétiques et leur impact sur l'expression génique à travers divers tissus (Lonsdale et al. 2013). Les échantillons biologiques proviennent de donneurs décédées. Ainsi, l'implantation d'une infrastructure a été nécessaire pour assurer une qualité des échantillons biologiques suffisamment élevée pour permettre leur analyse (Carithers et al. 2015). Notamment, la collecte des échantillons biologiques doit être effectuée dans les 24 heures suivant le décès du donneur. Bien que les échantillons biologiques

proviennent de donneurs décédés, les tissus sont qualifiés comme étant sain. Plusieurs critères d'éligibilité sont d'ailleurs en place:

- Donneur âgé entre 21-71 ans ;
- Indice de masse corporel entre 18,5 et 35 ;
- Aucune transfusion sanguine dans les 48h précédentes;
- Aucun diagnostique de cancer métastatique;
- Aucun traitement de chimiothérapie ou de radiothérapie dans les deux dernières années;
- Aucun antécédent d'abus de drogue par la voie intraveineuse dans les cinq dernières années;
- Aucune exposition et aucun diagnostic de virus de l'immunodéficience humaine (VIH), d'hépatite B et d'hépatite C.

La version utilisée est la version 8 (v8) du jeu de données. Elle comprend 17382 échantillons biologiques provenant de 54 tissus (Figure 1.8) de 948 individus, dont 834 ont des données de génotypage. Comparativement aux autres jeux de données d'expression disponibles, où l'expression était particulièrement disponible pour différents types de cellule du sang, le projet GTEx comprend deux points forts: son grand nombre de tissus ainsi que le grand nombre de donneurs. Il a permis de démontrer que des *eQTLs* sont trouvés dans presque tous les gènes (Consortium 2020), en plus de décrire les mécanismes moléculaires sous-jacents.

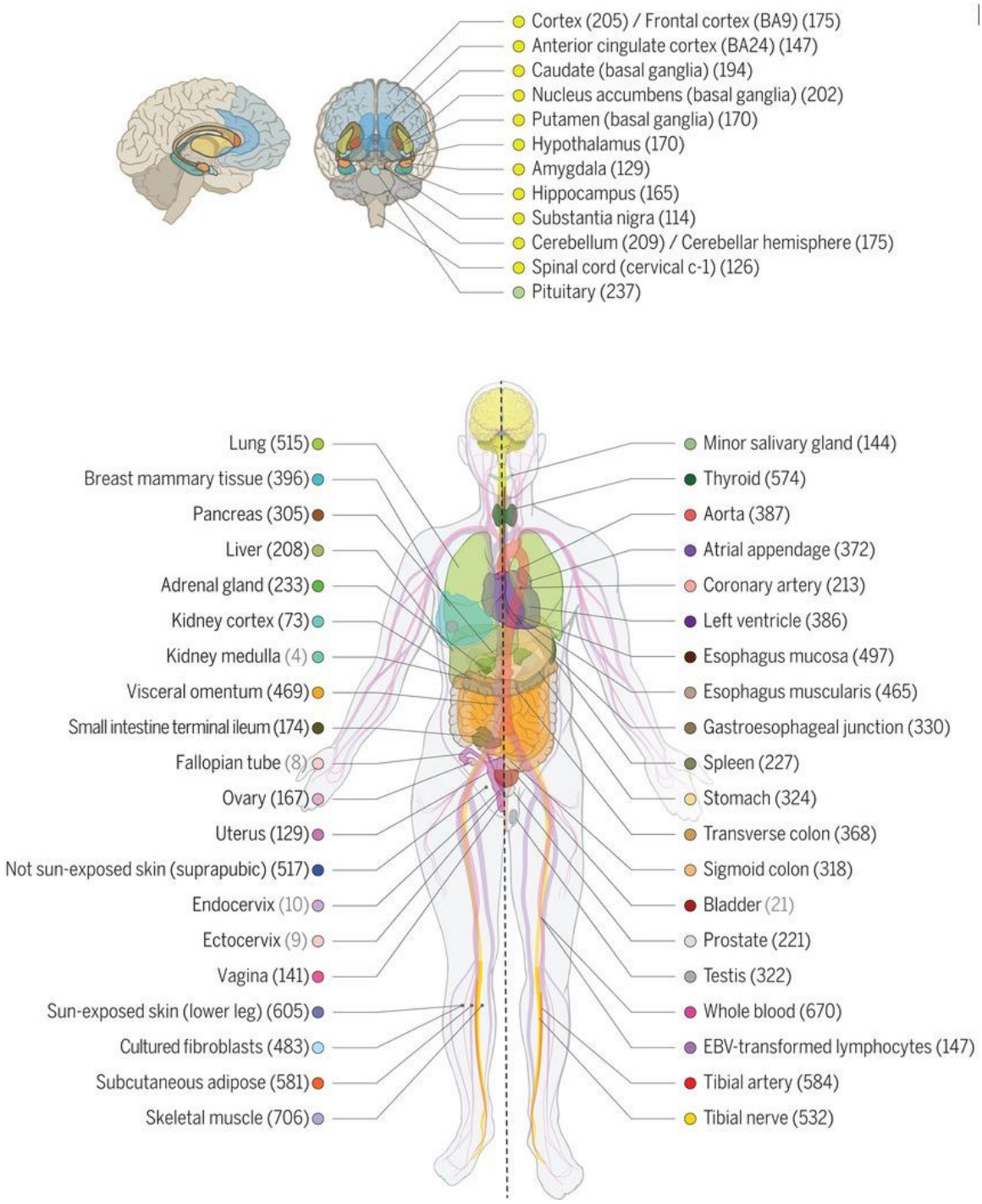


Fig. 1.8. Illustration des 54 types de tissus examinés dans le jeu de données de GTEx. Le nombre d'échantillons de donneurs génotypés est identifié entre parenthèses. Tirée de (Consortium 2020)

Chapitre 2

Article[1]

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

par

Alex Richard-St-Hilaire[a,b], Justin Pelletier[a,b], Isabel Gamache[a,b], Jean-Christophe Grenier[a], Raphael Poujol[a], Julie Hussin[a,c,*]

[a]Montreal Heart Institute, Research Center, Montreal, Qc, Canada

[b]Département de biochimie et medecine moleculaire, Université de Montréal, Montreal, Qc, Canada

[c]Département de Medecine, Université de Montréal, Montreal, Qc, Canada

[*]Corresponding author: Julie, Hussin; julie.hussin@umontreal.ca

[1]En préparation

Author contributions:

- Alex Richard-St-Hilaire did all analyses and prepared the manuscript and figures;
- Justin Pelletier initially did the Beta score analysis on the CYP4F cluster;
- Isabel Gamache normalized the GTEx dataset and calculated PEER factors on the normalized expressions;
- Jean-Christophe Grenier pre-processed genetic data to obtain iHS values and assisted in the preparation of the manuscript;
- Raphael Poujol assisted in the preparation of the manuscript;
- Julie Hussin supervised the project and assisted in the preparation of the manuscript.

2.1. Abstract

Cytochromes P450 (CYP450) are hemoproteins generally involved in the detoxification of the body of xenobiotic molecules and participate in the metabolism of many drugs. Genetic polymorphisms have been found to impact drugs responses and metabolic functions. However, genes encoding CYP450 proteins are often under-analyzed in large-scale genomic studies because the difficulty of analysis due to of their high rate of polymorphism. In this study, we investigate the genetic diversity for 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. The CYP3A subfamily metabolizes approximately 50% of drugs while CYP4F enzymes are involved in the metabolism of endogenous compounds, nutrients and drugs. Indeed, we found signals of recent positive selection in CYP3A and CYP4F genes and signals of balancing selection in CYP4F genes. Futhermore, unusual linkage disequilibrium is detected in both cluster, suggesting co-evolution. eQTLs were also found in both clusters which indicate co-regulation and epistasis.

2.2. 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. Early in the 1960s, 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 oxidations for many structural classes of chemicals (including the majority of drugs) (Danielson 2002; Nebert and Dalton 2006). They metabolically activate parent compounds to electrophilic intermediates, while Phase II enzymes conjugate these intermediates to more easily excretable derivatives.

CYP450 genes are a super-family of genes which appeared more than 3.5 billion years ago, being present in fungi, plants, bacteria, animals and humans (G. W. M. Chang et al. 1999). Genes are grouped into families and subfamilies based on the similarity of their sequence: 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 % (D. R. Nelson et al. 1996). Due to the fact that there are numerous CYP450 genes across species, a nomenclature had to be put in place: the number following the symbol "CYP" represents the family while the following letter represents the subfamily (D. R. Nelson et al. 1996), with the last number represents the individual gene.

In humans, the CYP450 family comprises 57 genes and 58 pseudogenes (David R Nelson et al. 2004) grouped in 18 families (Nebert, Wikvall, et al. 2013). Several CYP450 genes are found in clusters in the human genome, such as CYP2C, CYP3A, CYP4F, but clusters do not always contain all the genes and pseudogenes from their subfamily, as some are spread out across the genome. For example, the CYP4F subfamily has genes on chromosome 19 and pseudogenes on multiple chromosomes. These clusters and families are believed to have appeared due to duplication events (Danielson 2002; D. R. Nelson et al. 1996). The CYP2D6 gene is the most widely studied CYP450 gene in humans, due to its involvement in the metabolism of many drugs (Gaedigk 2013; Gaedigk et al. 2017). CYP3A4 and CYP3A5 are two genes in the CYP3A subfamily that also have been largely studied for the same reason as CYP2D6 (Elens et al. 2012; Lamba et al. 2012; Rojas et al. 2015; Tavira et al. 2013; D.

Wang et al. 2011). 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 can be found in the large list of genes identified in genomic scans as potential targets of natural selection (Carlson et al. 2005; Voight et al. 2006). Other studies of the genetic diversity for specific CYP450 subfamilies in human populations confirmed the presence of signatures of positive (Bains et al. 2013; Qiu et al. 2008), balancing (Janha et al. 2014) or purifying selection (Yasukochi et al. 2015). One example among the CYP450 genes is CYP2C19, involved in the metabolism of clopidogrel (Brown et al. 2018; Scott et al. 2013), where signals of positive selection on CYP2C19 alleles conferring slow metabolism (CYP2C19*2 and CYP2C19*3) were detected using relative extended haplotype homozygosity (REHH)(Janha et al. 2014). CYP2C19*2 is detected worldwide, but CYP2C19*3 is only detected in Asia. The selective advantages may be due to 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 (Janha et al. 2014). 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 function.

Moreover, the detection of signals of natural selection the CYP450 genes suggests that selective advantage occurring at the molecular level. This selective advantage can act on polymorphisms that modulate gene expression, widely known as expression quantitative trait loci (eQTL) (Kudaravalli et al. 2009). Detecting eQTLs linked to selection signals helps clarifying how gene expression is regulated and can lead to a better understanding of variants' biological effects (Nica et al. 2013). Furthermore, analysing eQTLs can assist the detection of gene-gene interactions (Huang et al. 2013) and co-regulation between genes (Lehner 2011). Such gene-gene interactions can also be detected by looking at the patterns of linkage disequilibrium (LD), as evolution will maintain co-evolving polymorphisms on

the same haplotypes (Rohlfs et al. 2010), 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. 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.

2.3. Results

We obtained genotypic data from the 1000 genomes project (1000G) (Genomes Project et al. 2015). A total of 2,157 individuals were analyzed from 22 populations, which were grouped into 4 super-populations (ie. Africa, Europe, East Asia and South Asia). The American populations were excluded from this dataset due to their high admixture levels. A total of 61,739 biallelic SNPs were analyzed in the CYP450 genes. We defined the CYP450 gene lists according to the HUGO Gene Nomenclature Committee (HGNC) (Tweedie et al. 2021) and extracted the genes intervals from the UCSC Genome Browser (Kent et al. 2002).

2.3.1. 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 the 1000G dataset to provide insight into the non-neutral forces that act on these genes. Briefly, for each gene we computed the mean Tajima's D by gene and also in 1 Kb windows. We assessed significance based on the empirical (null) distribution (computed on chromosome 22, see Methods), which allows to determine whether any genes have values that are higher or lower than expected while taking population-specific demographic factors into account (Methods). Significantly low Tajima's D values indicate an excess of rare alleles, which suggests either negative/purifying or positive selection forces affecting

genetic diversity. Significantly high values of Tajima's D suggest an excess of intermediate frequency alleles, which can reflect the occurrence of balancing selection.

As a starting point, we evaluated global genetic diversity in the European populations. Nine genes had Tajima's D values consistently below 0 (Figure 2.1A). The proportion of 1 Kb-windows of each gene lying outside the null distribution is shown in Figure 2.1B. 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 particularly strong purifying selection pressures may be acting in this CYP450 family, however complete selective sweeps driven by positive selection can also create this lack of diversity (Kim et al. 2002). CYP3A5 has a low overall mean Tajima's D 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 an overall 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 in these genes. Of these ten genes, five are in the CYP4F subfamily: CYP4F3, CYP4F11, CYP4F12, CYP4F8 and CYP4F2. Theses genes are large ($>$ 14 Kb) and have a mean Tajima's D above 0 and have several windows showing significantly high Tajima's D values. The strongest of these signals is seen on CYP4F12 (Figure 2.1B). The only CYP4F gene that does not show this specific signature is CYP4F22, the ancestral gene of the CYP4F cluster (Kirischian et al. 2012).

Because population differentiation can also help identify 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 heterozygosity differs between populations. When the F_{ST} value is extreme compared to the rest of the genome, this may also indicate that an allele is favoured in one population compared to the other, since positive selection will increase

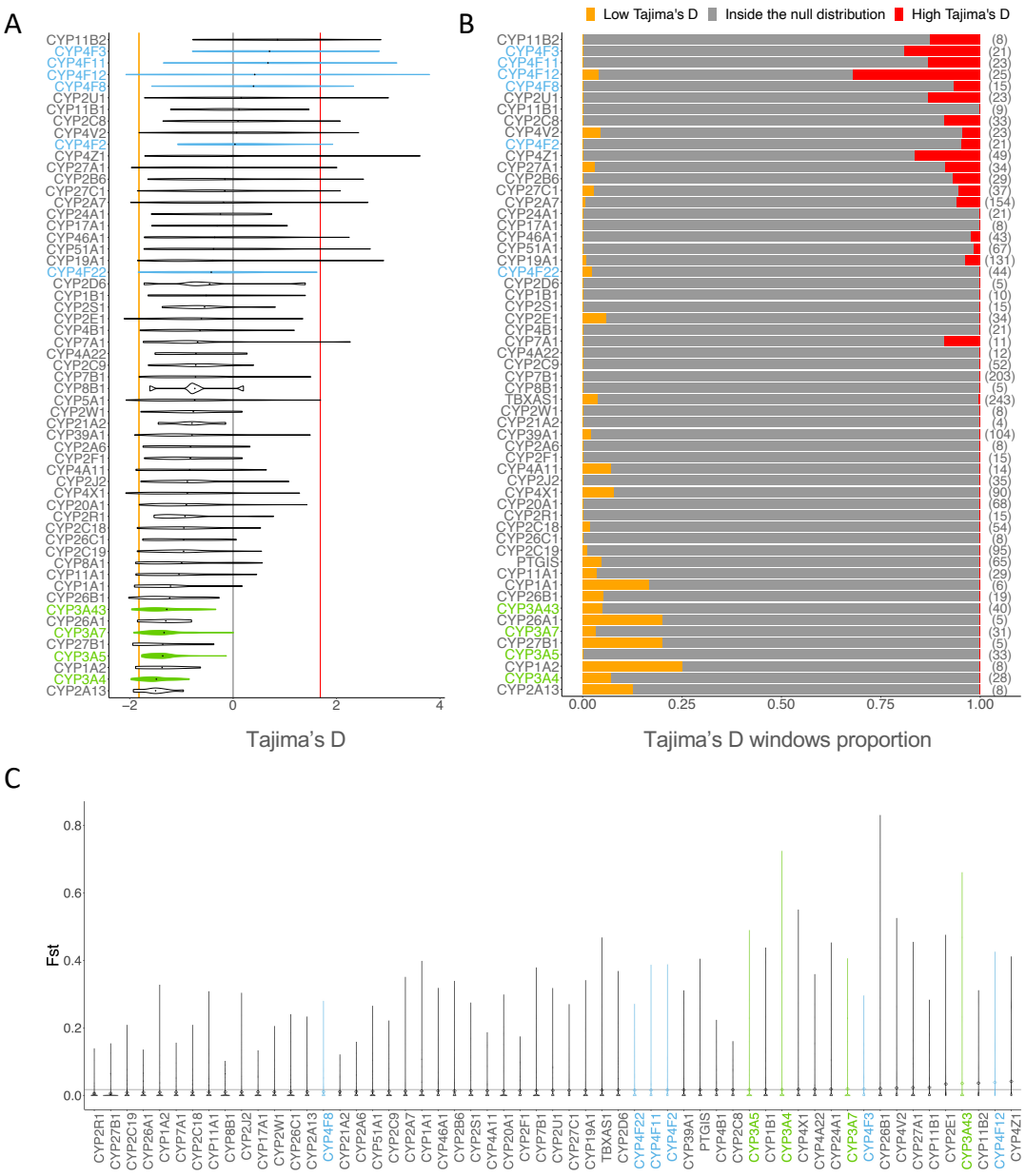


Fig. 2.1. A) Distribution of Tajima's D values computed on windows of 1 Kb for each CYP450 genes in the European populations. The 2.5th percentile is marked by the orange vertical line and the 97.5th percentile is marked by the red vertical line, representing the significance threshold. B) Proportion 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.

population differentiation at a specific locus. Figure 2.1C) 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 notably 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, which would have led to low F_{ST} instead.

2.3.2. Positive selection in CYP3A and CYP4F subfamilies

The global neutrality and differentiation analyses of CYP450 genes presented above lead us to hypothesize that positive selection, either directional (CYP3A) or balancing (CYP4F), is acting on subfamilies of CYP450 genes, possibly in a concerted fashion. Therefore, our next analysis focuses on identifying local signals in each subfamily. 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. Negative iHS scores indicates that the derived allele has swept up in frequency in that population and positive iHS score indicates that, in contrast, it is the ancestral allele (present in the common ancestor of all humans) that has increased in frequency. Typically, an absolute value of iHS greater than 2 at a SNP suggests that the region around the SNP is under selection (Voight et al. 2006).

In the CYP3A cluster, significant iHS values are detected (Figure 2.2A), but signals of positive selection differ between populations. Many signals are detectable in Africa, in East Asia and in Europe, while fewer signals are detectable in South Asia. 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 increased in frequency and are under positive selection. 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. This intergenic region also has a signal in the East Asian population. CYP3A43 is the only gene with signals in all super-populations. These results confirm that positive selective pressure is acting on CYP3A genes.

