Amandine Lecerf Defer

DÉVELOPPEMENT ET VALIDATION D'UN ALLOSCORE EN TRANSPLANTATION RÉNALE

Laboratoire d'accueil : CRTI - Inserm UMR1064 – ATIP Avenir, Translational Immuno Genetics in AutoImmunity and Tranplantation- CHU Nantes

Maître de stage : Pierre-Antoine Gourraud
Encadrants : Nicolas Vince - Sophie Limou

Master 2 Bioinformatique pour les Biologistes

Université de Nantes-UFR Sciences et Techniques

Mémoire de Stage réalisé
Du 18 Mars 2019 au 02 Septembre 2019

Remerciements

Avant d'exposer mon expérience professionnelle dans ce laboratoire, je voudrais tout d'abord remercier les personnes travaillant au CRTI pour leur bon accueil et leur gentillesse. Je remercie également les étudiants réalisant leurs thèses ou leurs stages présents avec moi et Madame Axelle Hermouet-Durand qui ont su me mettre à l'aise et permettre que mon stage se déroule dans une ambiance conviviale. Leur gentillesse a permis de faire de ce stage un moment très profitable.

Je remercie, le Professeur Pierre-Antoine Gourraud pour son accueil au sein de son équipe et d'avoir permis la réalisation de mon stage. Je remercie aussi Haritiana de son aide lors des demandes administratives.

Je remercie en particulier Nicolas Vince et Sophie Limou pour leur bonne humeur et leur professionnalisme. Je les remercie de m'avoir encadrée et accompagnée sur l'approfondissement de mes connaissances et sur l'utilisation de nouveaux outils informatiques tout au long de cette expérience professionnelle.

Je remercie également Monsieur Laurent Mesnard et son thésard Robert Clerc de m'avoir transmis leurs données et les éléments essentiels au calcul de l'Alloscore.

Résumé

L'insuffisance rénale chronique touche plus de 10 % de la population et se caractérise par une perte progressive de la fonction rénale qui peut mener graduellement à une insuffisance rénale terminale. La transplantation rénale est le meilleur traitement contre l'insuffisance rénale terminale. En 2015, 36 700 patients vivaient avec un rein fonctionnel transplanté en France. La survie à court terme du greffon est bien contrôlée par des traitements immunosuppresseurs, mais la survie à long terme reste insuffisante (après 10 ans, seulement 50% des greffons sont encore fonctionnels). Les mécanismes physiopathologiques moléculaires conduisant au rejet chronique sont encore mal compris. La compatibilité HLA donneur-receveur est le principal facteur associé à la survie du greffon. Néanmoins, l'échec à long terme d'une greffe, même pour des paires HLA entièrement appariées, suggère que d'autres facteurs autre que HLA pourraient être immunogènes. En 2016, le Pr. Mesnard et ses collègues ont proposé un score, appelé score de non-appariement allogénomique, en additionnant les différences d'acides aminés non synonymes sur les protéines transmembranaires non-HLA entre donneur et receveur. Cet Alloscore non-HLA obtenu à partir de données séquencées d'exome entier (WES) présente une corrélation significative avec la fonction rénale post-transplantation de 3 ans, indépendamment de la compatibilité HLA. Nous avons cherché à valider ce score par évaluation du poids des variants rares par rapport aux variants communs, des variants imputés par rapport aux variants non imputés et des facteurs HLA par rapport aux facteurs non-HLA sur la fonction rénale post-transplantation. L'imputation des SNPs a été réalisée à l'aide du Haplotypes Reference Consortium (HRC) pour créer un ensemble de données imputées et l'analyse a été réalisé grâce aux langages par R et python. Le rôle joué par la distance génétique donneur-receveur, en calculant l'IBS (identité par état) et l'IBD (identité par descendance) en utilisant PLINK, a été analysé.

Ici, nous avons confirmé que l'Alloscore est indépendant de la compatibilité HLA (qui n'est donc pas l'unique facteur à prendre en compte lors d'une greffe) et que ce score se base essentiellement sur les variants rares. Cependant, les protéines transmembranaires ne seraient pas les seules à prendre en compte lors de greffe. L'imputation des exomes ne semble pas aussi performante que l'imputation de données GWAS.

Nos résultats préliminaires doivent encore être affinés et confirmés par une cohorte indépendante. Dans l'ensemble, ces résultats peuvent avoir un impact majeur sur la sélection future des donneurs, car ils peuvent simultanément améliorer l'attribution des greffons et réduire le risque de rejet.