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 2.2B). 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. The Tajima's D analyses (Figure 2.1) suggested balancing selection in the CYP4F cluster, however iHS may be underpowered to detect these events.

2.3.3. Balancing selection in CYP3A and CYP4F subfamilies

The Beta score (Siewert et al. 2017) has been developed to specifically test whether balancing selection is present at specific loci. This statistic detects clusters of alleles with similar allele frequencies, which is a signature of balancing selection. Indeed, balancing selection maintains multiples alleles, resulting in clusters of SNPs at intermediate allele frequency. When the β score is significantly greater than 0, neutrality is rejected, suggesting the presence of balancing selection acting in the genetic region.

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. The same signal can been seen in the other populations, but it is weaker and do not pass our 1% threshold (Figure 2.3A). Overall, these results confirm 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 shows much more extreme β

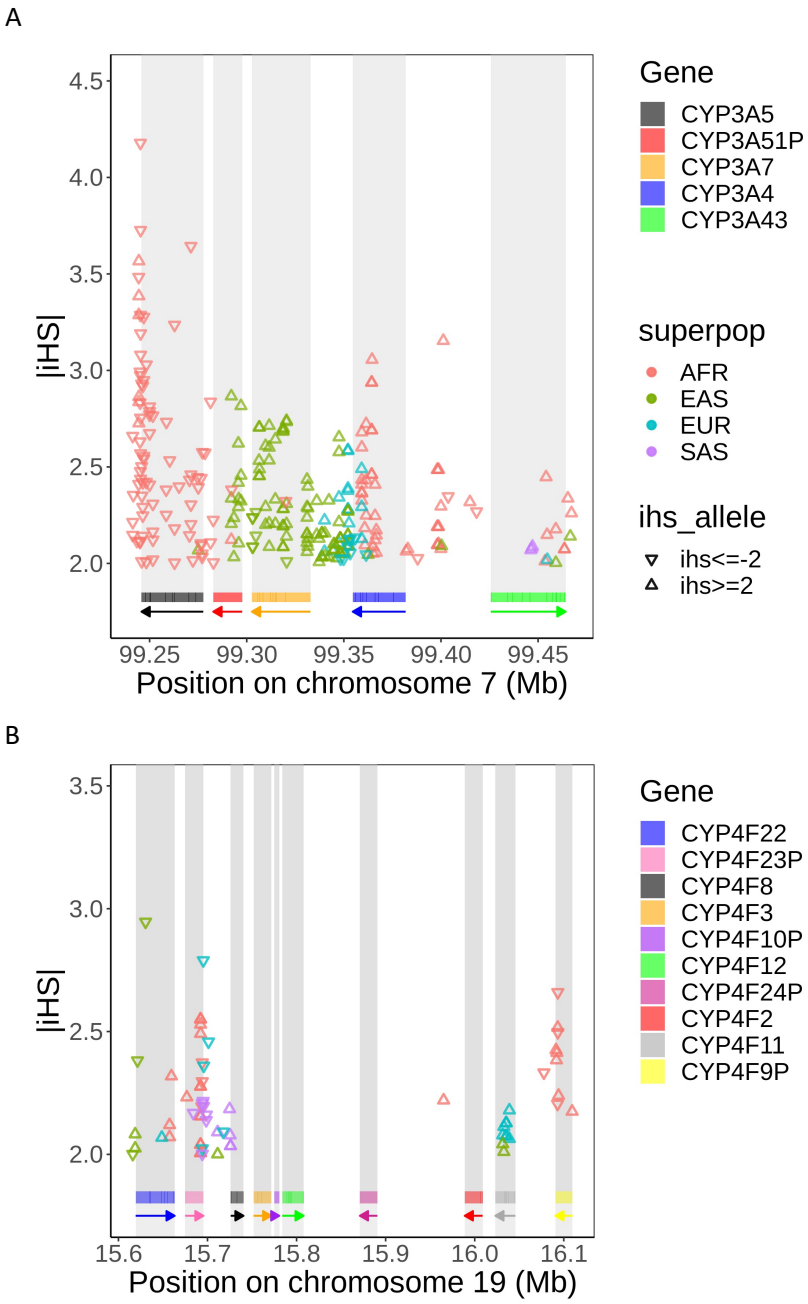


Fig. 2.2. Distribution of SNPs with high $|iHS|$ values ($|iHS| \geq 2$) in the A) CYP3A and B) CYP4F cluster. A triangle standing on its base means an iHS value ≥ 2, indicating that the ancestral allele has increased in frequency, and a triangle standing on its point means an iHS value $\leq$ -2, indicating 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.

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 2.3B). 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 confirm what the global neutrality tests and differentiation analyses suggested, namely that balancing selection does act on the CYP4F cluster, but not in the CYP3A cluster.

2.3.4. 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-occur during evolution can lead too concerted selective pressure, or co-evolution, depending on the resulting fitness of the individuals (Rohlfs et al. 2010). 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 MAF >0.05 in the two CYP450 clusters under investigation, 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 a null 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 potentially co-evolving.

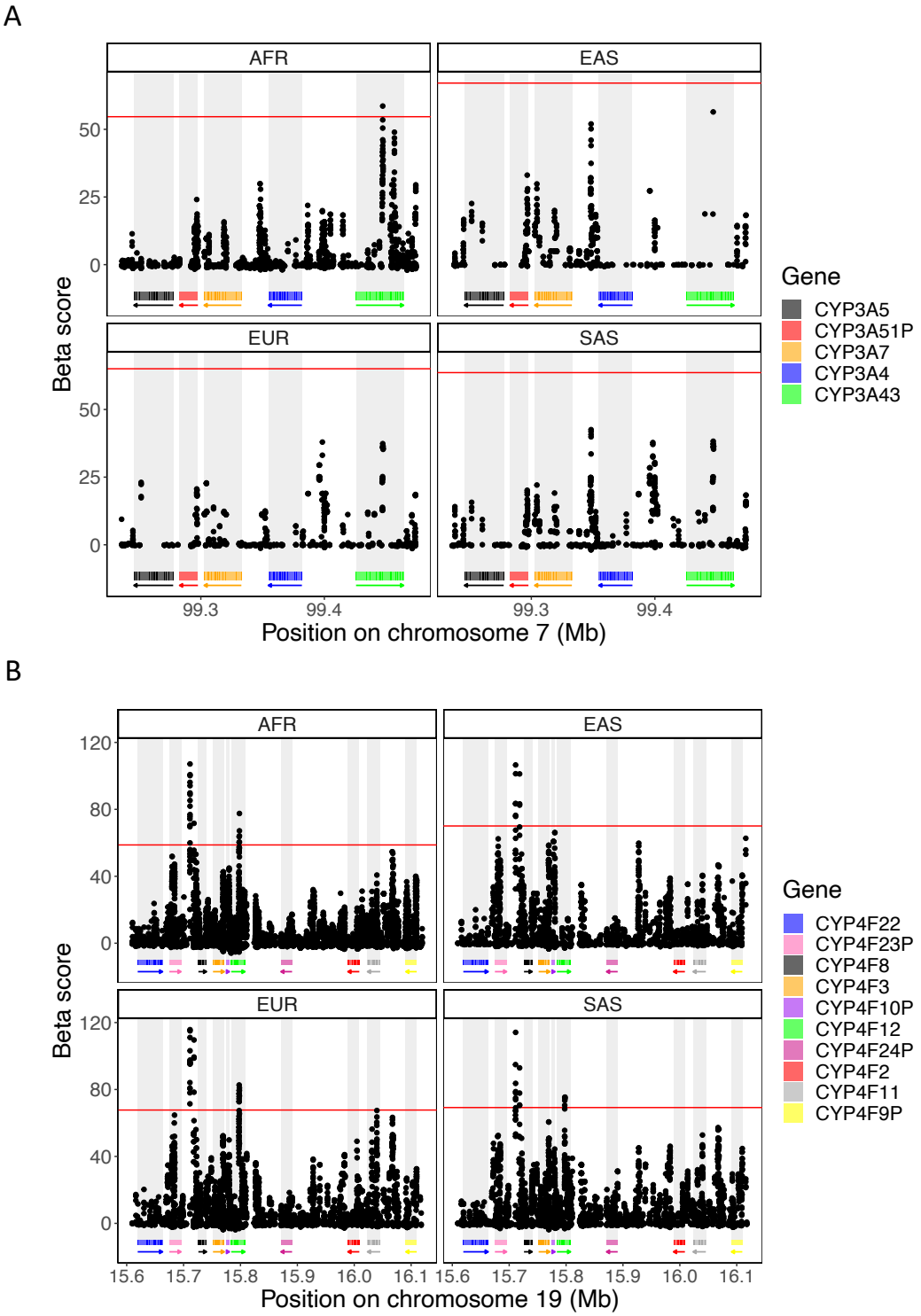


Fig. 2.3. β score in the chromosomal region of the A) CYP3A and B) CYP4F cluster for the 4 super-populations analyzed. The β score was calculated on the 1000G dataset on a per-site basis and the 99th percentile indicating the top 1% β score is displayed by the horizontal line in red. Rectangles below the plot show the position of each gene and arrows indicate on which strand the gene is located.

In both clusters, strong signals of uLD are present (Figure 2.4, Supplementary Figure 2.6) compared to matched gene clusters (Methods), but CYP4F shows 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 double (500 Kb vs 250 Kb, respectively). Significant uLD between CYP3A5 and CYP3A43 and between CYP3A7 and CYP3A43 can be seen in all European populations (Supplementary Figure 2.6A). 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) 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. Toscani 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 (Supplementary Figure 2.7). Among East Asians, uLD signals are seen almost exclusively between SNPs in CYP3A4 and CYP3A43, two genes that are next to each other, with no clear recombination hotspot separating them, meaning that linkage disequilibrium can be expected (Supplementary Figure 2.7). 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 2.4A). 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 2.4A, Supplementary Figure 2.6A).

In the CYP4F cluster, several pairs of SNPs have patterns of LD that deviate significantly from the empirical distribution (Supplementary Figure 2.6B). In almost every populations there is uLD for CYP4F22-CYP4F11 and CYP4F22-CYP4F12, even though these genes are far from each other (0.36 Mb, 0.12 Mb respectively). CYP4F22 and CYP4F2 are also in

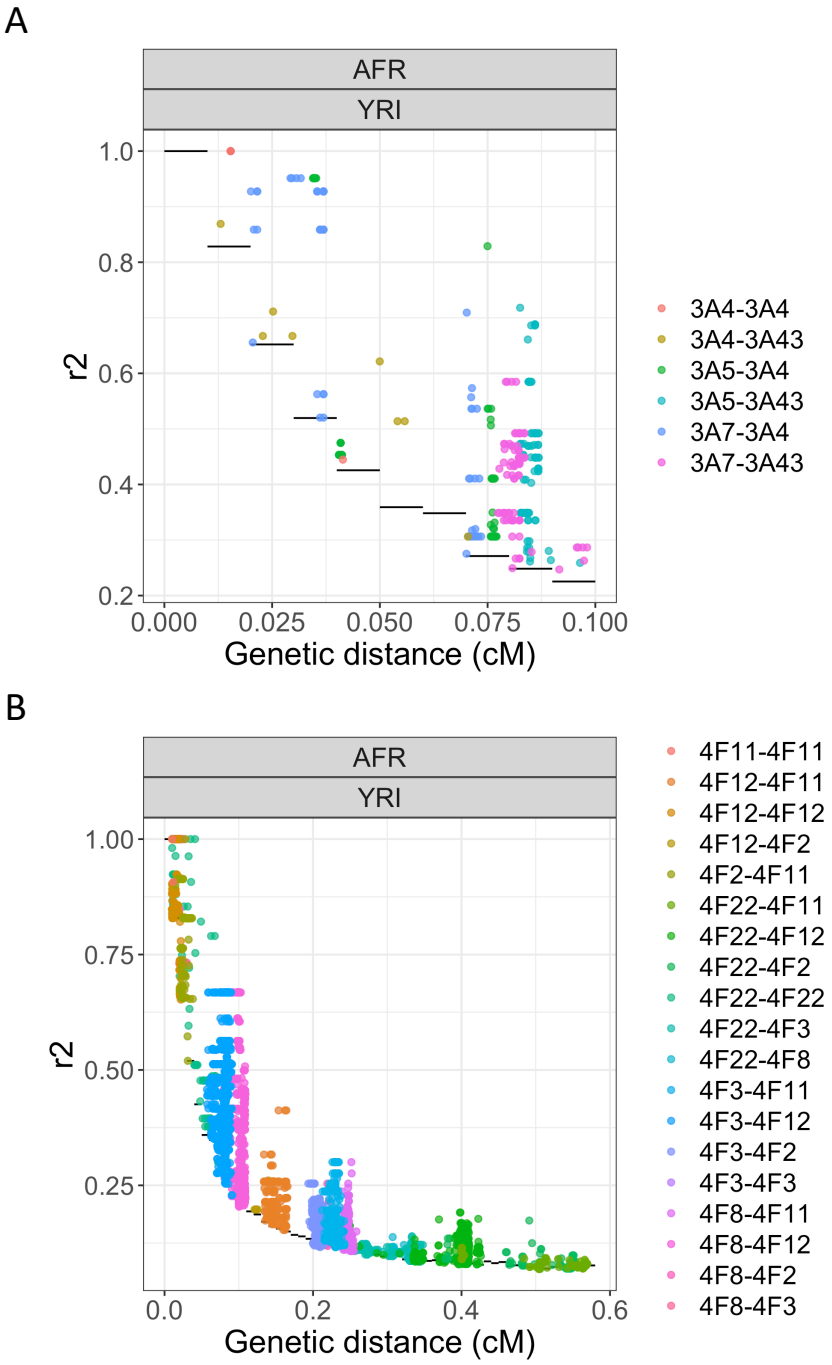


Fig. 2.4. r^2 values between each pairs of SNPs in the A) CYP3A and B) CYP4F cluster in the Yoruba (YRI, AFR) population. The distance between the SNPs is in centimorgan (cM). Only r^2 values over the null distribution are shown. The null distribution is shown with black horizontal lines. Dots are colored according to which genes are involved in the pair.

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 seems to be driven by association involving the CYP4F12 gene (Figure 2.4B). 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 (Supplementary Figure 2.8A) and CYP4F8 genes (Supplementary Figure 2.8B). The extreme signals in this gene cluster are in line with the hypothesis that balancing selection acts via gene-gene interactions, or epistasis (Llaurens et al. 2017). As these patterns could be due to sequencing errors (Akey et al. 2001), we used the latest 1000G dataset which has high-coverage sequencing and is aligned on hg38 (Methods). These results were replicated in this second dataset, greatly reducing the possibility that the observed signal is due to sequencing errors. 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 (Supplementary Figure 2.8A) and two SNPs in CYP4F8 (Supplementary Figure 2.8B). No specific SNPs are in uLD with CYP4F2.

2.3.5. Detection of eQTLs

We 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 expression quantitative trait loci (eQTL). We used GTEx dataset (v8), which contains gene expression of 54 tissues across 948 donors. We applied a linear regression model, correcting for age, sex, the first 5 Principal Components (PCs), time since death, collection center and PEER factors. To report genome-wide significant eQTL signals, we used a P-value threshold for significance at $10^{-}8$ (Methods).

In the CYP3A cluster, three SNPs are under positive selection in the Punjabi population (PJL, SAS): rs487813, rs679320 and rs568859. These SNPs are located in CYP3A43 and are significant eQTLs of CYP3A5 in lung (Figure 2.5A). The SNP under balancing selection in

the Luhya population (LWK, AFR), rs800667, in CYP3A43 is also an eQTL of CYP3A5 in lung (Figure 2.5B). These eQTLs affect CYP3A5 expression in lungs, even though CYP3A5 and CYP3A43 are at opposite ends of the cluster, 147.99 Kb apart. This result is in line with the LD analyses (Figure 2.4A), which suggested uLD between SNPs in CYP3A5 and CYP3A43 in Europeans, Africans and the Japanese. The four SNPs under selection are actually in uLD with exactly the same sixteen SNPs, of which eleven SNPs are in pairs of SNPs in uLD in Toscani (TSI, EUR) and five are in uLD in Americans of African Ancestry (ASW, AFR). The effect size estimate for every significant eQTL identified here is negative ($\beta_{eQTL} < 0$), indicating a reduction in CYP3A5 gene expression with each non-reference allele. No other significant eQTL associations are seen in this cluster whose SNPs show signatures of natural selection.

In the CYP4F cluster, a SNP under positive selection, rs74459786 (Supplementary Table 2.1), located in the intergenic region between CYP4F23P and CYP4F8, is an eQTL of CYP4F12 in adipose tissue (Figure 2.5C), with a negative effect size. SNPs under balancing selection (Supplementary Table 2.2) 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, there was a SNP with a balancing selection signal that is also an eQTL of CYP4F12 expression in adipose tissue, more precisely in adipose-subcutaneous (Figure 2.5D) 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 2.5C.

Another SNP under positive selection in this intergenic region, rs62115147 (Supplementary Table 2.1), is also associated with CYP4F3 expression in one of the brain tissues (Brain-Spinalcord-cervicalc-1) and in nerve tissue (Supplementary Figure 2.9A). 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.