Mots clé : Alloscore, GWAS, transplantation rénale, HLA, imputation

Abstract

Chronic kidney disease affects more than 10% of the population and is characterized by a progressive loss of kidney function that can gradually lead to end-stage renal disease. Renal transplantation is the best treatment for end-stage renal disease and in 2015, 36,700 patients were living with a functional kidney transplanted in France. Short-term graft survival is well controlled by immunosuppressive treatments, but long-term survival remains insufficient (after 10 years, only 50% of grafts are still functional). The molecular pathophysiological mechanisms leading to chronic rejection are still poorly understood. Donor-receiver HLA compatibility is the main factor associated with graft survival. Nevertheless, the long-term failure of a transplant, even for fully matched HLA pairs, suggests that factors other than HLA may be immunogenic. In 2016, Prof. Mesnard and colleagues proposed a score, called an allogenomic mismatch score, by adding the differences in non-synonymous amino acids on non-HLA transmembrane proteins between donor and recipient. This non-HLA Alloscore obtained from whole exome sequence data (WES) shows a significant correlation with 3-year post-transplant renal function, regardless of HLA compatibility. We sought to validate this score by evaluating the weight of rare variants versus common variants, imputed variants versus non-imputed variants, and HLA factors versus non-HLA factors on post-transplant renal function. Imputation of SNPs was performed using the Haplotypes Reference Consortium (HRC) to create a set imputed data and analysis was performed by R and python. The role played by donor-receiver genetic distance in calculating IBS (state identity) and IBD (progeny identity) using PLINK was analysed.

Here, we have confirmed that Alloscore is independent of HLA compatibility (which is therefore not the only factor to be taken into account when transplanting) and that this score is essentially based on rare variants. However, transmembrane proteins are not the only ones to be taken into account when transplanting. The imputation of exomes does not seem to be as efficient as the imputation of GWAS data.

Our preliminary results have yet to be refined and confirmed by an independent cohort. Overall, these results can have a major impact on future donor selection, as they can simultaneously improve graft allocation and reduce the risk of rejection.

Key-word : Alloscore, GWAS, kidney transplantation, HLA, imputation

Table des matières

Table des Figures

Introduction

1) Contexte Biologique

a. Les reins

Les reins sont situés de part et d'autre de la colonne vertébrale. Ces organes sont essentiels à l'humain et assurent de nombreuses fonctions.

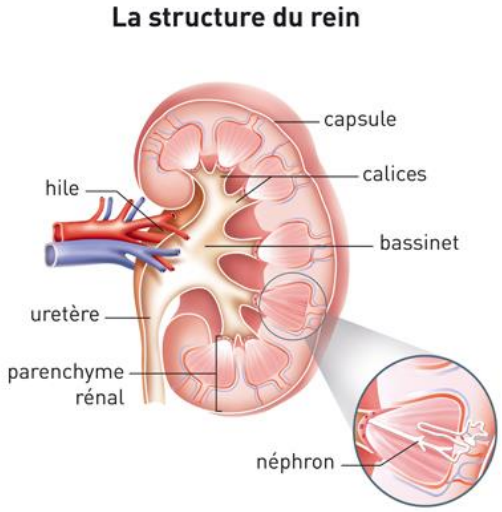


Figure 1 : Structure d'un rein : Organe en forme de haricot dont les éléments principaux sont les néphrons. [1]

La fonction première des reins est la filtration du sang et la production d'urine pour éliminer les déchets métaboliques produits par l'organisme (toxines, médicaments, polluants…). Les reins permettent aussi le maintien de l'équilibre homéostatique (PH, …). En effet, le volume et la composition de l'urine diffèrent en fonction des apports alimentaires et du métabolisme. Les glandes surrénales sécrètent des hormones indispensables à d'autres mécanismes de l'organisme (régulation de la pression artérielle, la maturation des globules rouges, …). Toutes ces fonctions sont contrôlées par l'unité fonctionnelle du rein : le néphron. Celui- ci se compose du corpuscule rénal qui assure la filtration du sang, des tubules rénaux et du tube collecteur qui absorbent de nombreux électrolytes. Chez l'Homme, un rein sain est composé d'environ un million de néphrons mais la fonction rénale est très variable selon l'individu (sexe, activité physique, …) et elle diminue avec l'âge.

La mesure des protéines présentes dans les urines, ainsi que le débit de filtration glomérulaire (GFR, glomerular filtration rate) permettent d'évaluer la fonction rénale. Le taux de protéine dans le sang est très faible lorsque les reins fonctionnent correctement mais si ces protéines se retrouvent en grande quantité, cela signifie qu'une pathologie rénale existe.

L'insuffisance rénale désigne la diminution plus ou moins importante des fonctions des reins qui perdent leur capacité à filtrer correctement le sang de l'organisme. En dessous d'un certain seuil de filtration des reins, on parle d'insuffisance rénale chronique (environ 10% de la population mondiale est touchée, soit entre 1,74 et 2,5 millions de personnes en France) qui peut aller jusqu'à l'insuffisance rénale terminale qui requiert une thérapie de remplacement (dialyse ou transplantation).

b. L'eGFR

Le volume sanguin filtré, lors de la filtration primaire, par le glomérule, par unité de temps, est la définition du taux de filtration glomérulaire (GFR, glomerule filtration rate). Pour la mesure du GFR, les urines sont récoltées pendant 24h ainsi qu'un échantillon sanguin pour doser la présence d'une substance exogène (iohexol) ou endogène (créatinine). L'eGFR correspond à une approximation du GFR (eGFR ou GFR estimé) basée uniquement sur un dosage sanguin de la créatinine et en prenant en compte l'âge, le sexe et l'origine ethnique de l'individu. Un test d'eGFR est employé pour surveiller les

personnes ayant une maladie rénale chronique (CKD, chronic kidney disease) et d'autres individus avec des facteurs de risque pour le CKD tel que le diabète, l'hypertension, les antécédents familiaux pour une maladie cardio-vasculaire. L'eGFR moyen chez un individu normal au repos est >90 mL/min/1,73m^2, bien que celui-ci baisse graduellement avec l'âge, sans pour autant avoir une maladie rénale[7] (Tableau 1).