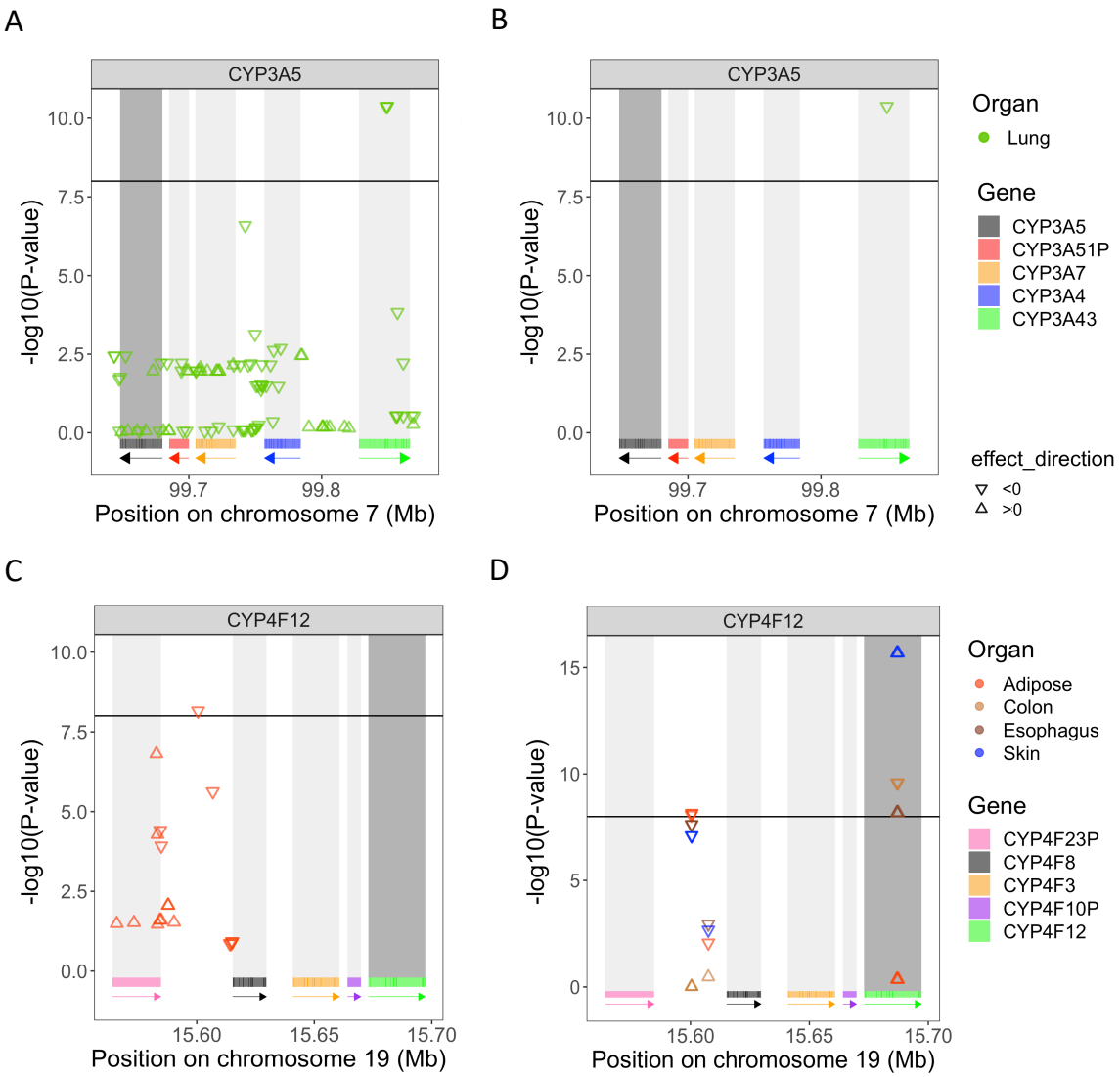


Fig. 2.5. P-values of the associations between SNPs under A) positive selection and B) balancing selection and CYP3A5's gene expression in lungs and p-values associated with SNPs C) under positive selection and D) balancing selection and tissue-specific gene expression of CYP4F12. CYP3A5 and CYP4F12 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.

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 (Supplementary Table 2.1) are eQTLs of CYP4F2 in brain and skin tissues (Supplementary Figure 2.9B) with consistent, negative effect sizes. Additionally, the same SNPs under positive selection within CYP4F11 are associated with expression of CYP4F11 itself in multiple tissues (Supplementary Figure 2.9C). The direction of effect on gene expression is the same for all significant associations.

2.4. Discussion

Drug metabolism is a rather complex system modulated by many CYP450 genes. As shown by others (X. Chen et al. 2009; J. Li et al. 2011; Qiu et al. 2008; Thompson, Kuttab-Boulos, Witonsky, et al. 2004), we 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 (Burk et al. 2004; X. Chen et al. 2009; Thompson, Kuttab-Boulos, Witonsky, et al. 2004; Thompson, Kuttab-Boulos, Yang, et al. 2006) and the less well known CYP4F cluster, identified because of their outlier patterns based on neutrality and population differentiation tests. The two CYP450 clusters have higher levels of natural selection forces (positive selection and balancing selection) acting on them as a whole, as well as population differentiation. The forces of natural selection differ between the two clusters; the CYP3A cluster is evolving under positive selection, while the CYP4F cluster is evolving more under balancing selection. Furthermore, both clusters have candidate sites for co-evolution and co-regulation, but especially in the CYP4F cluster. In the literature, the CYP450 genes are often studied independently. In our study, genes are studied as families to determine if there is evidence of epistasis between the genes within each cluster. As these clusters of genes are involved in drug metabolism (Danielson 2002; Liang et al. 2012; Nebert and Dalton 2006; J. E. Zhang et al. 2017), it is important to understand the impact of genetic variants on their gene expression, as this could help understand how

these variants might impact drug response and disease treatments. These events are known to influence allele frequencies and suggest that, across populations, the impact of specific variants on drug metabolism may differ, which could lead to a deeper understanding of differences in drug response (Guttman et al. 2019; Singh et al. 2011).

The CYP3A cluster, located on chromosome 7, contains 4 genes: CYP3A4, CYP3A5, CYP3A7 and CYP3A43. This cluster is involved in the metabolism of xenobiotic compounds, such as drugs. Signals of positive selection were detected in the CYP3A cluster with neutrality tests (Tajima's D, Fu and Li's D* and F*, and Fay and Wu's H). Specifically, CYP3A4 and CYP3A7 have been or are under recent positive selection in Africans, Europeans and Chinese, whereas CYP3A5 was under positive selection in Europeans and CYP3A43 in non-Africans (X. Chen et al. 2009). The ratio of nonsynonymous and synonymous substitution rates (dn/ds) also detected positive selection in CYP3A4 and CYP3A7, where CYP3A4 was under positive selection in humans and CYP3A7 was under positive selection during the origin of Hominoidea (human, chimpanzee, gorilla, and orangutan) (Qiu et al. 2008). CYP3A5 has a positive selection signal detected in Europeans by high iHS score(Voight et al. 2006) and the haplotype structure and the excess of rare variants in the non-Africans suggest that variants associated with salt homeostasis were targets of selective pressure in non-Africans (Thompson, Kuttab-Boulos, Witonsky, et al. 2004). In Eurasian populations, signals were also detected in CYP3A4 and CYP3A5 (J. Li et al. 2011). Our analyses confirmed that CYP3A genes are evolving under positive selection as previously reported. We found that the locus known to cause low expression, rs776746/CYP3A5*3, is under positive selection ($|iHS| \geq 2$) in two African populations (YRI, GWD) and that another locus, known to cause non-expression of CYP3A5 (Kuehl et al. 2001), rs10264272/CYP3A5*6, is also under positive selection ($|iHS| \geq 2$) in African populations (YRI, GWD, LWK). This means that the derived allele has swept up in frequency in Africans. Rs776746 (low-expresser of CYP3A5) is not an eQTL, but is in uLD in Toscani (TSI, EUR) with the four SNPs under selective pressure in the CYP3A cluster acting as eQTLs of CYP3A5 in lungs. However, the three SNPs under positive selection in the Punjabi (PJL, SAS), rs487813, rs679320 and

rs568859, are not mentioned as having any clinical association or publications nor is the SNP under balancing selection, rs800667, in an African population (LWK). The ancestral alleles of these SNPs have a selective advantage, as their iHS values are above 2. However, the negative effect size of the eQTLs indicates that each non-reference allele is decreasing CYP3A5 gene expression. Given that, positive selection is not acting on the eQTLs themselves. Additionally, the function of CYP3A43 is not well known, unlike other CYP3A genes. Our analysis suggest that SNPs in CYP3A43 regulate CYP3A5 gene expression, at least in lungs. CYP3A43 is the ancestor gene of this cluster (McArthur et al. 2003; Qiu et al. 2008). Therefore, additional analysis could be done to explore if the regulation is the result of the common regulatory elements, if it is the result of epistasis, or both. Furthermore, a study suggested that CYP3A43 might be under ongoing pseudogenization (Qiu et al. 2008). This could support our observation that CYP3A43 is acting as a regulator of gene expression, which is one suggested function of pseudogenes (Pink et al. 2011).

The CYP4F cluster, found on chromosome 19, is made up of 6 genes: CYP4F2, CYP4F3, CYP4F8, CYP4F11, CYP4F12 and CYP4F22. CYP4F enzymes are involved in the metabolism of endogenous compounds: arachidonic acid, leukotriene B4, nutrients (vitamins K1 and E) and drugs (pafuramidine and fingolimod) (Hardwick 2008; Jin et al. 2011; A. Kalsotra et al. 2006; M. Z. Wang et al. 2007). We found both positive and balancing selection acting on the CYP4F cluster. CYP4F12 is evolving under balancing selection and CYP4F22, CYP4F2, CYP4F23P and CYP4F9P are evolving under positive selection. These SNPs under selective pressure are associated with differential gene expression across the cluster in several tissues. The SNP under positive selection associated with expression of CYP4F12 in adipose tissue, rs74459786, is detected to be under positive selection in the Kinh population (KHV) in East Asians. Also, SNPs in CYP4F11 are associated with differential gene expression of CYP4F11 and CYP4F2. However, the selective advantage is on the ancestral allele for SNPs under positive selection in CYP4F11 and CYP4F2, suggesting that the ancestral allele increases expression and confers a selective advantage. A previous study also found high linkage disequilibrium in the CYP4F cluster and that a SNP in the 3' UTR of CYP4F11,

rs1060467, was causing a lower expression of CYP4F11 (J. E. Zhang et al. 2017). Although this SNP, rs1060467, is not in uLD with any of the SNPs identified as eQTL of CYP4F2 (Supplementary Table 2.1), we also found that SNPs in CYP4F11 are associated with the expression of CYP4F2. Both genes are implicated in common metabolic function, such as the synthesis of 20-HETE (20-hydroxyeicosatetraenoic acid) from arachidonic acid (Yi et al. 2017). Thus, this could indicate a possible regulatory mechanism of common functions. Finally, one region in the cluster emerged multiple times during our analysis. The intergenic region between CYP4F23P and CYP4F8 shows strong signals for positive selection and for balancing selection. The SNPs under selective pressure are also clear eQTLs of CYP4F12 in adipose tissue and of CYP4F3 in nerve. In addition, the derived allele of rs62115147 is under positive selection and also is decreasing CYP4F3 expression in nerve. Our work therefore highlights a novel regulatory element that may be important for the metabolism of fatty acid as the fatty acid metabolism is linked to the adipose tissue (Frayn et al. 2006) and that CYP4F12 is implicated in the metabolism of fatty acid (Stark et al. 2005).

In the future, such cluster-wide analyses could be applied to other CYP450 gene clusters, such as the CYP2C and the CYP2D clusters. As both clusters are involved in the drug metabolism (Danielson 2002; Jin et al. 2011; Liang et al. 2012; Nebert and Dalton 2006; M. Z. Wang et al. 2007; J. E. Zhang et al. 2017), which would provide a deeper understanding of how the genetic diversity of theses gene clusters affects drug metabolism. Another avenue of future work would be in determining the phenotypic and disease associations of the regions and SNPs we found, which will improve the understanding their clinical impact.

Finally, our results demonstrate that there is heterogeneity across human populations, both in terms of variants and of interactions between variants and genes for the CYP3A and CYP4F clusters. So, as studies typically focus on European populations, there could be an impact on metabolic functions and drug response in individuals with a different genetic profile. In particular, these variants could cause impaired efficacy, as well as side effects. This therefore underlines the importance of including individuals from several populations in biomedical research in order to capture all of the genetic diversity.

2.5. Methods

2.5.1. 1000 Genomes genetic data

The data analyzed is from the Phase III 1000 Genomes project (1000G) (Genomes Project et al. 2015). The 1000 Genomes Project includes 2,504 individuals from 26 populations. These populations can be split into 5 super-populations: 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) file of 1000G uses the hg19 genome build. VCFtools v0.1.14 (Danecek et al. 2011) was used to filter the 1000G dataset. Indels and non-biallelic alleles were removed (`-remove-indels -max-alleles 2 -min-alleles 2`) and only SNPs located in the 57 CYP450 genes were kept (`-bed`). After filtering, the CYP450 dataset included a total of 61,739 SNPs and 2157 individuals. We refer to this as the "1000G CYP450 dataset". Another 1000 Genomes dataset used is the "1000 Genomes 30x on GRCh38" (Byrska-Bishop et al. 2021). This dataset has new, high-coverage (30x) sequencing data aligned on hg38 and includes the 2157 individuals from the "1000G CYP450 dataset".

2.5.2. Genetic diversity and population differentiation

Both Tajima's D and F_{ST} statistics were obtained with VCFtools 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 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 annotation file (`https://cadd.gs.washington.edu/`) 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). The F_{ST} values, from Weir and Cockerham derivation (Weir et al. 1984), 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. Violin plots were generated using the library `ggplot2` in R (v.3.6.0)(`https://www.r-project.org/`).

2.5.3. Detecting natural selection

The method used to detect balancing selection is the β score (Siewert et al. 2017). This score has already been calculated on the whole 1000 Genomes project data (`http://coruscant.itmat.upenn.edu/data/SiewertEA_Full_BetaScores.tar.gz`). The approach used to detect signal of recent positive selection was iHS (integrated haplotype score) (Voight et al. 2006). The iHS computation was performed by us on the Phase III 1000 Genomes project dataset (`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/`), 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 (Heng Li 2011). The hapbin program v.1.3.0 (Maclean et al. 2015) was then used to compute iHS using per population-specific genetic maps computed by Adam Auton on the 1000 Genomes OMNI dataset (`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20130507_omni_recombination_rates`). 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 phase III dataset. For all natural selection analyses, SNPs annotated to be in a repetitive region were identified using the RepeatMasker track on the UCSC genome browser (Kent et al. 2002) and were removed.

2.5.4. 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 (C. C. Chang et al. 2015) using the population-specific genetic maps. The genetic map from the closest sub-population was selected when the genetic map was not available for a subpopulation based on their global F_{ST} value.

To compute a null distribution to detect unusual linkage disequilibrium (uLD), the Human hg19 Gene transfer format (GTF) file from Ensembl v87 was split for each chromosome. Each chromosome was screened using an in-house python script to find windows matching the CYP4F cluster: windows of 430 Kb containing 6 genes were kept. The 1000G dataset was filtered to exclude 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 we calculated the 99th percentile of r^2 values of each pair of SNPs lying in the bin. This process was done separately for each sub-population of 1000G, 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 filtered to keep only those pairs where one SNP was located in CYP4F12. Using VCFTools, CYP4F genetic data was extracted from the newer, resequenced, hg38 1000 Genomes dataset (Byrska-Bishop et al. 2021) and r^2 values were calculated as described above.

2.5.5. eQTLs analysis of SNPs under selection

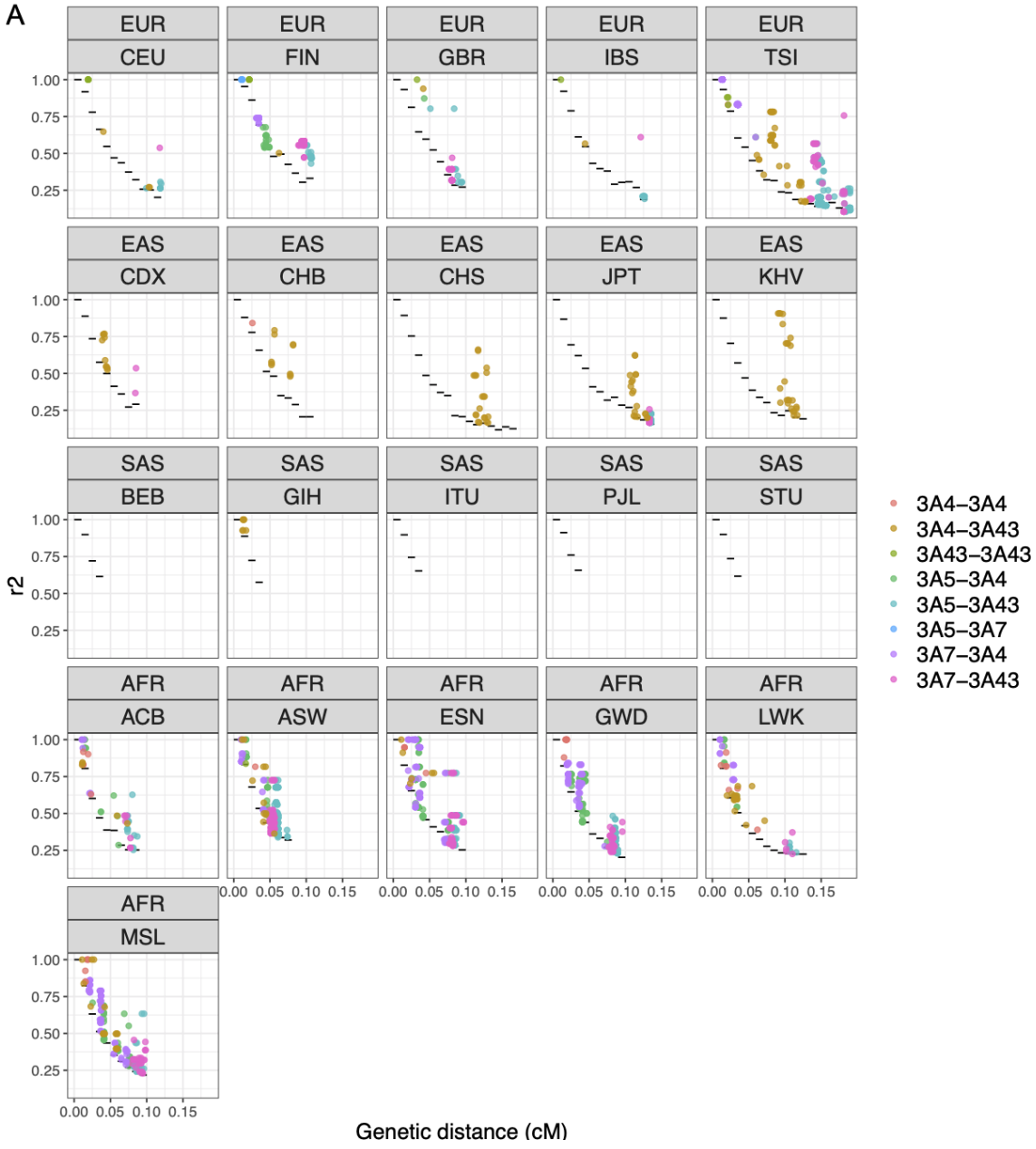
The Genotype-Tissue Expression v8 (GTEx)(Lonsdale et al. 2013) was accessed through dbGaP (phs000424.v8.p2, dbgap project #19088) and contains gene expression across 54

tissues and 838 donors as well as genotyping information, compiled in a VCF file[2] by GTEx on the hg38 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 expression quantitative loci (eQTL) analyses, we selected only SNPs that were identified to be under positive or balancing selection in CYP3A and CYP4F clusters in previous analyses. Since the positions of these SNPs were in the hg19 genome build, we converted these positions to the hg38 genome build to match GTEx v8 data, using the `liftOver` function of the `rtracklayer` R library (Lawrence et al. 2009). P-values of associations between each selected SNP and gene expression of every gene in the cluster were calculated with a linear regression 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). All eQTL plots were generated using the `ggplot2` library in R.