Tableau 1 : Classification des stades d'évolution de la maladie rénale chronique [7]

Stade	eGFR (ml/min/1,73 m^2)	Définition
1	≥ 90	Maladie rénale chronique * avec eGFR normal ou augmenté
2	entr 60 et 89	Maladie rénale chronique * avec eGFR légèrement diminué
3A	entre 45 et 59	Insuffisance rénale chronique modérée
3B	entre 30 et 44	Insuffisance rénale chronique modérée
4	entre 15 et 29	Insuffisance rénale chronique sévère
5	<15	Insuffisance rénale chronique terminale

*avec marqueurs d'atteinte rénale : albuminurie, hématurie, leucocyturie, anomalies morphologiques ou histologiques, ou marqueurs de dysfonctionnement tubulaire, persistant plus de 3 mois (deux ou trois examens consécutifs)

Un taux d'eGFR supérieur à 90 mL/min/1,73m^2 signifie que les reins fonctionnent correctement. Si ce taux est compris entre 60 et 89 la fonction rénale est légèrement diminuée. De 59 à 30, l'insuffisance rénale est modérée mais cette insuffisance devient sévère si le taux d'eGFR est compris entre 29 et 15. Un taux d'eGFR inférieur à 15 traduit une insuffisance rénale chronique terminale qui nécessite une greffe ou un traitement par dialyse. L'insuffisance rénale chronique (IRC) correspond à une altération du fonctionnement des reins qui ne filtrent plus correctement le sang.

L'IRC est définie par une baisse du taux de l'eGFR inférieure à 60[2] (selon la Haute Autorité de Santé (HAS)) pendant une période de trois mois ou un eGFR de 60 avec la présence de dégâts dans le rein. Dans ce cas, le fonctionnement du rein est dit nui ou réduit. Le taux de créatinine dans le sang n'augmente pas jusqu'à ce que la fonction rénale soit devenue altérée. La diminution de l'eGFR est corrélée à la réduction du nombre de néphrons fonctionnels. Les deux principales causes de l'IRC sont l'hypertension artérielle et le diabète. La troisième cause (due aux glomérulonéphrites ou syndromes glomérulaires), se caractérise par une inflammation ou par des lésions des glomérules (Figure 2).

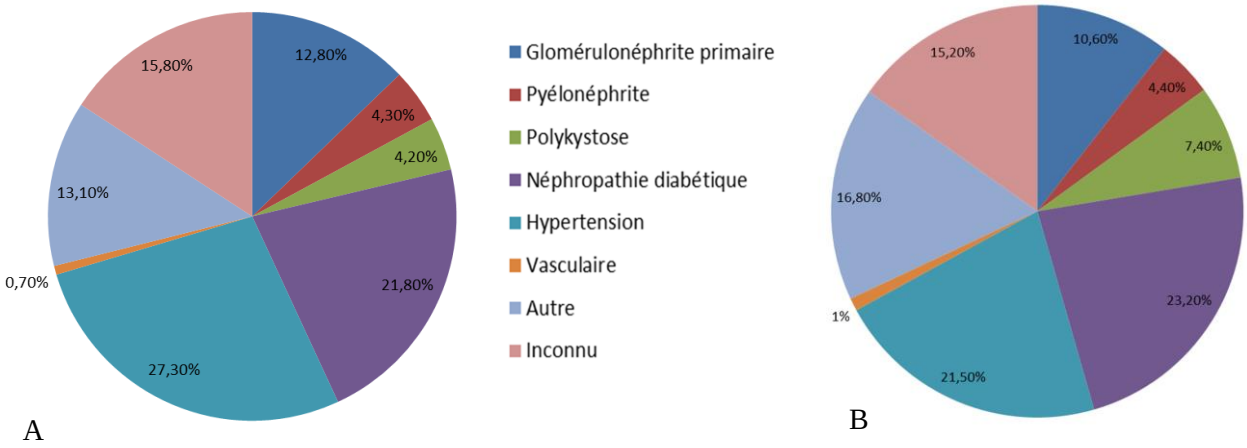


Figure 2 : répartition des causes de l'IRC chez l'homme (A) et la femme (B). [3]

c. La transplantation rénale

La transplantation rénale consiste à remplacer un rein défectueux par un rein sain venant d'un donneur qui peut être vivant ou mort. Souvent, le rein sain est posé sans que les reins malades ne soient enlevés (Figure 3).

Figure 3 : Schéma représentant une greffe de rein [4]

Pour qu'une greffe soit possible, il faut que le donneur et le receveur soient compatibles. Pour cela, trois critères sont pris en compte :

- la compatibilité ABO : Le groupe sanguin du donneur doit être compatible avec le groupe sanguin du receveur.
- la compatibilité tissulaire : plus le donneur et le receveur ont d'antigènes HLA en commun, meilleure sera la greffe.
- un crossmatch négatif : le receveur ne doit pas présenter d'anticorps contre le donneur.