[2]GTEx_Analysis_2017-06-05_v8_WGS_VCF_files_GTEx_Analysis_2017-06-05_v8_WholeGenomeSeq_838Indiv_Analysis_Freeze.vcf.gz

2.6. Supplementary Figures



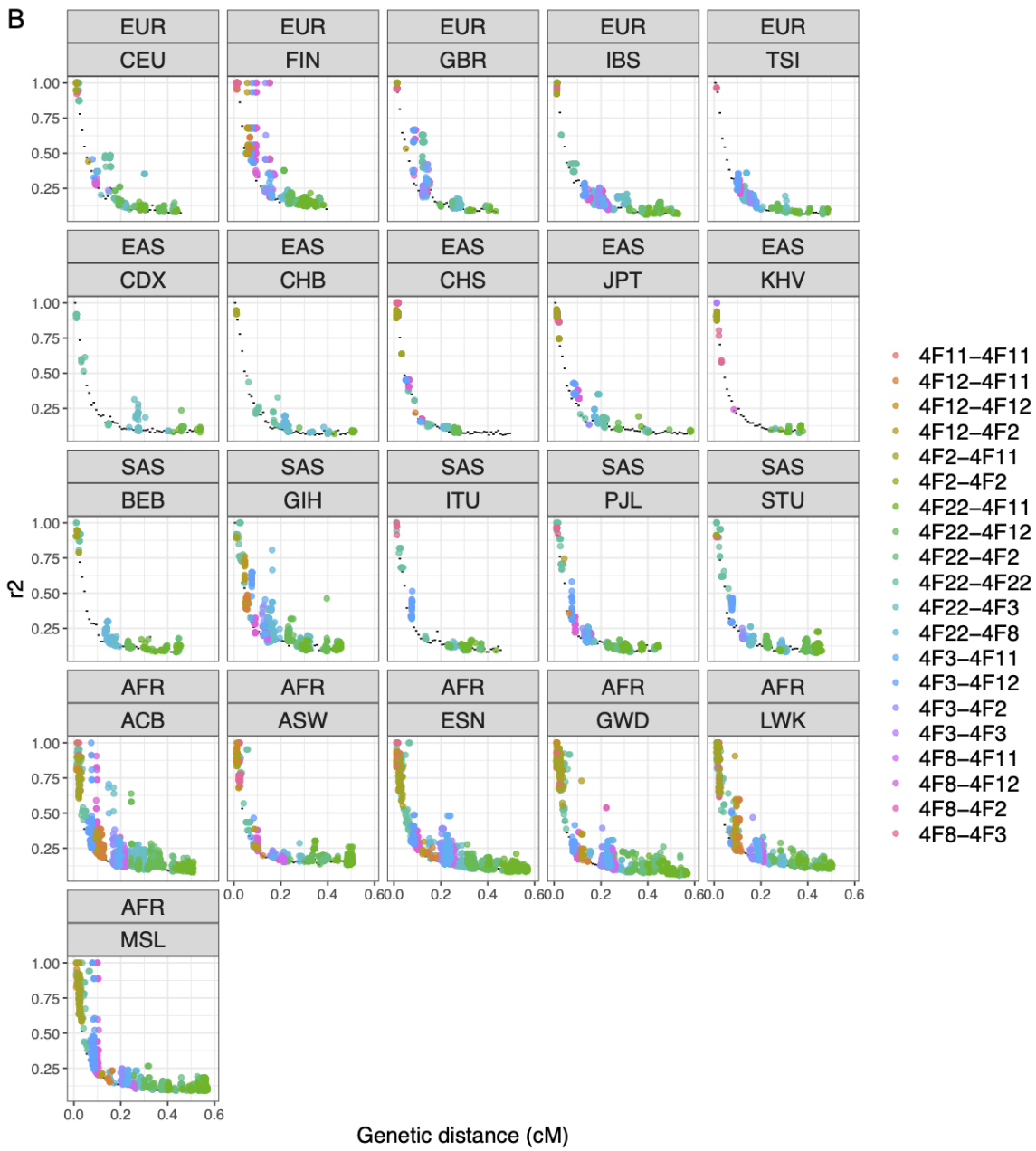


Fig. 2.6. r^2 values between each pairs of SNPs in the A) CYP3A and B) CYP4F cluster for each 1000G populations, except YRI (AFR). The genetic distance between the SNPs is in centimorgan (cM). Only r^2 values over the empirical threshold are shown. The empirical distribution is shown with black horizontal lines. Dots are colored according to which genes are involved in the pair.

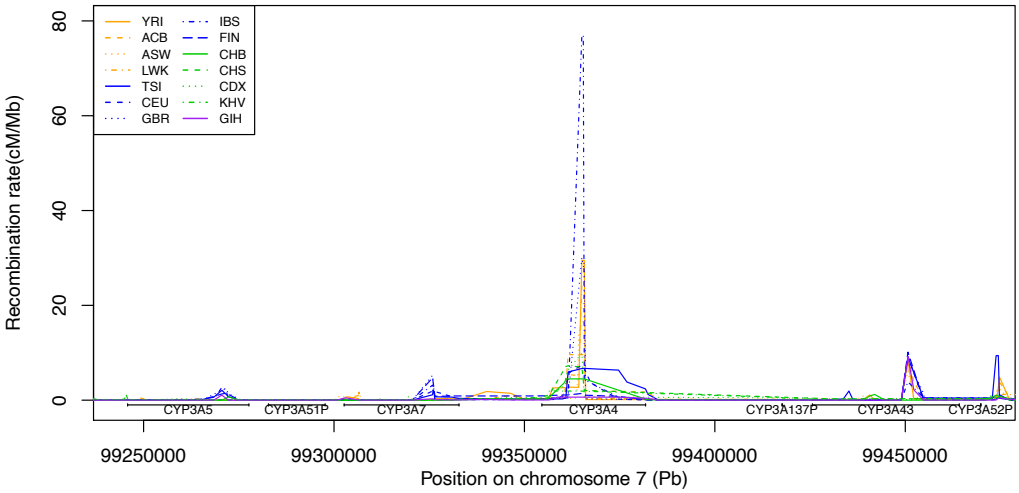
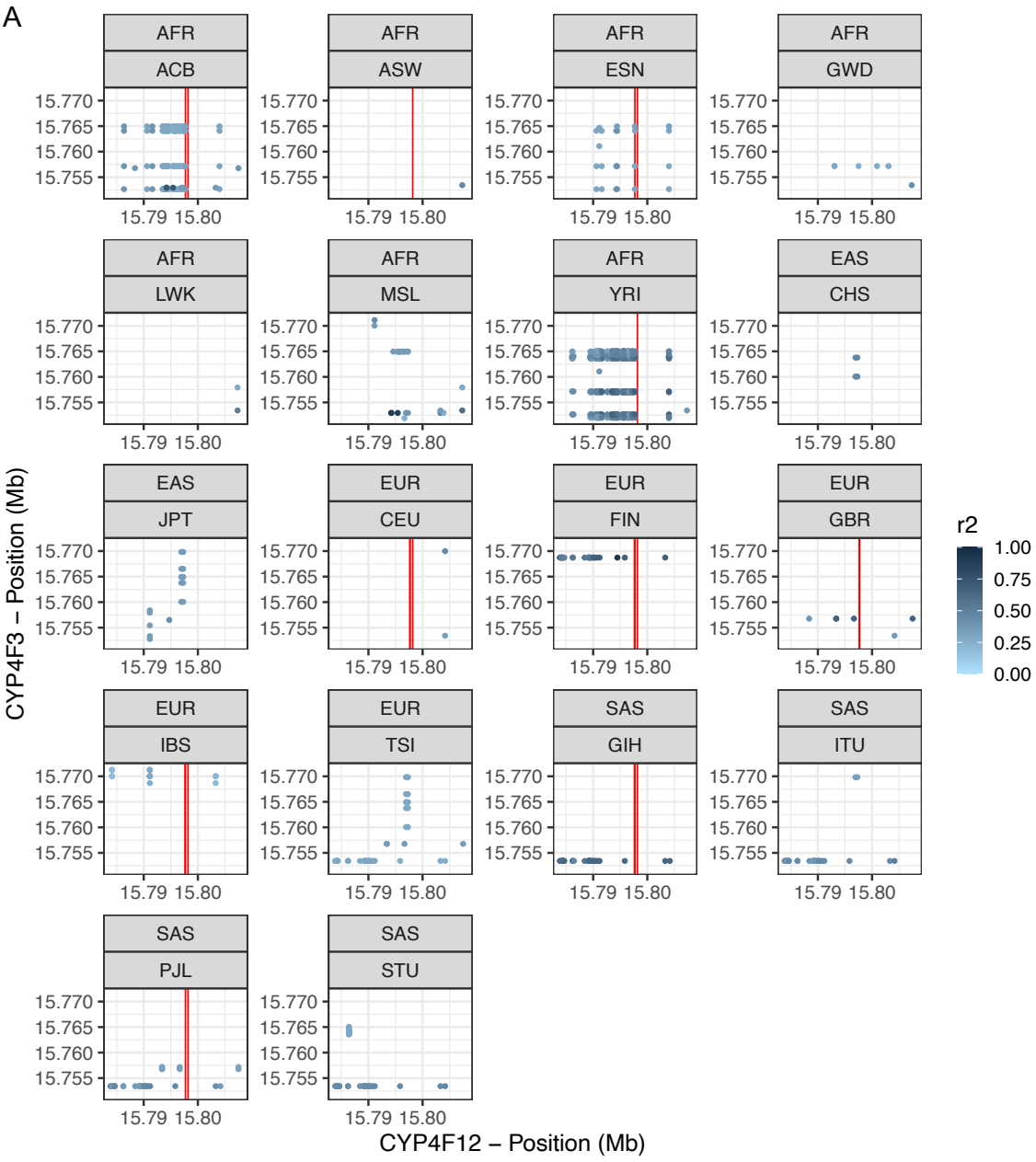


Fig. 2.7. Recombination map in the CYP3A gene cluster. Each line, with a different line pattern, represents a population and is colored according to the super-population. Each gene and pseudogene are shown below the plot with horizontal line.


A
AFR
ACB
AFR
ASW
AFR
ESN
AFR
GWD
AFR
LWK
AFR
MSL
AFR
YRI
EAS
CHS
EAS
JPT
EUR
CEU
EUR
FIN
EUR
GBR
EUR
IBS
EUR
TSI
SAS
GIH
SAS
ITU
SAS
PJL
SAS
STU
15.770
15.765
15.760
15.755
15.79 15.80
CYP4F3 – Position (Mb)
CYP4F12 – Position (Mb)
r2
1.00
0.75
0.50
0.25
0.00

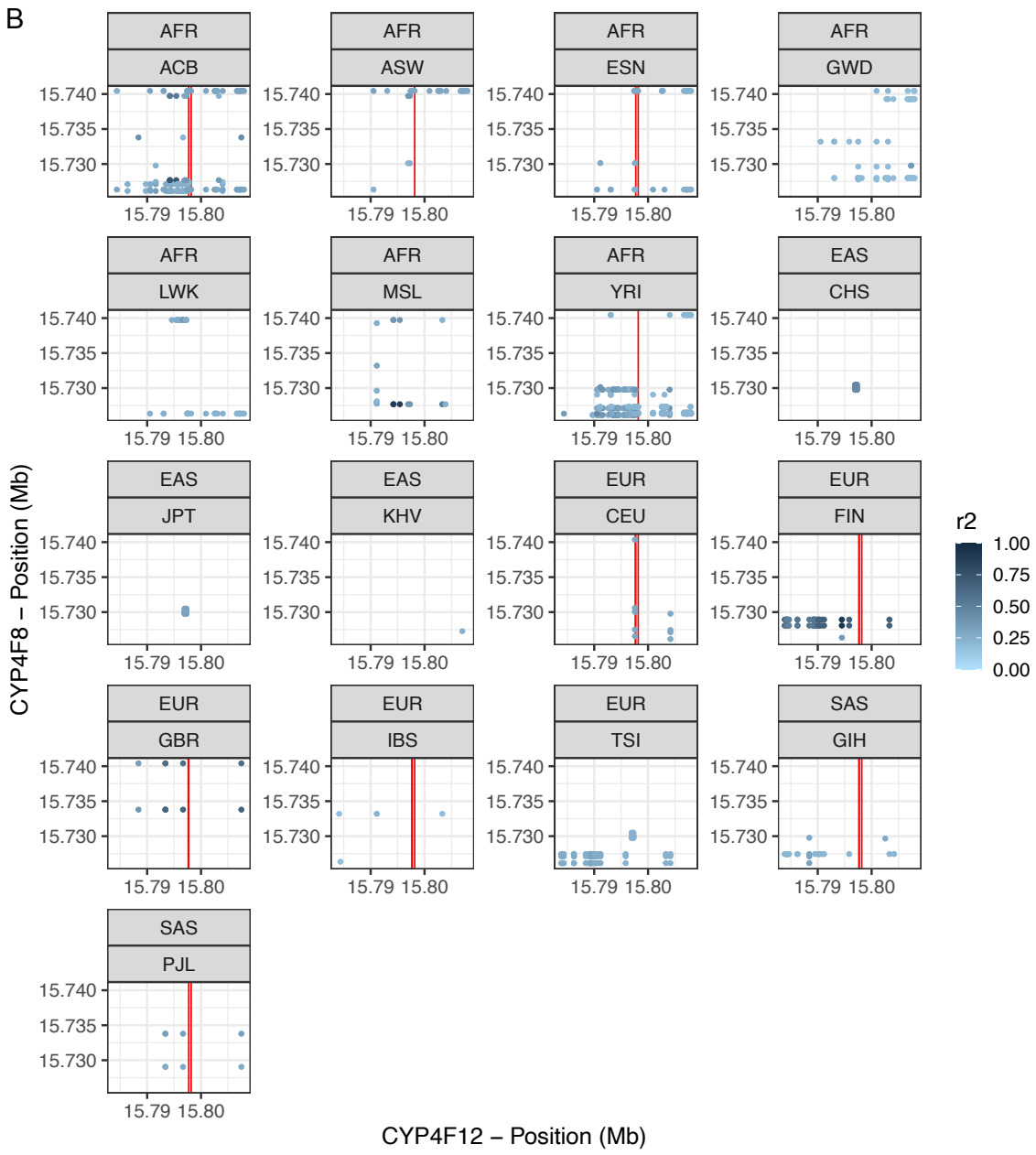


Fig. 2.8. Coordinates of each SNP that is in a pair of SNPs with r^2 values in the extremes of the empirical distribution for each subpopulation of 1000G. The displayed SNPs pairs have one SNP in CYP4F12 and the other is in A) CYP4F3 and in B) CYP4F8. We took r^2 values from the previous analysis and filtered to keep only values where one SNP was located in CYP4F12. The graph is generated using the *ggplot*2 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*.

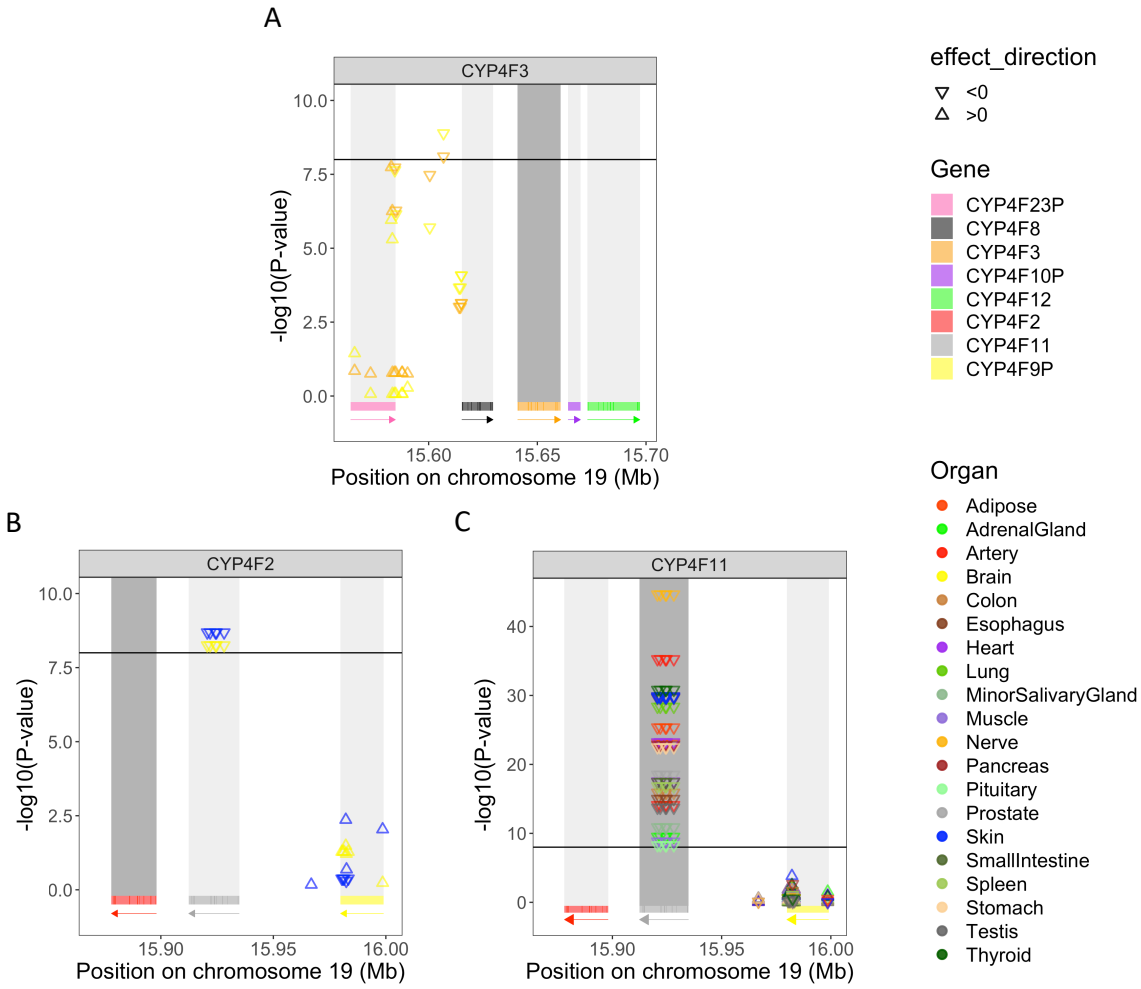


Fig. 2.9. P-values associated with SNPs under positive selection ($|iHS| \geq 2$) explaining variation of gene expression of A) CYP4F3 B) CYP4F2 and C) CYP4F11. The tested gene is shown in dark gray and the effect size is represented either by a triangle standing on its base or a triangle standing on 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	

Tableau 2.1. 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
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
	74.91468	ITU	
	114.13019	GIH	
	78.72150	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

Table 2.2 continued from previous page

Variant identifier	β score	Population	Super-population
	94.87801	GIH	SAS
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	
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	

Table 2.2 continued from previous page

Variant identifier	β score	Population	Super-population
rs16980720	60.32447	ACB	AFR
	63.67802	ESN	
	77.61884	CEU	EUR
	79.43404	FIN	
	75.75824	GBR	
	82.68951	IBS	
	75.28558	GIH	SAS
	73.43462	PJL	
rs644584	72.91370	CEU	EUR
	74.40043	FIN	
	72.64120	GBR	
	77.31849	IBS	
	70.56299	GIH	SAS

Tableau 2.2. 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.