2) Le HLA

Le HLA[5], « human leukocyte antigen », correspond au complexe majeur d'histocompatibilité (CMH) chez l'Homme. C'est le groupe de marqueurs génétiques de la compatibilité tissulaire humaine. Le HLA, codé par des gènes du chromosome 6, est exprimé à la surface des cellules ; ces cellules présentent des peptides antigéniques aux récepteurs des lymphocytes T. C'est-à-dire qu'il permet au système immunitaire de distinguer ses propres tissus (soi) de ce qu'il considère comme étranger (non-soi ; organes venant d'un donneur pour une greffe, virus, …). De ce fait, le système immunitaire s'attaque à tout corps étranger ou cellule qui ne présente pas un HLA complémentaire à sa surface. Ainsi, pour éviter le rejet lors d'une greffe, il faut trouver la plus grande compatibilité HLA possible.

3) Contexte du stage

La transplantation rénale est le meilleur traitement de l'insuffisance rénale terminale. En 2015, 36 700 patients vivaient avec un rein greffé fonctionnel en France. Les immunosuppresseurs permettent de contrôler la survie du greffon à court terme, cependant, la survie à long terme reste insuffisante (après 10 ans, seulement 50% des greffons sont encore fonctionnels) et les mécanismes conduisant au rejet chronique du greffon sont encore mal compris. Le facteur le plus important pour la survie du greffon est la compatibilité HLA entre le donneur et le receveur. En effet, plus le nombre de différences alléliques HLA entre le donneur et le receveur augmente, plus le risque de rejet augmente. Néanmoins, l'échec à long terme d'une greffe, même pour des paires HLA entièrement appariées, suggère que d'autres facteurs que HLA pourraient être immunogènes et qu'il existe de très nombreuses autres protéines portées par le greffon potentiellement immunogènes.

En 2016, Pr. Laurent Mesnard et ses collègues ont proposé un score, appelé Alloscore (score de non appariement génomique), en mesurant les différences d'acides aminés des protéines transmembranaires non-HLA entre le donneur et le receveur [6]. Cet Alloscore non-HLA obtenu à partir de données de séquences d'exomes entiers (WES, whole exome sequencing) se concentre sur des variantes rares et présente une corrélation significative avec la fonction rénale post-transplantation de 3 ans, indépendamment de la compatibilité HLA.

Au début de cette année, Reindl-Schwaighofer et ses collègues ont développé un score similaire (appelé SNP MM) à partir de données GWAS se concentrant sur des variants communs [7] et ont montré que le score SNP MM non-HLA est en corrélation significative avec la survie du greffon à 5 ans.

Objectifs du stage

L'objectif de ce stage est de démontrer la faisabilité d'un Alloscore à partir de données de type WES et de le comparer à des scores existants. Il est nécessaire de comprendre comment est mesuré l'Alloscore, de récolter et imputer des données de type WES en transplantation, d'appliquer le calcul de l'Alloscore à ces diverses données, de valider ce score et enfin de le comparer à des scores déjà existants. Le but final est de valider l'Alloscore sur des données de type GWAS.

Description des données

1) Les patients

L'échantillon principal (cohorte discovery) se compose de 10 couples de donneur-receveur de transplantation rénale ayant accepté de participer à l'essai clinique en transplantation d'organe 04 (CTOT-04, étude observationnelle multicentrique sur le diagnostic non invasif du rejet de greffe rénale par l'étude de l'ARNm des cellules urinaires). Parmi ces couples, les receveurs ont reçu une greffe de rein d'un donneur vivant. L'ADN de ces individus a été extrait du sang périphérique stocké à l'aide du kit de prélèvement de sang EZ1 DNA (Qiagen) pour ensuite permettre d'effectuer le test et le séquençage d'exome. Les informations démographiques et cliniques de cet échantillon sont présentées dans la table suivante (Tableau 2). Nous avons également à notre disposition deux autres cohortes. La cohorte Cornell qui est composée de 24 paires de donneur-receveur provenant de New-York et la cohorte Paris qui est composée de 19 paires de donneur-receveur avec les mêmes conditions.

Tableau 2 : Caractéristiques démographiques et cliniques de l'échantillon Discovery[6]