2.7. Supplementary text

2.7.1. Pre-processing of GTEx genetic data

Starting from the imputed genotyping dataset, we kept bi-allelic SNPs and removed positions with more than 5% genotype missingness, leaving 100,986 SNPs which were used to perform a PCA using flashPCA2 (Abraham et al. 2017). To retain the non-admixed individuals of European descent, we reduced the dimensionality of the top 10 PCs using the R package `UMAP` (McInnes et al. 2018) (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 our all subsequent analyses with these 699 individuals. Next we separated each tissue using the file named *GTEx_Analysis_v8_Annotations_SampleAttributesDS.txt*, then removed tissues with fewer than 50 samples (bladder, endocervix, ectocervix, fallopian tube, kidney medulla), leaving samples from 49 different tissues. We kept in our analyses genes which had more than 6 reads in at least 20% of the sample[3]. We then normalized expression data using limma (TMM normalization) (Ritchie et al. 2015) and voom (Law et al. 2014). We calculated PEER factors (Stegle et al. 2012) 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 $N \geq 350$.

[3]GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_reads.gct.gz

Chapitre 3

Synthèse

3.1. Discussion

Le projet reporté dans ce mémoire permet de mieux caractériser les deux sous-familles de gènes CYP3A et CYP4F. Au niveau des forces de pression sélective, les analyses ont confirmé la présence de sélection naturelle dans les gènes CYP3A, ce qui a auparavant été observée dans plusieurs études (X. Chen et al. 2009; J. Li et al. 2011; Qiu et al. 2008; Thompson, Kuttab-Boulos, Witonsky, et al. 2004). Pour CYP3A5, contrairement aux signaux qui ont été reportés chez les Européens (J. Li et al. 2011; Thompson, Kuttab-Boulos, Witonsky, et al. 2004), nos résultats reportent plutôt des signaux de sélection positive en Afrique. En particulier, nous avons trouvé que deux locus connus pour causer une faible expression (rs776746/CYP3A5*3), ou une non-expression (rs10264272/CYP3A5*6) (Kuehl et al. 2001) sont sous sélection positive en Afrique. Pour CYP3A4, nos résultats confirment la présence de sélection positive chez les Africains et les Européens (X. Chen et al. 2009; J. Li et al. 2011), mais ne réplique pas les signaux chez les Asiatiques alors que pour CYP3A7, nos analyses confirment la sélection positive chez les Asiatiques (X. Chen et al. 2009), mais pas pour les Européens et les Africains. Pour le gène CYP3A43, les forces sélectives détectées se trouvaient chez les non-Africains (X. Chen et al. 2009). Or, nos analyses ont plutôt détecté des signaux dans les quatre super-populations, particulièrement en Afrique et en Asie de l'Est. Pour la sous-famille CYP4F, les pressions de sélection n'étaient pas connues. Nos analyses ont démontré qu'il y avait présence de sélection positive et de sélection balancée. En

effet, nous avons détecté des signaux de sélection positive pour CYP4F22 en Afrique, Europe et Asie de l'Est et pour CYP4F11 en Europe et en Asie de l'Est. La région intergénique entre CYP4F23P et CYP4F8 est sous sélection positive pour les quatre super-populations alors que le pseudogène CYP4F9P semble être sous sélection positive en Afrique. Cela indique que cette région a une fonction potentiellement bénéfique, bien que cette fonction ne soit pas établie. De plus, la présence d'eQTLs dans cette région pourrait indiquer qu'elle serait impliquée dans la régulation de l'expression génique. D'ailleurs, cette même région intergénique entre CYP4F23P et CYP4F8 a également un signal significatif de sélection balancée dans les quatre super-populations. Enfin, un signal de sélection balancée a été observé pour CYP4F12 dans les populations Africaines, Européennes et Asiatiques du Sud. En somme, les pressions de sélection diffèrent entre les deux sous-familles, suggérant des histoires évolutives distinctes.

Ensuite, l'analyse du déséquilibre de liaison a identifié un grand nombre de variants ayant un déséquilibre de liaison anormal, que nous avons également nommé *unusual linkage disequilibrium* (uLD) dans l'article présenté au chapitre 2. Dans la sous-famille CYP3A, de nombreuses paires de variants sont en uLD en Afrique, où nous observons le plus de uLD. CYP3A4 est en uLD avec tous les autres gènes de la sous-famille. En Asie de l'Est, les paires en uLD se retrouvent le plus souvent entre CYP3A4 et CYP3A43 alors qu'en Asie du Sud, très peu de uLD est observé. En Europe, la Finlande et la Toscane sont les deux populations avec le plus de paires de variants en uLD. Tout comme la sous-famille CYP3A, les populations Africaines ont le plus de paires de variants en uLD chez les gènes CYP4F. Une population Africaine en particulier, la population Yoruba d'Ibadan du Nigeria (YRI), démontre un fort déséquilibre de liaison entre CYP4F12 et les autres gènes de la sous-famille. Notamment, nos analyses ont démontrés que des variants se situant dans CYP4F12 dans une région spécifique (chr19:15.79-18.00 Mb,GRCh38) était en uLD avec des quelques SNPs de CYP4F8 et CYP4F3. En Finlande, ce patron est également détecté. Cependant, la région de CYP4F12 en uLD avec des variants de CYP4F8 et CYP4F3 est différente. Puis, CYP4F22 est également en uLD dans la majorité des populations avec CYP4F11 et CYP4F12. Comme

la distance physique entre ces gènes est grande, ce résultat n'était pas attendu. De plus, la sous-famille CYP4F comporte un plus grand pourcentage de paires de variants ayant un déséquilibre de liaison anormal comparativement à la sous-famille CYP3A (8.1% vs 4.7% de paires de SNPs en uLD). Toutefois, comme les gènes CYP4F sont sous sélection balancée, il est attendu que les variants de ces gènes soient en déséquilibre de liaison. En effet, la sélection balancée a pour effet de maintenir des allèles à fréquence intermédiaire lorsqu'il y a un avantage de l'hétérozygote, ce qui peut augmenter le déséquilibre de liaison (Slatkin 2008). De surcroît, l'analyse de déséquilibre de liaison permet d'identifier des variants co-évoluant ensemble. Par conséquent, nos résultats suggèrent donc une co-évolution entre plusieurs gènes de cette sous-famille ainsi qu'une co-évotion entre CYP3A5 et CYP3A43.

Par la suite, la présence d'*eQTLs* a été détectée dans les 2 sous-familles à partir des variants ayant été identifiés comme étant sous pression sélective. Pour la sous-famille CYP3A, peu d'*eQTLs* sont présents. Les seuls variants étant des *eQTLs* se situent dans CYP3A43 et cause une réduction de l'expression de CYP3A5 dans les poumons. Les gènes CYP3A5 et CYP3A43 étant à l'opposé physiquement l'un de l'autre dans la sous-famille, ce résultat n'était pas attendu. En effet, le déséquilibre de liaison tend à diminuer plus la distance augmente. Or, lors de l'analyse de déséquilibre de liaison, des paires de variants avec un LD anormal entre ces deux gènes ont été détectées. Il semblerait donc avoir une co-évolution et une interaction entre CYP3A5 et CYP3A43 qui passerait par l'expression génique. Pour la sous-famille CYP4F, plusieurs *eQTLs* sont détectés, particulièrement au niveau de l'expression de CYP4F12 dans le colon, l'oesophage, les tissus adipeux et la peau. L'expression de CYP4F12 dans le colon avait d'ailleurs été rapportée (J. Bylund et al. 2001). En plus d'être ressorti dans ces analyses, CYP4F12 est ressorti durant les analyses de déséquilibre de liaison et de sélection balancée. Les variants sous sélection balancée dans CYP4F12 se retrouvent également en uLD avec CYP4F22 et CYP4F8, malgré la présence de nombreux *hotspots* de recombinaison entre CYP4F12 et les deux autres gènes (Figure 3.1). Ces *hotspots* sont présents dans l'ensemble des populations, à différentes intensités, et devraient faire en sorte que le LD entre ces variants devrait être plus faible. Cependant, ce n'est pas ce que

nous observons, nous indiquant une interaction possibles entre ces variants. Par ailleurs, les variants sous sélection positive dans CYP4F11 régulent l'expression génique de CYP4F11 dans plusieurs tissus, mais régulent également l'expression de CYP4F2 dans le cerveau et la peau. Un variant se situant dans CYP4F11, rs1060467, a auparavant été associé avec une baisse d'expression de CYP4F2 (J. E. Zhang et al. 2017). Ce variant n'étant ni sous sélection positive ou balancée, il n'est pas inclus dans nos analyses d'eQTLs et il n'est pas en uLD avec aucun de nos variants identifiés comme étant des eQTLs de CYP4F2. Nous avons cependant observé une interaction similaire où des variants de CYP4F11 sont associés à une baisse d'expression de CYP4F2. Ensuite, le variant rs74459786, se trouvant dans la région intergénique entre CYP4F23P et CYP4F8, est sous sélection positive dans les populations japonaise (JPT, EAS) et sri-lankaise (STU, SAS). Rs74459786 est également sous sélection balancée. En effet, ce variant est situé dans la région avec le haut score β et ce, à travers les quatre super-populations. Néanmoins, le variant n'est pas sous sélection balancée dans la population japonaise et sri-lankaise. Ce variant est également un *eQTL* de CYP4F12 dans les tissus adipeux, ce qui suggère une interaction variant-gène conférant un avantage sélectif. De surcroît, un seul allèle dérivé étant sous sélection positive est associé significativement à une variation de l'expression génique. Ce variant, rs62115147, est sous sélection positive dans la population Ibérienne d'Espagne (IBS) cause une diminution de l'expression de CYP4F3 dans les nerfs et le cerveau. D'ailleurs, la majorité de ces *eQTLs* cause une baisse d'expression plutôt qu'une augmentation de l'expression. Cette différence dans l'expression peut être causé par une modification du motif de liaison, affectant ainsi l'adhésion de facteur de transcription. La recherche de motifs autour des eQTLs permettrait de valider cette hypothèse. L'analyse des *eQTLs* démontre donc une interaction entre CYP3A5-CYP3A43, CYP4F11-CYP4F12 ainsi qu'entre la région intergénique entre CYP4F23P et CYP4F8 et l'expression de CYP4F12 et CYP4F3. Certes, bien que les *eQTLs* identifient les tissus dont l'expression génique varie, nous ne pouvons interpréter fonctionnellement ces résultats avec précision. Davantage d'analyses seraient requises pour valider les hypothèses générées par notre étude. Notamment, une analyse PheWAS (*Phenome-Wide Association Study*)

permettrait une meilleure interprétation des résultats, ainsi qu'une analyse de l'expression allèle-spécifique (ASE).

Dans le même ordre d'idées, il serait intéressant d'analyser la présence d'eQTLs pour les variants rares de nos gènes. Dans un contexte pharmacogénomique, les variants rares sont sous-analysés dans les grandes études comparativement aux variants communs. Or, avec la venue des nouvelles méthodes de séquençage, il est désormais possible de détecter les variants rares. Il a notamment été démontré que les variants rares causeraient des altérations fonctionnelles et qu'ils peuvent être responsable de la variabilité observée. D'ailleurs, la variabilité inter-individuelle qui n'est pas expliquée pour le moment pourrait être due à des variants rares. Par exemple, il a été démontré que des variants rares dans CYP2C9 influencerait la réponse à la warfarine et qu'inclure ces variants dans les algorithmes de dosage pourrait aider à améliorer leur performance (Magnus Ingelman-Sundberg et al. 2018). De plus, dans certaines populations, comme la population africaine, les variants rares sont présents en grand nombre. Il serait donc pertinent de les étudier afin de comprendre l'impact de ces variants dans cette population (Drögemöller et al. 2014).

Comme décrit précédemment, un fort déséquilibre de liaison est présent dans les deux sous-familles de gènes. Or, ce fort déséquilibre de liaison peut compliquer l'identification des variants causaux. Notamment, comme les variants sont en déséquilibre de liaison, l'effet observé pourrait être causé par un SNP en déséquilibre de liaison avec le SNP identifié comme «d'intérêt». Il y a donc une limite à l'interprétation des variants identifiés.

Ensuite, comme énoncé dans la section 1.1.4.2, le gène CYP4F3 a deux isoformes connus au niveau de l'expression génique: CYP4F3A et CYP4F3B. CYP4F3A est exprimé au niveau des neutrophiles alors que CYP4F3B est exprimé au niveau du foie et des reins (Christmas, Jones, et al. 2001; Corcos et al. 2012). Or, selon Gencode (Frankish et al. 2019), il existerait plus que deux isoformes. Nous n'avons cependant pas effectué l'analyse sur les isoformes. De surcroît, l'analyse des variants reliés à l'épissage alternatif, (*Splicing Quantitative Trait Loci*, sQTLs) serait également importante à faire, particulièrement pour CYP3A43, puisqu'une étude suggère qu'il y aurait soit un épissage alternatif ou par un épissage aberrant pour

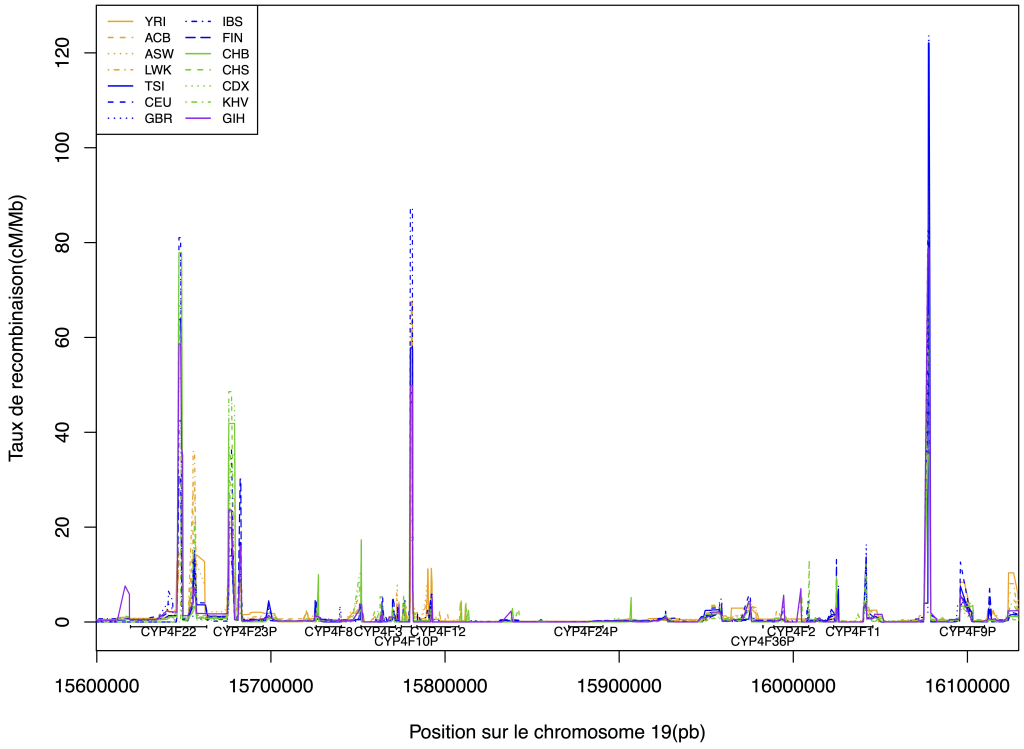


Fig. 3.1. Taux de recombinaison des gènes CYP4F pour les super-populations du projet des 1000 Génomes - Chaque ligne, avec un motif de ligne différent, représente une population et est colorée en fonction de la super-population. Les gènes et pseudogènes sont représentés sous le graphique avec une ligne horizontale.

ce gène (Gellner et al. 2001). Afin de vérifier s'il y a de l'épissage aberrant, il existe des logiciels tels que LeafCutter (Jenkinson et al. 2020), FRASER (Mertes et al. 2021) et MAJIQ (Vaquero-Garcia et al. 2016)).

De plus, un concept n'ayant pas été étudié dans notre projet est la pléiotropie. Un gène (ou variant) est pléiotropique s'il influence plusieurs traits ou qu'il induit plusieurs phénotypes (Cundy et al. 2017). Nos résultats de l'analyse d'*eQTLs* suggèrent la présence de pléiotropie puisque l'expression diffèrent dans plusieurs tissus différents. Notamment, comme les gènes CYP3A et CYP4F peuvent avoir plusieurs fonctions et qu'ils sont impliqués dans plusieurs processus biologiques, ceci suggère qu'ils pourraient y avoir de la pléiotropie. Ils pourraient également être régulés par d'autres gènes en *trans*. Or, nous ne pouvons pas

confirmer la présence de pléiotropie dans nos gènes avec seulement les analyses que nous avons effectuées. Toutefois, une analyse PheWAS (*Phenome-Wide Association Study*) pourrait permettre d'étudier l'association entre les SNPs étant identifiés comme des *eQTLs* sous sélection et les phénotypes qu'ils confèrent et d'ainsi, déterminer la présence de pléiotropie. D'ailleurs, cette analyse permettrait également une meilleure compréhension de l'impact des *eQTLs* sur les processus biologiques.

3.2. Perspective

Dans ce projet, nous avons appliqué des méthodes en génétique des populations afin d'identifier des loci sous pressions sélectives et analysé ces loci afin de déterminer s'ils étaient des *eQTLs*. Dans le futur, ces méthodes pourraient être appliquées sur d'autres jeux de données publiques en génétique des populations, soit des données génomiques de la population québécoise en utilisant les données de la biobanque de l'ICM (Low-Kam et al. 2016) et de CARTaGENE (Awadalla et al. 2013). Ceci pourrait permettre une meilleure caractérisation des variants présents dans la population québécoise. Également, comme il existe plusieurs familles et sous-familles de gènes CYP450, ces analyses pourraient être appliquées sur d'autres sous-familles telles CYP2C ou CYP2D. Ces deux sous-familles sont également impliquées dans le métabolisme des médicaments et des variants connus affectent la réponse aux médicaments.