Characteristic	Discovery cohort
Number of Transplant Pairs with living donors	10/10
Allogenomics mismatch score AMS(SD)[range]	1335(304)[994–2033]
Clinical factors	
Age	
Donor (SD)	41 (13)
Recipient (SD)	48 (10)
Living Donor type	
Living related N (AMS) [SD])	4 (1116 [143])
Living unrelated N (AMS) [SD])	6 (1481 [300])
Donor sex	
Male (%)	2 (20%)
Female (%)	8 (80%)
Donor Race	
Black (%)	4(40%)
Non-Black (%)	6(60%)
Recipient sex	
Male (%)	9 (90%)
Female (%)	1 (10%)
Recipient Race	
Black (%)	4 (40%)
Non-Black (%)	6 (60%)
Number of HLA mismatches ABDR (SD)	3.9 (1.91)
Functional Factors	
Number of Patients at 12 months	10
Serum creatinine level at 12 months mg/dL (SD)	1.51 (0.35)
eGFR at 12 months ml/min/1.73m^2 (SD)	54.3(10)
Number of Patients at 24 months	9
Serum creatinine level at 24 months mg/dL (SD)	1.36 (0.19)
eGFR at 24 months ml/min/1.73m^2 (SD	59 (7.7)
Number of Patients at 36 months	8
Serum creatinine level at 36 months mg/dL(SD)	1.62 (0.50)
eGFR at 36 months ml/min/1.73m^2 (SD)	53.4 (15)

Tout au long de mon étude, j'ai utilisé deux formats de fichier : des fichiers VCF et des fichiers PLINK.

2) Fichiers au format VCF et données WES

Les données génétiques de type WES de ces cohortes sont disponibles dans des fichiers au format VCF. Un VCF (Variant Call Format[8]) est un fichier texte contenant les variations génétiques d'un ou plusieurs individus. Ce format de fichier se compose de deux parties : un en-tête et un corps de fichier.

- Un en-tête (header) qui fournit des lignes de données décrivant le corps du fichier débute par le symbole #. Elles contiennent des mots-clés relatifs au fichier (le format, la référence pour les données, la date de formation mais également une description des colonnes utilisées dans le corps du fichier).
- le corps du fichier est composé de huit colonnes obligatoires quel que soit le fichier ainsi qu'un nombre illimité de colonnes facultatives (permettant de regrouper des informations sur le format et sur d'autres informations sur l'échantillon ainsi que les données génomiques de chaque individu). Les huit colonnes obligatoires sont :

1 : CHROM Numéro du chromosome sur lequel se trouve le variant
2 : POS Position du variant sur le chromosome
3 : ID l'identifiant du variant qui peut être sa position sur la plaque illumina ou son rsID (numéro d'accès unique des SNP) ou si l'ID n'est pas connu, il est remplacé par un ".".
4 : REF Nucléotide de référence à la position donnée.
5 : ALT La liste des allèles alternatifs à cette position.
6 : QUAL Un score de qualité associé aux allèles donnés.
7 : FILTRE Filtre selon lequel la variation a été choisie.
8 : INFO Liste d'arguments pour décrire la variation.

Le séquençage d'exome entier (Whole Exome Sequencing) est une technique génomique permettant de séquencer toutes les régions de gènes codantes pour des protéines (les exons). Les exons sont au nombre de 180 000 chez l'Homme, représentent moins de 2% du génome, mais contiennent environ 85% des variants liés à une maladie [9]. Le but de cette approche est d'identifier les variants génétiques codants rares, dont la fréquence d'apparition dans le génome est inférieure à 1%, pouvant modifier les séquences protéiques associés à des maladies. Le séquençage uniquement des régions codantes du génome permet aux chercheurs de concentrer leurs ressources sur les gènes les plus susceptibles d'affecter le phénotype et il est ainsi possible, comme dans le cas d'une GWAS, de corréler la présence d'une variation avec un trait phénotypique. Le séquençage de l'ADN consiste à déterminer l'ordre d'enchaînement des nucléotides pour un fragment d'ADN donné [10].

3) Fichiers au format PLINK et données GWAS

Durant cette étude, des fichiers au format PLINK[11] ont été utilisés pour certaines étapes. PLINK est un ensemble d'outils d'analyse d'association du génome entier (GWAS), à source libre, conçu pour effectuer un ensemble d'analyses de base à grande échelle de manière efficace en termes de calcul. PLINK se concentre uniquement sur l'analyse des données de génotype / phénotype. PLINK est en constante évolution et il est en cours d'élaboration par Shaun Purcell au Centre pour la recherche en génétique humaine (CHGR), au Massachusetts General Hospital (MGH) et au Broad Institute of Harvard & MIT. Plink produit des fichiers qui lui sont propres. Les fichiers utilisés dans ce projet sont les fichiers Bed, Bim, Fam [12, 13].

Un fichier au format bed (signifiant Browser Extensible Format) est un fichier binaire représentant les variants biallélicques d'un génotype. Ce fichier doit être accompagné des fichiers bim et fam pour être exploité.

Un fichier bim est un fichier texte contenant les informations de chaque variant. Il accompagne une table de génotypes binaires contenue dans le fichier bed. Il s'agit d'un fichier sans en-tête et chaque ligne du fichier représente un variant. Chaque ligne est divisée en 6 colonnes séparées par une tabulation.