Un autre aspect des CYP450 est la grande quantité de pseudogènes. En effet, chez l'humain, il y a 57 gènes CYP450 et 58 pseudogènes, et donc les pseudogènes représentent une grande proportion de cette famille. Une sous-famille en particulier, la sous-famille CYP4F, contient un grand nombres de pseudogènes et ils se situent sur plusieurs chromosomes (chr 2, 8, 9, 13, 18, 19 et 21. Ceci nous mène à supposer que lorsqu'ils ne sont pas en *cluster*, qu'ils ne sont pas fonctionnel. Notre hypothèse repose sur le fait que certains pseudogènes se retrouvent isolés sur différents chromosomes, plutôt qu'en *cluster*. Pour tester cette hypothèse, une première étape serait d'analyser la configuration en 3D afin de vérifier s'ils sont en proximité physique et une deuxième étape seraient de vérifier si ces pseudogènes

sont exprimés, ce qui serait un premier pas permettant de déterminer leur fonctionnalité. Également, dans le même ordre d'idées, une analyse de transcriptomique pourrait permette d'identifier si les pseudogènes ont des *eQTLs* et si ces *eQTLs* sont des éléments régulateurs des gènes environnants. Dans le cas où les lectures de *RNA-Seq* s'alignent parfaitement aux pseudogènes et aux gènes, il sera cependant difficile de déterminer si les lectures appartiennent au gène ou au pseudogène. Comme ils s'alignent à plusieurs endroits, leurs scores devraient en être affectés et donc, diminués. De plus, si la similarité de séquence entre les pseudogènes et les gènes est très élevée, il se pourrait que les variants identifiés soient des erreurs d'alignement puisque la lecture s'aligne possiblement au mauvais endroit. En bref, les pseudogènes CYP450 pourraient donc être étudiés afin d'aider à déterminer leur fonction dans les différentes familles de gènes.

Pour conclure, nos résultats démontrent qu'il y a une hétérogénéité à travers les populations humaines, autant au niveau des variants que de l'interaction entre les variants et les gènes pour les sous-familles CYP3A et CYP4F. Ainsi, comme les études portent généralement sur les Européens, il pourrait y avoir un impact sur les fonctions métaboliques et sur la réponse aux médicaments chez les individus ayant un profil génétique différent. Notamment, ces variants pourraient causer une efficacité altérée, de même que des effets secondaires. Ceci souligne donc l'importance d'inclure des individus provenant de plusieurs populations en recherche biomédicale afin de capturer l'ensemble de la diversité génétique.

Références bibliographiques

Abraham, Gad et al. (Sept. 2017). "FlashPCA2: principal component analysis of Biobank-scale genotype datasets". In: *Bioinformatics* 33.17, pp. 2776–2778. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btx299.

Akey, Joshua M. et al. (June 2001). "The Effect That Genotyping Errors Have on the Robustness of Common Linkage-Disequilibrium Measures". en. In: *The American Journal of Human Genetics* 68.6, pp. 1447–1456. ISSN: 0002-9297. DOI: 10.1086/320607.

Andrés, Aida M. (2011). "Balancing Selection in the Human Genome". In: *Encyclopedia of Life Sciences*. DOI: 10.1002/9780470015902.a0022863.

Awadalla, P. et al. (Oct. 2013). "Cohort profile of the CARTaGENE study: Quebec's population-based biobank for public health and personalized genomics". In: *Int J Epidemiol* 42.5. Edition: 2012/10/17, pp. 1285–99. ISSN: 1464-3685 (Electronic) 0300-5771 (Linking). DOI: 10.1093/ije/dys160.

Bailey, David G et al. (Aug. 1998). "Grapefruit juice–drug interactions". In: *British Journal of Clinical Pharmacology* 46.2, pp. 101–110. ISSN: 0306-5251. DOI: 10.1046/j.1365-2125.1998.00764.x.

Bains, Ripudaman K. et al. (2013). "Molecular diversity and population structure at the Cytochrome P450 3A5 gene in Africa". eng. In: *BMC genetics* 14, pp. 34–34. ISSN: 1471-2156. DOI: 10.1186/1471-2156-14-34.

Barnes, M.R. (2007). *Bioinformatics for Geneticists: A Bioinformatics Primer for the Analysis of Genetic Data*. Wiley. ISBN: 978-0-470-02619-9.

Barreiro, Luis (2017). *BCM2004- Évolution moléculaire: La sélection naturelle*. présentation PowerPoint. Université de Montréal.

Bigos, K. L. et al. (June 2011). "Genetic variation in CYP3A43 explains racial difference in olanzapine clearance". eng. In: *Molecular Psychiatry* 16.6, pp. 620–625. ISSN: 1476-5578. DOI: 10.1038/mp.2011.38.

Bland, J. M. et al. (Jan. 1995). "Multiple significance tests: the Bonferroni method." In: *BMJ : British Medical Journal* 310.6973, p. 170. ISSN: 0959-8138.

Bolger, A. M. et al. (2014). *Trimmomatic: A flexible read trimming tool for Illumina NGS data*.

Brown, Sherry-Ann et al. (Jan. 2018). "Pharmacogenomic Impact of CYP2C19 Variation on Clopidogrel Therapy in Precision Cardiovascular Medicine". en. In: *Journal of Personalized Medicine* 8.1, p. 8. ISSN: 2075-4426. DOI: 10.3390/jpm8010008.

Burk, Oliver et al. (Jan. 2004). "Cytochrome P450 3A and their regulation". In: *Naunyn-Schmiedeberg's Archives of Pharmacology* 369.1, pp. 105–124. ISSN: 1432-1912. DOI: 10.1007/s00210-003-0815-3.

Bylund, J. et al. (2001). "cDna cloning and expression of CYP4F12, a novel human cytochrome P450." In: *Biochemical and biophysical research communications*. DOI: 10.1006/BBRC.2000.4191.

Bylund, Johan, Niklas Finnström, et al. (July 1999). "Gene Expression of a Novel Cytochrome P450 of the CYP4F Subfamily in Human Seminal Vesicles". en. In: *Biochemical and Biophysical Research Communications* 261.1, pp. 169–174. ISSN: 0006-291X. DOI: 10.1006/bbrc.1999.1011.

Bylund, Johan, Mats Hidestrand, et al. (July 2000). "Identification of CYP4F8 in Human Seminal Vesicles as a Prominent 19-Hydroxylase of Prostaglandin Endoperoxides*". en. In: *Journal of Biological Chemistry* 275.29, pp. 21844–21849. ISSN: 0021-9258. DOI: 10.1074/jbc.M001712200.

Byrska-Bishop, Marta et al. (Feb. 2021). "High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios". en. In: *bioRxiv*. Publisher:

Cold Spring Harbor Laboratory Section: New Results, p. 2021.02.06.430068. DOI: 10.1101/2021.02.06.430068.

Cadzow, Murray et al. (Aug. 2014). "A bioinformatics workflow for detecting signatures of selection in genomic data". English. In: *Frontiers in Genetics* 5.293. ISSN: 1664-8021. DOI: 10.3389/fgene.2014.00293.

Cardon, Lon R et al. (Feb. 2003). "Population stratification and spurious allelic association". en. In: *The Lancet* 361.9357, pp. 598–604. ISSN: 0140-6736. DOI: 10.1016/S0140-6736(03)12520-2.

Carithers, Latarsha J. et al. (Oct. 2015). "A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project". In: *Biopreservation and Biobanking* 13.5. Publisher: Mary Ann Liebert, Inc., publishers, pp. 311–319. ISSN: 1947-5535. DOI: 10.1089/bio.2015.0032.

Carlson, Christopher S. et al. (Nov. 2005). "Genomic regions exhibiting positive selection identified from dense genotype data". In: *Genome Research* 15.11, pp. 1553–1565. ISSN: 1088-9051. DOI: 10.1101/gr.4326505.

Chang, Christopher C et al. (Dec. 2015). "Second-generation PLINK: rising to the challenge of larger and richer datasets". In: *GigaScience* 4.s13742-015-0047-8. ISSN: 2047-217X. DOI: 10.1186/s13742-015-0047-8.

Chang, G. W. M. et al. (1999). "The physiological and pharmacological roles of cytochrome P450 isoenzymes". en. In: *Anaesthesia* 54.1, pp. 42–50. ISSN: 1365-2044. DOI: https://doi.org/10.1046/j.1365-2044.1999.00602.x.

Chen, B. et al. (2017). "Departure from Hardy Weinberg Equilibrium and Genotyping Error". In: *Front Genet* 8. Edition: 2017/11/23, p. 167. ISSN: 1664-8021 (Print) 1664-8021 (Linking). DOI: 10.3389/fgene.2017.00167.

Chen, Xiaoping et al. (Oct. 2009). "Molecular Population Genetics of Human CYP3A Locus: Signatures of Positive Selection and Implications for Evolutionary Environmental Medicine". In: *Environmental Health Perspectives* 117.10, pp. 1541–1548. ISSN: 0091-6765. DOI: 10.1289/ehp.0800528.

Choudhuri, S. (2014). *Bioinformatics for Beginners: Genes, Genomes, Molecular Evolution, Databases and Analytical Tools*. Elsevier Science. ISBN: 978-0-12-410510-2.

Christiansen, F.B. (2000). *Population Genetics of Multiple Loci*. Wiley. ISBN: 978-0-471-97979-1.

Christmas, Peter, Jeffrey P. Jones, et al. (Oct. 2001). "Alternative Splicing Determines the Function of CYP4F3 by Switching Substrate Specificity *". English. In: *Journal of Biological Chemistry* 276.41. Publisher: Elsevier, pp. 38166–38172. ISSN: 0021-9258, 1083-351X. DOI: 10.1074/jbc.M104818200.

Christmas, Peter, Sonia R. Ursino, et al. (July 1999). "Expression of the CYP4F3 Gene: TISSUE-SPECIFIC SPLICING AND ALTERNATIVE PROMOTERS GENERATE HIGH AND LOW Km FORMS OF LEUKOTRIENE B4-HYDROXYLASE*". en. In: *Journal of Biological Chemistry* 274.30, pp. 21191–21199. ISSN: 0021-9258. DOI: 10.1074/jbc.274.30.21191.

Churchill, G. A. et al. (Nov. 1994). "Empirical Threshold Values for Quantitative Trait Mapping". In: *Genetics* 138.3, pp. 963–971. ISSN: 0016-6731.

Consortium, The GTEx (Sept. 2020). "The GTEx Consortium atlas of genetic regulatory effects across human tissues". en. In: *Science* 369.6509. Publisher: American Association for the Advancement of Science Section: Research Article, pp. 1318–1330. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aaz1776.

Corcos, Laurent et al. (June 2012). "Human cytochrome P450 4F3: structure, functions, and prospects". en. In: *Drug Metabolism and Personalized Therapy* 27.2. Publisher: De Gruyter Section: Drug Metabolism and Personalized Therapy, pp. 63–71. ISSN: 0792-5077, 2191-0162. DOI: 10.1515/dmdi-2011-0037.

Crawford, M.H. (2007). *Anthropological Genetics: Theory, Methods and Applications*. Cambridge University Press. ISBN: 978-0-521-54697-3.

Cui, Xiaoming et al. (Sept. 2000). "A Novel Human Cytochrome P450 4F Isoform (CYP4F11): cDNA Cloning, Expression, and Genomic Structural Characterization". en. In: *Genomics* 68.2, pp. 161–166. ISSN: 0888-7543. DOI: 10.1006/geno.2000.6276.

Cundy, A.S. et al. (2017). *Découvrir la biologie*. De Boeck supérieur. ISBN: 978-2-8073-0287-7.

Cvijovic, I. et al. (Aug. 2018). "The Effect of Strong Purifying Selection on Genetic Diversity". In: *Genetics* 209.4. Edition: 2018/05/31, pp. 1235–1278. ISSN: 1943-2631 (Electronic) 0016-6731 (Linking). DOI: 10.1534/genetics.118.301058.

Daigneault, Jocelyne et al. (1991). "Genetic epidemiology of cystic fibrosis in Saguenay-Lac-St-Jean (Quebec, Canada)". In: *Clinical Genetics* 40.4, pp. 298–303. ISSN: 0009-9163. DOI: 10.1111/j.1399-0004.1991.tb03099.x.

Danecek, Petr et al. (Aug. 2011). "The variant call format and VCFtools". en. In: *Bioinformatics* 27.15. Publisher: Oxford Academic, pp. 2156–2158. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr330.

Danielson, P. B. (Dec. 2002). "The cytochrome P450 superfamily: biochemistry, evolution and drug metabolism in humans". eng. In: *Current Drug Metabolism* 3.6, pp. 561–597. ISSN: 1389-2002. DOI: 10.2174/1389200023337054.

Darwin, C. (1870). *De l'origine des espèces par sélection naturelle: ou, Des lois de transformation des étres organisés*. Victor Masson et Fils.

Dixon, Anna L. et al. (Oct. 2007). "A genome-wide association study of global gene expression". en. In: *Nature Genetics* 39.10. Number: 10 Publisher: Nature Publishing Group, pp. 1202–1207. ISSN: 1546-1718. DOI: 10.1038/ng2109.

Domanski, Tammy L. et al. (2001). "cDNA Cloning and Initial Characterization of CYP3A43, a Novel Human Cytochrome P450". In: *Molecular Pharmacology* 59.2, pp. 386–392. DOI: 10.1124/mol.59.2.386.

Drögemöller, Britt I et al. (June 2014). "Considerations for rare variants in drug metabolism genes and the clinical implications". In: *Expert Opinion on Drug Metabolism & Toxicology* 10.6. Publisher: Taylor & Francis _eprint: https://doi.org/10.1517/17425255.2014.903239, pp. 873–884. ISSN: 1742-5255. DOI: 10.1517/17425255.2014.903239.

Eichelbaum, M. et al. (2006). "Pharmacogenomics and individualized drug therapy". In: *Annu Rev Med* 57. Edition: 2006/01/18, pp. 119–37. ISSN: 0066-4219 (Print) 0066-4219 (Linking). DOI: 10.1146/annurev.med.56.082103.104724.

Elens, Laure et al. (Dec. 2012). "CYP3A4*22: promising newly identified CYP3A4 variant allele for personalizing pharmacotherapy". In: *Pharmacogenomics* 14.1. Publisher: Future Medicine, pp. 47–62. ISSN: 1462-2416. DOI: 10.2217/pgs.12.187.

Foulkes, A.S. (2009). *Applied Statistical Genetics with R: For Population-based Association Studies*. Springer New York. ISBN: 978-0-387-89554-3.

Frankish, Adam et al. (Jan. 2019). "GENCODE reference annotation for the human and mouse genomes". eng. In: *Nucleic Acids Research* 47.D1, pp. D766–D773. ISSN: 1362-4962. DOI: 10.1093/nar/gky955.

Frayn, Keith N. et al. (2006). "Fatty acid metabolism in adipose tissue, muscle and liver in health and disease". eng. In: *Essays in Biochemistry* 42, pp. 89–103. ISSN: 0071-1365. DOI: 10.1042/bse0420089.

Gaedigk, Andrea (Oct. 2013). "Complexities of CYP2D6 gene analysis and interpretation". In: *International Review of Psychiatry* 25.5, pp. 534–553. ISSN: 0954-0261. DOI: 10.3109/09540261.2013.825581.

Gaedigk, Andrea et al. (Jan. 2017). "Prediction of CYP2D6 phenotype from genotype across world populations". en. In: *Genetics in Medicine* 19.1, pp. 69–76. ISSN: 1530-0366. DOI: 10.1038/gim.2016.80.

Gazave, E. et al. (Nov. 2013). "Population growth inflates the per-individual number of deleterious mutations and reduces their mean effect". In: *Genetics* 195.3. Edition: 2013/08/28, pp. 969–78. ISSN: 1943-2631 (Electronic) 0016-6731 (Linking). DOI: 10.1534/genetics.113.153973.

Gellner, Klaus et al. (2001). "Genomic organization of the human CYP3A locus: identification of a new, inducible CYP3A gene". In: *Pharmacogenetics and Genomics* 11.2, pp. 111–121. ISSN: 1744-6872.

Genomes Project, Consortium et al. (Oct. 2015). “A global reference for human genetic variation”. In: *Nature* 526.7571. Edition: 2015/10/04, pp. 68–74. ISSN: 1476-4687 (Electronic) 0028-0836 (Linking). DOI: 10.1038/nature15393.

Gilad, Yoav et al. (Aug. 2008). “Revealing the architecture of gene regulation: the promise of eQTL studies”. In: *Trends in genetics : TIG* 24.8, pp. 408–415. ISSN: 0168-9525. DOI: 10.1016/j.tig.2008.06.001.

Gonzalez, Frank J. et al. (Jan. 1990). “Evolution of the P450 gene superfamily:: animal-plant ‘warfare’, molecular drive and human genetic differences in drug oxidation”. In: *Trends in Genetics* 6, pp. 182–186. ISSN: 0168-9525. DOI: 10.1016/0168-9525(90)90174-5.

Guengerich, F. Peter (Apr. 1999). “CYTOCHROME P-450 3A4: Regulation and Role in Drug Metabolism”. In: *Annual Review of Pharmacology and Toxicology* 39.1. Publisher: Annual Reviews, pp. 1–17. ISSN: 0362-1642. DOI: 10.1146/annurev.pharmtox.39.1.1.

Guttman, Yelena et al. (Mar. 2019). “Polymorphism in Cytochrome P450 3A4 Is Ethnicity Related”. In: *Frontiers in Genetics* 10. ISSN: 1664-8021. DOI: 10.3389/fgene.2019.00224.

Hall, B. (2011). *Evolution: Principles and Processes*. Jones & Bartlett Learning. ISBN: 978-0-7637-6039-7.

Hamilton, M. (2011). *Population Genetics*. Wiley. ISBN: 978-1-4443-6245-9.

Hardwick, J. P. (June 2008). “Cytochrome P450 omega hydroxylase (CYP4) function in fatty acid metabolism and metabolic diseases”. In: *Biochem Pharmacol* 75.12. Edition: 2008/04/25, pp. 2263–75. ISSN: 1873-2968 (Electronic) 0006-2952 (Linking). DOI: 10.1016/j.bcp.2008.03.004.

Hartl, D.L. (1988). *A Primer of Population Genetics*. Grove/Atlantic, Incorporated. ISBN: 978-0-87893-301-3.

Hashizume, Takanori, Susumu Imaoka, Toyoko Hiroi, et al. (Feb. 2001). “cDNA Cloning and Expression of a Novel Cytochrome P450 (CYP4F12) from Human Small Intestine”. en. In: *Biochemical and Biophysical Research Communications* 280.4, pp. 1135–1141. ISSN: 0006-291X. DOI: 10.1006/bbrc.2000.4238.

Hashizume, Takanori, Susumu Imaoka, Masashi Mise, et al. (Jan. 2002). "Involvement of CYP2J2 and CYP4F12 in the metabolism of ebastine in human intestinal microsomes". eng. In: *The Journal of Pharmacology and Experimental Therapeutics* 300.1, pp. 298–304. ISSN: 0022-3565. DOI: 10.1124/jpet.300.1.298.

He, Hang et al. (2016). "Developmental regulation of CYP3A4 and CYP3A7 in Chinese Han population". In: *Drug Metabolism and Pharmacokinetics* 31.6, pp. 433–444. ISSN: 1347-4367. DOI: 10.1016/j.dmpk.2016.08.008.

Hellwege, Jacklyn et al. (Oct. 2017). "Population Stratification in Genetic Association Studies". In: *Current protocols in human genetics* 95, pp. 1.22.1–1.22.23. ISSN: 1934-8266. DOI: 10.1002/cphg.48.

Henry, J.P. et al. (2008). *Précis de génétique des populations: cours, exercices et problèmes résolus*. Dunod. ISBN: 978-2-10-051928-6.

Herzog, Michael H. et al. (2019). "The Multiple Testing Problem". en. In: *Understanding Statistics and Experimental Design : How to Not Lie with Statistics*. Ed. by Michael H. Herzog et al. Learning Materials in Biosciences. Cham: Springer International Publishing, pp. 63–66. ISBN: 978-3-030-03499-3. DOI: 10.1007/978-3-030-03499-3_5.

Holsinger, K. E. et al. (Sept. 2009). "Genetics in geographically structured populations: defining, estimating and interpreting F(ST)". In: *Nat Rev Genet* 10.9. Edition: 2009/08/19, pp. 639–50. ISSN: 1471-0064 (Electronic) 1471-0056 (Linking). DOI: 10.1038/nrg2611.

Hosking, Louise et al. (May 2004). "Detection of genotyping errors by Hardy–Weinberg equilibrium testing". en. In: *European Journal of Human Genetics* 12.5. Number: 5 Publisher: Nature Publishing Group, pp. 395–399. ISSN: 1476-5438. DOI: 10.1038/sj.ejhg.5201164.

Hotz, Alrun et al. (2018). "Mutation update for CYP4F22 variants associated with autosomal recessive congenital ichthyosis". en. In: *Human Mutation* 39.10. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/humu.23594, pp. 1305–1313. ISSN: 1098-1004. DOI: https://doi.org/10.1002/humu.23594.

Huang, Yang et al. (May 2013). "eQTL Epistasis – Challenges and Computational Approaches". In: *Frontiers in Genetics* 4. ISSN: 1664-8021. DOI: 10.3389/fgene.2013.00051.

Hussin, Julie G. et al. (Apr. 2015). "Recombination affects accumulation of damaging and disease-associated mutations in human populations". eng. In: *Nature Genetics* 47.4, pp. 400–404. ISSN: 1546-1718. DOI: 10.1038/ng.3216.

Ingelman-Sundberg, M. et al. (Dec. 2007). "Influence of cytochrome P450 polymorphisms on drug therapies: pharmacogenetic, pharmacoepigenetic and clinical aspects". In: *Pharmacol Ther* 116.3. Edition: 2007/11/16, pp. 496–526. ISSN: 0163-7258 (Print) 0163-7258 (Linking). DOI: 10.1016/j.pharmthera.2007.09.004.

Ingelman-Sundberg, Magnus et al. (May 2018). "Integrating rare genetic variants into pharmacogenetic drug response predictions". In: *Human Genomics* 12.1, p. 26. ISSN: 1479-7364. DOI: 10.1186/s40246-018-0157-3.

Institute, Babraham (2021). *FastQC A Quality Control tool for High Throughput Sequence Data*.

Janha, Ramatoulie E et al. (Apr. 2014). "Inactive alleles of cytochrome P450 2C19 may be positively selected in human evolution". In: *BMC Evolutionary Biology* 14, p. 71. ISSN: 1471-2148. DOI: 10.1186/1471-2148-14-71.

Jenkinson, Garrett et al. (Nov. 2020). "LeafCutterMD: an algorithm for outlier splicing detection in rare diseases". In: *Bioinformatics* 36.17, pp. 4609–4615. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btaa259.

Jin, Yi et al. (Feb. 2011). "CYP4F Enzymes Are Responsible for the Elimination of Fingolimod (FTY720), a Novel Treatment of Relapsing Multiple Sclerosis". en. In: *Drug Metabolism and Disposition* 39.2. Publisher: American Society for Pharmacology and Experimental Therapeutics Section: Article, pp. 191–198. ISSN: 0090-9556, 1521-009X. DOI: 10.1124/dmd.110.035378.

Kalsotra, A. et al. (Dec. 2006). "Cytochrome P450 4F subfamily: at the crossroads of eicosanoid and drug metabolism". In: *Pharmacol Ther* 112.3. Edition: 2006/08/24, pp. 589–

611. ISSN: 0163-7258 (Print) 0163-7258 (Linking). DOI: 10.1016/j.pharmthera.2006.03.008.

Kalsotra, Auinash et al. (Sept. 2004). "Expression and characterization of human cytochrome P450 4F11: Putative role in the metabolism of therapeutic drugs and eicosanoids". en. In: *Toxicology and Applied Pharmacology*. Mechanisms Regulating Enzymes Involved in Xenobiotic Disposition: A Tribute to Ed Bresnick 199.3, pp. 295–304. ISSN: 0041-008X. DOI: 10.1016/j.taap.2003.12.033.

Kawashima, A. et al. (2014). "Substrate-dependent evolution of cytochrome P450: rapid turnover of the detoxification-type and conservation of the biosynthesis-type". In: *PLoS One* 9.6. Edition: 2014/07/01, e100059. ISSN: 1932-6203 (Electronic) 1932-6203 (Linking). DOI: 10.1371/journal.pone.0100059.

Keinan, A. et al. (May 2012). "Recent explosive human population growth has resulted in an excess of rare genetic variants". In: *Science* 336.6082. Edition: 2012/05/15, pp. 740–3. ISSN: 1095-9203 (Electronic) 0036-8075 (Linking). DOI: 10.1126/science.1217283.

Kent, W. James et al. (June 2002). "The human genome browser at UCSC". eng. In: *Genome Research* 12.6, pp. 996–1006. ISSN: 1088-9051. DOI: 10.1101/gr.229102.

Kim, Yuseob et al. (Feb. 2002). "Detecting a Local Signature of Genetic Hitchhiking Along a Recombining Chromosome". en. In: *Genetics* 160.2. Publisher: Genetics Section: INVESTIGATIONS, pp. 765–777. ISSN: 0016-6731, 1943-2631.

Kimura, M. (Aug. 1991). "The neutral theory of molecular evolution: a review of recent evidence". eng. In: *Idengaku Zasshi* 66.4, pp. 367–386. ISSN: 0021-504X. DOI: 10.1266/jjg.66.367.

Kimura, Motoo (Nov. 1987). "Molecular evolutionary clock and the neutral theory". en. In: *Journal of Molecular Evolution* 26.1, pp. 24–33. ISSN: 1432-1432. DOI: 10.1007/BF02111279.

Kirischian, N. L. et al. (Jan. 2012). "Phylogenetic and functional analyses of the cytochrome P450 family 4". In: *Mol Phylogenet Evol* 62.1. Edition: 2011/11/15, pp. 458–71. ISSN: 1095-9513 (Electronic) 1055-7903 (Linking). DOI: 10.1016/j.ympev.2011.10.016.

Kitada, Mitsukazu et al. (Aug. 1985). "Purification and properties of cytochrome P-450 from homogenates of human fetal livers". en. In: *Archives of Biochemistry and Biophysics* 241.1, pp. 275–280. ISSN: 0003-9861. DOI: 10.1016/0003-9861(85)90383-2.

Komori, Masayuki et al. (1989). "Molecular Cloning and Sequence Analysis of cDNA Containing the Entire Coding Region for Human Fetal Liver Cytochrome P-450". en. In: 105.2, p. 3.

Krishnamurthy, K.V. (2003). *Textbook of Biodiversity*. Taylor & Francis. ISBN: 978-1-57808-325-1.

Kudaravalli, Sridhar et al. (Mar. 2009). "Gene expression levels are a target of recent natural selection in the human genome". eng. In: *Molecular Biology and Evolution* 26.3, pp. 649–658. ISSN: 1537-1719. DOI: 10.1093/molbev/msn289.

Kuehl, Peter et al. (Apr. 2001). "Sequence diversity in CYP3A promoters and characterization of the genetic basis of polymorphic CYP3A5 expression". en. In: *Nature Genetics* 27.4. Number: 4 Publisher: Nature Publishing Group, pp. 383–391. ISSN: 1546-1718. DOI: 10.1038/86882.

Kukurba, Kimberly R. et al. (Apr. 2015). "RNA Sequencing and Analysis". In: *Cold Spring Harbor protocols* 2015.11, pp. 951–969. ISSN: 1940-3402. DOI: 10.1101/pdb.top084970.

Lacroix, Dan et al. (1997). "Expression of CYP3A in the Human Liver — Evidence that the Shift between CYP3A7 and CYP3A4 Occurs Immediately After Birth". In: *European Journal of Biochemistry* 247.2, pp. 625–634. ISSN: 0014-2956. DOI: 10.1111/j.1432-1033.1997.00625.x.

Lamba, Jatinder et al. (July 2012). "PharmGKB summary: very important pharmacogene information for CYP3A5". In: *Pharmacogenetics and genomics* 22.7, pp. 555–558. ISSN: 1744-6872. DOI: 10.1097/FPC.0b013e328351d47f.

Law, Charity W. et al. (Feb. 2014). "voom: precision weights unlock linear model analysis tools for RNA-seq read counts". en. In: *Genome Biology* 15.2. Number: 2 Publisher: BioMed Central, pp. 1–17. ISSN: 1474-760X. DOI: 10.1186/gb-2014-15-2-r29.

Lawrence, Michael et al. (July 2009). "rtracklayer: an R package for interfacing with genome browsers". In: *Bioinformatics* 25.14, pp. 1841–1842. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp328.

Le Quéré, Valérie et al. (Aug. 2004). "Human CYP4F3s are the main catalysts in the oxidation of fatty acid epoxides". en. In: *Journal of Lipid Research* 45.8, pp. 1446–1458. ISSN: 00222275. DOI: 10.1194/jlr.M300463-JLR200.

Lefevre, T. et al. (2016). *Biologie évolutive*. De Boeck supérieur. ISBN: 978-2-8073-0296-9.

Lehner, Ben (Aug. 2011). "Molecular mechanisms of epistasis within and between genes". en. In: *Trends in Genetics* 27.8, pp. 323–331. ISSN: 0168-9525. DOI: 10.1016/j.tig.2011.05.007.

Lewontin, R. C. (1964). "The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models". eng. In: *Genetics* 49.1, pp. 49–67. ISSN: 0016-6731.

Li, H. et al. (Sept. 2019). "Neonatal cytochrome P450 CYP3A7: A comprehensive review of its role in development, disease, and xenobiotic metabolism". In: *Arch Biochem Biophys* 673. Edition: 2019/08/26, p. 108078. ISSN: 1096-0384 (Electronic) 0003-9861 (Linking). DOI: 10.1016/j.abb.2019.108078.

Li, Heng (Nov. 2011). "A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data". In: *Bioinformatics* 27.21, pp. 2987–2993. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr509.

Li, J. et al. (Feb. 2011). "Global patterns of genetic diversity and signals of natural selection for human ADME genes". In: *Hum Mol Genet* 20.3. Edition: 2010/11/18, pp. 528–40. ISSN: 1460-2083 (Electronic) 0964-6906 (Linking). DOI: 10.1093/hmg/ddq498.

Li, Lun et al. (2012). "eQTL". en. In: *Quantitative Trait Loci (QTL): Methods and Protocols*. Ed. by Scott A. Rifkin. Methods in Molecular Biology. Totowa, NJ: Humana Press, pp. 265–279. ISBN: 978-1-61779-785-9. DOI: 10.1007/978-1-61779-785-9_14.

Liang, Ruijuan et al. (July 2012). "Influence of CYP4F2 genotype on warfarin dose requirement–a systematic review and meta-analysis". en. In: *Thrombosis Research* 130.1, pp. 38–44. ISSN: 0049-3848. DOI: 10.1016/j.thromres.2011.11.043.

Lichten, Michael et al. (Dec. 1995). "Meiotic recombination hotspots". In: *Annual Review of Genetics* 29.1. Publisher: Annual Reviews, pp. 423–444. ISSN: 0066-4197. DOI: 10.1146/annurev.ge.29.120195.002231.

Liu, Jinbo et al. (Sept. 2014). "Serum Free Fatty Acid Biomarkers of Lung Cancer". en. In: *Chest* 146.3, pp. 670–679. ISSN: 0012-3692. DOI: 10.1378/chest.13-2568.

Llaurens, V. et al. (May 2017). "Genetic architecture and balancing selection: the life and death of differentiated variants". In: *Mol Ecol* 26.9. Edition: 2017/02/09, pp. 2430–2448. ISSN: 1365-294X (Electronic) 0962-1083 (Linking). DOI: 10.1111/mec.14051.

Lonsdale, John et al. (June 2013). "The Genotype-Tissue Expression (GTEx) project". en. In: *Nature Genetics* 45.6. Number: 6 Publisher: Nature Publishing Group, pp. 580–585. ISSN: 1546-1718. DOI: 10.1038/ng.2653.

Low-Kam, C. et al. (Nov. 2016). "Whole-genome sequencing in French Canadians from Quebec". In: *Hum Genet* 135.11. Edition: 2016/07/05, pp. 1213–1221. ISSN: 1432-1203 (Electronic) 0340-6717 (Linking). DOI: 10.1007/s00439-016-1702-6.

Lowe, Rohan et al. (May 2017). "Transcriptomics technologies". In: *PLoS Computational Biology* 13.5. ISSN: 1553-734X. DOI: 10.1371/journal.pcbi.1005457.

Lu, BingXin et al. (Feb. 2013). "Comparative study of de novo assembly and genome-guided assembly strategies for transcriptome reconstruction based on RNA-Seq". en. In: *Science China Life Sciences* 56.2, pp. 143–155. ISSN: 1869-1889. DOI: 10.1007/s11427-013-4442-z.

Maclean, Colin A. et al. (Nov. 2015). "hapbin: An Efficient Program for Performing Haplotype-Based Scans for Positive Selection in Large Genomic Datasets". In: *Molecular Biology and Evolution* 32.11, pp. 3027–3029. ISSN: 0737-4038. DOI: 10.1093/molbev/msv172.

McArthur, Andrew G. et al. (Aug. 2003). "Phylogenetic Analysis of the Cytochrome P450 3 (CYP3) Gene Family". en. In: *Journal of Molecular Evolution* 57.2, pp. 200–211. ISSN: 1432-1432. DOI: 10.1007/s00239-003-2466-x.

McInnes, Leland et al. (Feb. 2018). "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction". en. In:

Menozzi, P. et al. (Sept. 1978). "Synthetic maps of human gene frequencies in Europeans". eng. In: *Science (New York, N.Y.)* 201.4358, pp. 786–792. ISSN: 0036-8075. DOI: 10.1126/science.356262.

Mertes, Christian et al. (Jan. 2021). "Detection of aberrant splicing events in RNA-seq data using FRASER". en. In: *Nature Communications* 12.1. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Computational models;Disease genetics;RNA sequencing;RNA splicing Subject_term_id: computational-models;disease-genetics;rna-sequencing;rna-splicing, p. 529. ISSN: 2041-1723. DOI: 10.1038/s41467-020-20573-7.

Nebert, Daniel W. and Timothy P. Dalton (Dec. 2006). "The role of cytochrome P450 enzymes in endogenous signalling pathways and environmental carcinogenesis". en. In: *Nature Reviews Cancer* 6.12. Number: 12 Publisher: Nature Publishing Group, pp. 947–960. ISSN: 1474-1768. DOI: 10.1038/nrc2015.

Nebert, Daniel W., Kjell Wikvall, et al. (Feb. 2013). "Human cytochromes P450 in health and disease". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 368.1612. Publisher: Royal Society, p. 20120431. DOI: 10.1098/rstb.2012.0431.

Nelson, D. R. et al. (Feb. 1996). "P450 superfamily: update on new sequences, gene mapping, accession numbers and nomenclature". eng. In: *Pharmacogenetics* 6.1, pp. 1–42. ISSN: 0960-314X. DOI: 10.1097/00008571-199602000-00002.

Nelson, David R et al. (2004). "Comparison of cytochrome P450 (CYP) genes from the mouse and human genomes, including nomenclature recommendations for genes, pseudogenes

and alternative-splice variants". In: *Pharmacogenetics and Genomics* 14.1, pp. 1–18. ISSN: 1744-6872.

Nica, Alexandra C. et al. (June 2013). "Expression quantitative trait loci: present and future". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 368.1620. ISSN: 0962-8436. DOI: 10.1098/rstb.2012.0362.

Novembre, John et al. (Nov. 2008). "Genes mirror geography within Europe". In: *Nature* 456.7218, pp. 98–101. ISSN: 0028-0836. DOI: 10.1038/nature07331.

Ogu, Chris C. et al. (Oct. 2000). "Drug interactions due to cytochrome P450". In: *Proceedings (Baylor University. Medical Center)* 13.4, pp. 421–423. ISSN: 0899-8280.

Ohno, Yusuke et al. (June 2015). "Essential role of the cytochrome P450 CYP4F22 in the production of acylceramide, the key lipid for skin permeability barrier formation". en. In: *Proceedings of the National Academy of Sciences* 112.25. Publisher: National Academy of Sciences Section: Biological Sciences, pp. 7707–7712. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1503491112.

Orr, H. Allen (Aug. 2009). "Fitness and its role in evolutionary genetics". In: *Nature reviews. Genetics* 10.8, pp. 531–539. ISSN: 1471-0056. DOI: 10.1038/nrg2603.

Pan, S. T. et al. (June 2016). "Computational Identification of the Paralogs and Orthologs of Human Cytochrome P450 Superfamily and the Implication in Drug Discovery". In: *Int J Mol Sci* 17.7. Edition: 2016/07/02. ISSN: 1422-0067 (Electronic) 1422-0067 (Linking). DOI: 10.3390/ijms17071020.