1 : Nom ou code du chromosome (numéro de 1 à 22 ou MT pour les chromosomes mitochondriaux et X/Y pour les chromosomes sexuels)
2 : Identifiant (ID ou rsID) du variant
3 : Distance génétique du variant en morgans ou centimorgans sur le chromosome
4 : Coordonnée de la paire de base du variant
5 : Allèle de référence
6 : Allèle alternatif

Un fichier fam est un fichier texte contenant les informations de chaque individu qui accompagne une table de génotypes binaires contenue dans le fichier bed. C'est un fichier sans en-tête

et chaque ligne correspond à un individu. Chaque ligne est composée de 6 champs séparés par une tabulation.
1 : identifiant de la famille ("FID")
2 : Identifiant intrafamilial ("IID")
3 : ID du père au sein de la famille ('0' si le père n'est pas dans l'ensemble de données)
4 : ID de la mère au sein de la famille ('0' si la mère n'est pas dans l'ensemble de données)
5 : Code de sexe ('1' = homme,'2' = femme, '0' = inconnu)
6 : Valeur du phénotype ('1' = témoin,'2' = cas)

Une étude d'association pangénomique (Genome Wide Association Study) est une analyse de nombreuses variations génétiques chez de nombreux individus (génotypage ADN), afin d'étudier leurs corrélations avec des traits phénotypiques. Ces études se concentrent sur les associations entre les polymorphismes nucléotidiques (SNP, variants génétiques avec une fréquence d'apparition dans le génome supérieur à 1%) et des phénotypes tels que les maladies humaines (analyse cas/témoin ou traits quantitatifs). Néanmoins, les GWAS ne donnent pas une relation de causalité du SNP à son caractère mais renseignent plutôt sur une région génomique corrélée avec ce trait [(14)].

Le génotypage vise à déterminer une variation génétique, à une position spécifique, pour un individu ou un groupe d'individus.

Étapes de modification des fichiers et les outils associés

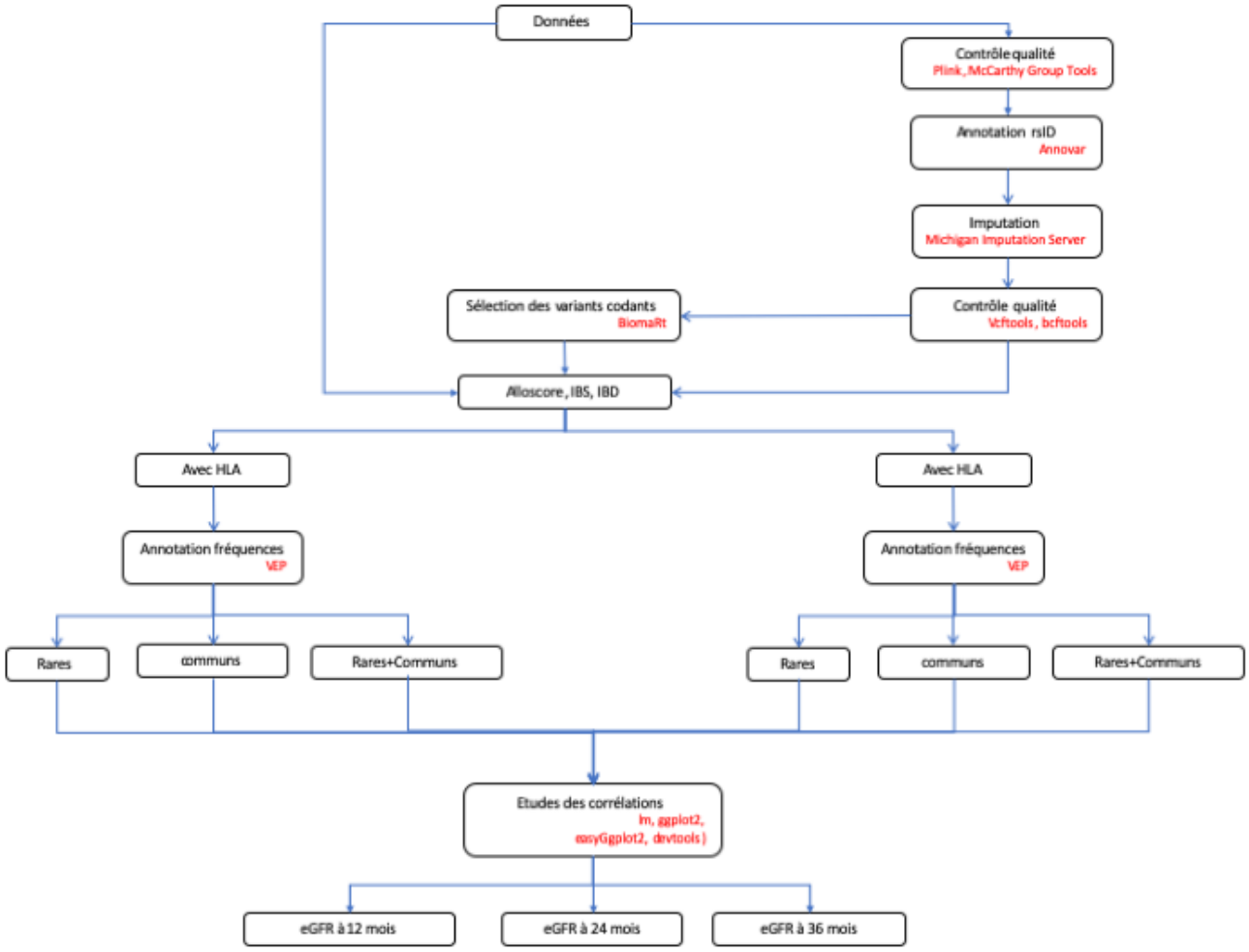


Figure 4 : Pipeline des étapes à réaliser et les outils associés.