Petrov, Anton et al. (Nov. 2004). "Microarray Image Processing and Quality Control". en. In: *Journal of VLSI signal processing systems for signal, image and video technology* 38.3, pp. 211–226. ISSN: 0922-5773. DOI: 10.1023/B:VLSI.0000042488.08307.ad.

Pink, Ryan Charles et al. (May 2011). "Pseudogenes: Pseudo-functional or key regulators in health and disease?" In: *RNA* 17.5, pp. 792–798. ISSN: 1355-8382. DOI: 10.1261/rna.2658311.

"Population Stratification" (2006). en. In: *Encyclopedic Reference of Genomics and Proteomics in Molecular Medicine*. Berlin, Heidelberg: Springer, pp. 1444–1445. ISBN: 978-3-540-29623-2. DOI: 10.1007/3-540-29623-9_8284.

Powell, Pnina K. et al. (June 1998). "Metabolism of Arachidonic Acid to 20-Hydroxy-5,8,11,14-eicosatetraenoic Acid by P450 Enzymes in Human Liver: Involvement of CYP4F2 and CYP4A11". en. In: *Journal of Pharmacology and Experimental Therapeutics* 285.3. Publisher: American Society for Pharmacology and Experimental Therapeutics Section: DRUG METABOLISM AND DISPOSITION, pp. 1327–1336. ISSN: 0022-3565, 1521-0103.

Pritchard, J. K. et al. (2001). "Linkage disequilibrium in humans: models and data". eng. In: *American journal of human genetics* 69.1. Edition: 2001/06/14, pp. 1–14. ISSN: 0002-9297 1537-6605. DOI: 10.1086/321275.

Qiu, Huan et al. (2008). "CYP3 phylogenomics: evidence for positive selection of CYP3A4 and CYP3A7". In: *Pharmacogenetics and Genomics* 18.1, pp. 53–66. ISSN: 1744-6872. DOI: 10.1097/FPC.0b013e3282f313f8.

Quiver, Melanie H. et al. (Oct. 2018). "Adaptive eQTLs reveal the evolutionary impacts of pleiotropy and tissue-specificity, while contributing to health and disease in human populations". en. In: *bioRxiv*. Publisher: Cold Spring Harbor Laboratory Section: New Results, p. 444737. DOI: 10.1101/444737.

Reece, J.B. et al. (2012). *Campbell Biologie*. 4e. Pearson ERPI. ISBN: 2-7613-2856-6.

Reich, David et al. (May 2008). "Principal component analysis of genetic data". en. In: *Nature Genetics* 40.5. Number: 5 Publisher: Nature Publishing Group, pp. 491–492. ISSN: 1546-1718. DOI: 10.1038/ng0508-491.

Relethford, J.H. (2012). *Human Population Genetics*. Wiley. ISBN: 978-1-118-18162-1.

Richman, Adam (2000). "Evolution of balanced genetic polymorphism". In: *Molecular Ecology* 9.12, pp. 1953–1963. ISSN: 0962-1083. DOI: 10.1046/j.1365-294X.2000.01125.x.

Ritchie, Matthew E. et al. (Apr. 2015). "limma powers differential expression analyses for RNA-sequencing and microarray studies". In: *Nucleic Acids Research* 43.7, e47–e47. ISSN: 0305-1048. DOI: 10.1093/nar/gkv007.

Rohlfs, Rori V. et al. (May 2010). "Detecting Coevolution through Allelic Association between Physically Unlinked Loci". In: *American Journal of Human Genetics* 86.5, pp. 674–685. ISSN: 0002-9297. DOI: 10.1016/j.ajhg.2010.03.001.

Rojas, L. et al. (Feb. 2015). "Effect of CYP3A5*3 on kidney transplant recipients treated with tacrolimus: a systematic review and meta-analysis of observational studies". en. In: *The Pharmacogenomics Journal* 15.1. Number: 1 Publisher: Nature Publishing Group, pp. 38–48. ISSN: 1473-1150. DOI: 10.1038/tpj.2014.38.

Sabeti, Pardis C. et al. (Oct. 2002). "Detecting recent positive selection in the human genome from haplotype structure". In: *Nature* 419.6909, pp. 832–837. ISSN: 1476-4687. DOI: 10.1038/nature01140.

Scott, S. A. et al. (2013). "Clinical Pharmacogenetics Implementation Consortium Guidelines for CYP2C19 Genotype and Clopidogrel Therapy: 2013 Update". en. In: *Clinical Pharmacology & Therapeutics* 94.3, pp. 317–323. ISSN: 1532-6535. DOI: https://doi.org/10.1038/clpt.2013.105.

Scriver, Charles R. (2001). "Human Genetics: Lessons from Quebec Populations". In: *Annual Review of Genomics and Human Genetics* 2.1, pp. 69–101. DOI: 10.1146/annurev.genom.2.1.69.

Sevrioukova, Irina F. et al. (Mar. 2013). "UNDERSTANDING THE MECHANISM OF CYTOCHROME P450 3A4: RECENT ADVANCES AND REMAINING PROBLEMS". In: *Dalton transactions (Cambridge, England : 2003)* 42.9, pp. 3116–3126. ISSN: 1477-9226. DOI: 10.1039/c2dt31833d.

Shendre, Aditi et al. (Mar. 2016). "Race-specific influence of CYP4F2 on dose and risk of hemorrhage among warfarin users". In: *Pharmacotherapy* 36.3, pp. 263–272. ISSN: 0277-0008. DOI: 10.1002/phar.1717.

Shiotani, Akiko et al. (2013). "Novel Single Nucleotide Polymorphism Markers for Low Dose Aspirin-Associated Small Bowel Bleeding". en. In: *PLOS ONE* 8.12. Publisher: Public Library of Science, e84244. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0084244.

Siewert, K. M. et al. (Nov. 2017). "Detecting Long-Term Balancing Selection Using Allele Frequency Correlation". In: *Mol Biol Evol* 34.11. Edition: 2017/10/06, pp. 2996–3005. ISSN: 1537-1719 (Electronic) 0737-4038 (Linking). DOI: 10.1093/molbev/msx209.

Sim, Sarah C. et al. (Sept. 2005). "CYP3A7 protein expression is high in a fraction of adult human livers and partially associated with the CYP3A7*1C allele". en-US. In: *Pharmacogenetics and Genomics* 15.9, pp. 625–631. ISSN: 1744-6872. DOI: 10.1097/01.fpc.0000171516.84139.89.

Singh, Onkar et al. (2011). "Influence of CYP4F rs2108622 (V433M) on Warfarin Dose Requirement in Asian Patients". In: *Drug Metabolism and Pharmacokinetics* 26.2, pp. 130–136. DOI: 10.2133/dmpk.DMPK-10-RG-080.

Slatkin, Montgomery (June 2008). "Linkage disequilibrium — understanding the evolutionary past and mapping the medical future". In: *Nature reviews. Genetics* 9.6, pp. 477–485. ISSN: 1471-0056. DOI: 10.1038/nrg2361.

Smit, Pauline et al. (Sept. 2005). "A Common Polymorphism in the CYP3A7 Gene Is Associated with a Nearly 50% Reduction in Serum Dehydroepiandrosterone Sulfate Levels". In: *The Journal of Clinical Endocrinology & Metabolism* 90.9, pp. 5313–5316. ISSN: 0021-972X. DOI: 10.1210/jc.2005-0307.

Sontag, Timothy J. et al. (July 2002). "Cytochrome P450 -Hydroxylase Pathway of Tocopherol Catabolism: NOVEL MECHANISM OF REGULATION OF VITAMIN E STATUS*". en. In: *Journal of Biological Chemistry* 277.28, pp. 25290–25296. ISSN: 0021-9258. DOI: 10.1074/jbc.M201466200.

Stark, Katarina et al. (Sept. 2005). "Oxygenation of polyunsaturated long chain fatty acids by recombinant CYP4F8 and CYP4F12 and catalytic importance of Tyr-125 and Gly-328 of CYP4F8". en. In: *Archives of Biochemistry and Biophysics* 441.2, pp. 174–181. ISSN: 0003-9861. DOI: 10.1016/j.abb.2005.07.003.

Stegle, Oliver et al. (Feb. 2012). "Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses". In: *Nature protocols* 7.3, pp. 500–507. ISSN: 1754-2189. DOI: 10.1038/nprot.2011.457.

Stone, Angie et al. (May 2005). "CYP3A43 Pro340Ala Polymorphism and Prostate Cancer Risk in African Americans and Caucasians". en. In: *Cancer Epidemiology and Prevention Biomarkers* 14.5. Publisher: American Association for Cancer Research Section: Research Articles, pp. 1257–1261. ISSN: 1055-9965, 1538-7755. DOI: 10.1158/1055-9965.EPI-04-0534.

Stranger, Barbara E. et al. (Oct. 2007). "Population genomics of human gene expression". In: *Nature genetics* 39.10, pp. 1217–1224. ISSN: 1061-4036. DOI: 10.1038/ng2142.

Sudmant, Peter H. et al. (Oct. 2015). "An integrated map of structural variation in 2,504 human genomes". en. In: *Nature* 526.7571. Number: 7571 Publisher: Nature Publishing Group, pp. 75–81. ISSN: 1476-4687. DOI: 10.1038/nature15394.

Tajima, F. (1989). "Statistical method for testing the neutral mutation hypothesis by DNA polymorphism". eng. In: *Genetics* 123.3, pp. 585–595. ISSN: 0016-6731.

Tavira, Beatriz et al. (Aug. 2013). "A search for new CYP3A4 variants as determinants of tacrolimus dose requirements in renal-transplanted patients". eng. In: *Pharmacogenetics and Genomics* 23.8, pp. 445–448. ISSN: 1744-6880. DOI: 10.1097/FPC.0b013e3283636856.

Thompson, E. E., H. Kuttab-Boulos, D. Witonsky, et al. (2004). "CYP3A variation and the evolution of salt-sensitivity variants". eng. In: *American journal of human genetics* 75.6. Edition: 2004/10/18, pp. 1059–1069. ISSN: 0002-9297 1537-6605. DOI: 10.1086/426406.

Thompson, E. E., H. Kuttab-Boulos, L. Yang, et al. (Mar. 2006). "Sequence diversity and haplotype structure at the human CYP3A cluster". In: *Pharmacogenomics J* 6.2. Edition: 2005/11/30, pp. 105–14. ISSN: 1470-269X (Print) 1470-269X (Linking). DOI: 10.1038/sj.tpj.6500347.

Tweedie, Susan et al. (Jan. 2021). "Genenames.org: the HGNC and VGNC resources in 2021". In: *Nucleic Acids Research* 49.D1, pp. D939–D946. ISSN: 0305-1048. DOI: 10.1093/nar/gkaa980.

Vaquero-Garcia, Jorge et al. (Feb. 2016). "A new view of transcriptome complexity and regulation through the lens of local splicing variations". In: *eLife* 5. Ed. by Juan Valcárcel. Publisher: eLife Sciences Publications, Ltd, e11752. ISSN: 2050-084X. DOI: 10.7554/eLife.11752.

Voight, Benjamin F. et al. (2006). "A Map of Recent Positive Selection in the Human Genome". In: *PLOS Biology* 4.3, e72. DOI: 10.1371/journal.pbio.0040072.

Wang, D. et al. (Aug. 2011). "Intronic polymorphism in CYP3A4 affects hepatic expression and response to statin drugs". en. In: *The Pharmacogenomics Journal* 11.4. Number: 4 Publisher: Nature Publishing Group, pp. 274–286. ISSN: 1473-1150. DOI: 10.1038/tpj.2010.28.

Wang, Michael Zhuo et al. (Nov. 2007). "Human Enteric Microsomal CYP4F Enzymes O-Demethylate the Antiparasitic Prodrug Pafuramidine". en. In: *Drug Metabolism and Disposition* 35.11. Publisher: American Society for Pharmacology and Experimental Therapeutics Section: Article, pp. 2067–2075. ISSN: 0090-9556, 1521-009X. DOI: 10.1124/dmd.107.016428.

Wang, Zhong et al. (Jan. 2009). "RNA-Seq: a revolutionary tool for transcriptomics". In: *Nature reviews. Genetics* 10.1, pp. 57–63. ISSN: 1471-0056. DOI: 10.1038/nrg2484.

Watterson, G. A. (1975). "On the number of segregating sites in genetical models without recombination". In: *Theoretical Population Biology* 7.2, pp. 256–276. ISSN: 0040-5809. DOI: 10.1016/0040-5809(75)90020-9.

Weedall, G. D. et al. (July 2010). "Detecting signatures of balancing selection to identify targets of anti-parasite immunity". In: *Trends Parasitol* 26.7. Edition: 2010/05/15, pp. 363–9. ISSN: 1471-5007 (Electronic) 1471-4922 (Linking). DOI: 10.1016/j.pt.2010.04.002.

Weir, B. S. et al. (1984). "Estimating F-Statistics for the Analysis of Population Structure". In: *Evolution* 38.6, pp. 1358–1370. ISSN: 00143820, 15585646. DOI: 10.2307/2408641.

Werk, A. N. et al. (2014). "Functional Gene Variants of CYP3A4". en. In: *Clinical Pharmacology & Therapeutics* 96.3, pp. 340–348. ISSN: 1532-6535. DOI: 10.1038/clpt.2014.129.

Wright, S. (1984). *Evolution and the Genetics of Populations, Volume 4: Variability Within and Among Natural Populations*. University of Chicago Press. ISBN: 978-0-226-91041-3.

Wright, Sewall (1943). "ISOLATION BY DISTANCE". In: *Genetics* 28.2, p. 114.

Xu, Jinrui et al. (Mar. 2016). "Are Human Translated Pseudogenes Functional?" In: *Molecular Biology and Evolution* 33.3, pp. 755–760. ISSN: 0737-4038. DOI: 10.1093/molbev/msv268.

Yasukochi, Yoshiki et al. (Mar. 2015). "Molecular evolution of the CYP2D subfamily in primates: purifying selection on substrate recognition sites without the frequent or long-tract gene conversion". eng. In: *Genome Biology and Evolution* 7.4, pp. 1053–1067. ISSN: 1759-6653. DOI: 10.1093/gbe/evv056.

Yi, Myeongjin et al. (Apr. 2017). "Functional characterization of a common CYP4F11 genetic variant and identification of functionally defective CYP4F11 variants in erythromycin metabolism and 20-HETE synthesis". en. In: *Archives of Biochemistry and Biophysics* 620, pp. 43–51. ISSN: 0003-9861. DOI: 10.1016/j.abb.2017.03.010.

Yin, Jieyun et al. (2017). "Pathway-analysis of published genome-wide association studies of lung cancer: A potential role for the CYP4F3 locus". en. In: *Molecular Carcinogenesis* 56.6, pp. 1663–1672. ISSN: 1098-2744. DOI: https://doi.org/10.1002/mc.22622.

Yoichi Ishida (2009). "Sewall Wright and Gustave Malécot on Isolation by Distance". In: *Philosophy of Science* 76.5, pp. 784–796. DOI: 10.1086/605802.

Yu, Haiyuan et al. (Jan. 2007). "Positional artifacts in microarrays: experimental verification and construction of COP, an automated detection tool". In: *Nucleic Acids Research* 35.2, e8. ISSN: 0305-1048. DOI: 10.1093/nar/gkl871.

Zdanowicz, M.M. et al. (2010). *Concepts in Pharmacogenomics*. American Society of Health-System Pharmacists. ISBN: 978-1-58528-234-0.

Zhang, J. E. et al. (2017). "Effect of Genetic Variability in the CYP4F2, CYP4F11, and CYP4F12 Genes on Liver mRNA Levels and Warfarin Response". In: *Front Pharmacol* 8. Edition: 2017/06/18, p. 323. ISSN: 1663-9812 (Print) 1663-9812 (Linking). DOI: 10.3389/fphar.2017.00323.

Zhang, Yaping et al. (Sept. 2014). "Serum Unsaturated Free Fatty Acids: Potential Biomarkers for Early Detection and Disease Progression Monitoring of Non-Small Cell Lung Cancer". In: *Journal of Cancer* 5.8, pp. 706–714. ISSN: 1837-9664. DOI: 10.7150/jca.9787.

Annexe A

Schéma illustrant les différentes étapes réalisées afin d'obtenir la distribution nulle lors de l'analyse de déséquilibre de liaison

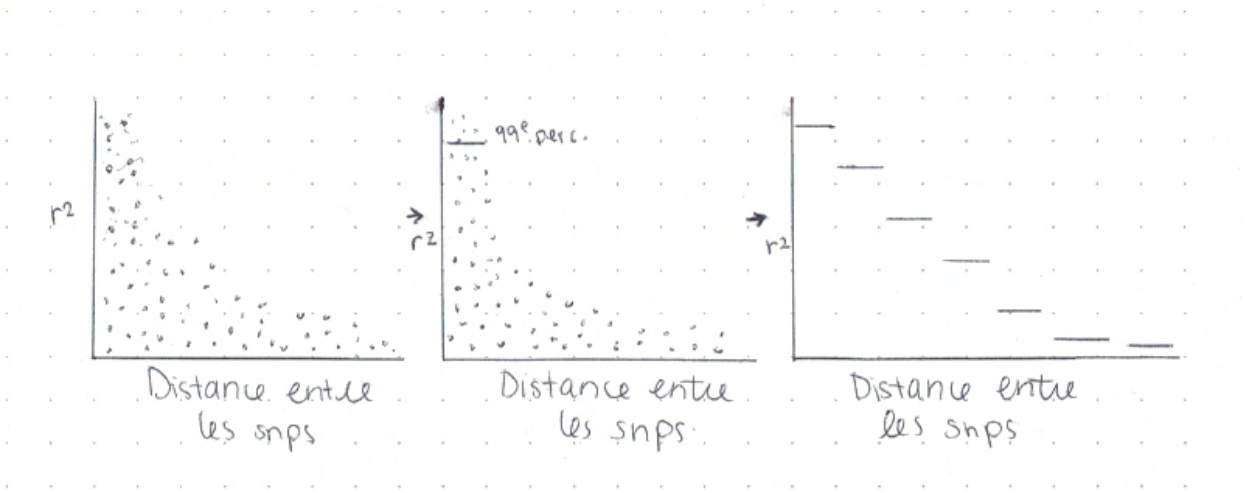


Fig. A.1. Schéma illustrant les différentes étapes réalisées afin d'obtenir la distribution nulle lors de l'analyse de déséquilibre de liaison. La première étape consiste à calculer le r^2 pour chaque paire de SNP. Lors de la deuxième étape, le 99e percentile est calculé sur des intervalles de distance fixes. Le dernier dessin représente la distribution nulle.