1) Les contrôles qualité

Un contrôle qualité de pré-imputation a été réalisé dans un premier temps avec l'outil PLINK[12] des données génomiques et donc des variants mais aussi des individus. Il a été possible de supprimer les chromosomes extra-chromosomiques, mitochondriaux et sexuels, d'écarter les données manquantes sur les genotypes de notre fichier, de supprimer les variants non disponibles pour l'imputation, de supprimer les indels. Dans un second temps, les outils BCFtools[15], VCFtools[16], les pipelines de McCarthy Group Tools[17] et le script python checkVCF.py[18] ont été utilisés pour réaliser un nettoyage préliminaire spécifique des données non imputées, afin notamment d'aligner les SNPs génotypés avec le panel HRC.

Un contrôle qualité de post-imputation a été réalisé avec les outils vcftools et bcftools pour supprimer les variants imputés dont la prédiction serait fausse.

VCFtools (vcftools.sourceforge.net/) est un progiciel conçu pour travailler avec les fichiers VCF, tels que ceux générés par le projet 1000 Genomes. L'objectif de VCFtools est de fournir des méthodes facilement accessibles pour travailler avec des données complexes sur les variations génétiques sous forme de fichiers VCF. Cet ensemble d'outils peut être utilisé pour, filtrer les variantes spécifiques, comparer les fichiers, convertir en différents types de fichiers, valider et fusionner des fichiers, créer des sous-ensembles de variantes. VCFtools se compose d'un module perl et d'un exécutable binaire. Le module perl est une API Perl générale pour manipuler les fichiers VCF, alors que l'exécutable binaire fournit des routines d'analyses générales.

BCFtools (samtools.github.io/bcftools/bcftools.html) est un ensemble d'outils qui manipulent un fichier VCF et son équivalent binaire BCF. Toutes les commandes fonctionnent avec les VCF et les BCF, qu'ils soient compressés en BGZF ou non. Les VCF et BCF compressés et indexés sont universels pour tous les programmes. La compression est réalisable avec bcftoolsview.
Il permet entre autres d'éditer un fichier, ajouter et supprimer des annotations, concaténer des fichiers, les filtrer, les convertir, les indexer, réaliser des statistiques, …

Les pipelines de McCarthy Group Tools (well.ox.ac.uk/~wrayner/tools/) sont de petits programmes qui analysent le VCF pour que l'imputation soit possible.

Le script python checkVCF.py (genome.sph.umich.edu/wiki/CheckVCF.py) est un petit programme qui vérifie la conformité du VCF et de ses variants avant les tests d'association.

Tous ces outils ont été utilisés via Bash et python.

2) L'imputation

L'imputation (Figure 5) des SNPs est une technique statistique qui permet de compléter les génotypes manquants dans nos jeux de données génotypés à partir d'un panel de référence [19]. Les outils les plus utilisés pour l'imputation, sont les serveurs en ligne du ''SANGER Institute'' (situé à Cambridge au Royaume-Uni) [20] et celui de " Michigan Imputation Server" [21]. Leurs algorithmes utilisent des outils d'imputation de référence (tels que Shape-IT et Eagle) ainsi que plusieurs panels de référence dont The Haplotype Reference Consortium (HRC) [19] qui est le plus grand panel de référence disponible à ce jour. Ce panel comprend 38 820 individus (majoritairement d'origine européenne), 64 976 haplotypes et 39 235 157 SNPs.

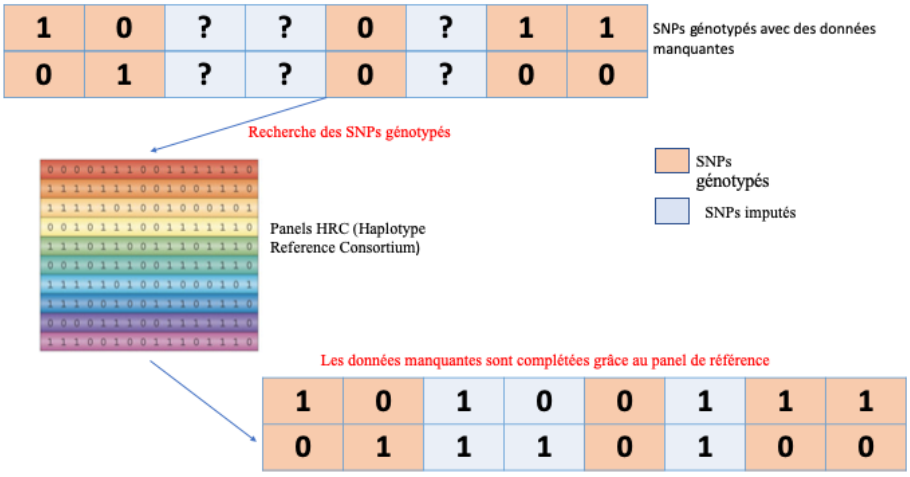


Figure 5 : Principe de l'imputation

Pour chaque SNP imputé, la fiabilité de la prédiction est déterminée par le calcul d'un score appelé le score de post-probabilité). Si ce score est égal à 0.8, il y a 20% de chance que la prédiction soit fausse. Le serveur nécessite que le fichier vcf de données à imputées soit séparé en plusieurs vcf, où chaque vcf correspond à un chromosome.

Michigan Imputation Server[22] fournit un service gratuit d'imputation de génotype utilisant Minimac4 (c'est la dernière version de la série de logiciels d'imputation de génotypes, qui utilise moins de mémoire pour une imputation de meilleure qualité). L'imputation est réalisée à partir de données de séquençage provenant du HRC (Haplotype Reference Consortium). Il réalisera un contrôle qualité complet sur le jeu de données. Le serveur d'imputation du Michigan utilise le modèle de programmation MapReduce pour une parallélisation efficace des tâches de calcul, ainsi les 22 chromosomes sont imputés simultanément.

3) Annotation

Une annotation de l'identifiant (rsID) pour chaque variant est réalisée par l'outil Annovar. Annovar est un outil logiciel java qui permet d'annoter fonctionnellement les variants génétiques détectés dans divers génomes (y compris les génomes humains hg18, hg19, hg38, ainsi que ceux des souris, vers, mouches, levures et autres). Il a pour base une liste de variants avec le chromosome, la position de départ, la position finale, le nucléotide de référence et les nucléotides observés. Annovar[23] permet différents types d'annotations dont l'annotation régionale (identification des variantes dans des régions génomiques spécifiques) qui est celle utilisée dans cette étude.

Une annotation des conséquences génétiques, des noms de gènes et de transcrits de chaque variant a été effectuée post-imputation à l'aide de l'outil BiomaRt[24] qui est un package de R. Cet outil a pour but d'interroger un grand nombre de bases de données et en particulier Ensembl. BioMart est devenu une solution générique pour l'intégration d'algorithmes et de données au travers d'une application simple et rapide à mettre en place. Il permet aux laboratoires de créer et de consulter leurs bases de données avec cet outil. BiomaRt est disponible en version web mais également grâce à un package R.

Enfin, une annotation de la fréquence de chaque variant imputé ou non a été réalisé grâce à l'outil VEP[25]. VEP est un outil du site Ensembl[26] (ensembl.org/index.html)qui détermine l'effet des variants qu'on lui propose (snp, insertion, délétion, …) sur les gènes, les transcrits et la séquence protéique, ainsi que sur les régions régulatrices. En entrant les coordonnées ou les identifiants rsID des variants, on retrouve les gènes et les transcrits affectés par ceux-ci, leur localisation, les conséquences qu'entraine cette variation. Les fréquences des allèles secondaires associées au projet « 1000 genomes » peuvent être retrouvées dans les quatre populations continentales, AFR (africain), AMR (américain), ASN (asiatique) et EUR (européen). VEP peut aussi fournir des identifiants

supplémentaires pour les gènes, les transcrits, les protéines et les variants (symbole de gène, version de la transcription, CCDS, Ensembl, Uniprot, ...).

Tous ces outils ont été utilisés à l'aide de Bash et Awk.

Les scores étudiés : l'Alloscore, l'IBD et l'IBS

1) L'Alloscore

Ce score se base sur le calcul du nombre d'acides aminés différents entre 2 individus. Chaque individu possède sa propre combinaison d'acides aminés. Dans le cas d'une greffe, il est intéressant de voir s'il y a un intérêt à comparer les compatibilités d'acides aminés (compatibilité non HLA) pour la survie du greffon.

Ce score provient de l'article de Mesnard et al. publié en 2016 [6]. Le script "Alloscore.py" qui permet son calcul nécessite :

- Le VCF à étudier
- Un fichier contenant la localisation des protéines codées par les gènes (secrétés, intracellulaires…) créé à partir du site web The human protein atlas [27] (proteinatlas.org/about) qui est un programme suédois lancé en 2003 dans le but de cartographier toutes les protéines humaines dans les cellules, les tissus et les organes en utilisant l'intégration de diverses technologies omiques.
- Un fichier (transmembranaire.txt) contenant le nom des transcrits transmembranaires créé grâce au site Ensembl et à l'outil Tmhmm (outils de prédiction des domaines transmembranaires) pour l'annotation des hélices transmembranaires.

Ce script permet tout d'abord de filtrer les données du vcf pour ne garder que les variants codants qui ont pour "consequence" stop_gained ou stop_lost ou initiator_codon ou inframe_insertion ou inframe_deletion ou missense_variant ou splice_donor_variant ou frameshift_variant ou splice_acceptor_variant. Un second filtre ne sert qu'à sélectionner les individus dont la profondeur de reads, pour le séquençage, est compris entre 10 et 500. Le script permet le calcul du score de mismatch entre le donneur et le receveur pour chaque acide aminé. La contribution du mismatch (Figure 6) entre acides aminés est différente si les patients étudiés sont homozygotes ou hétérozygotes :

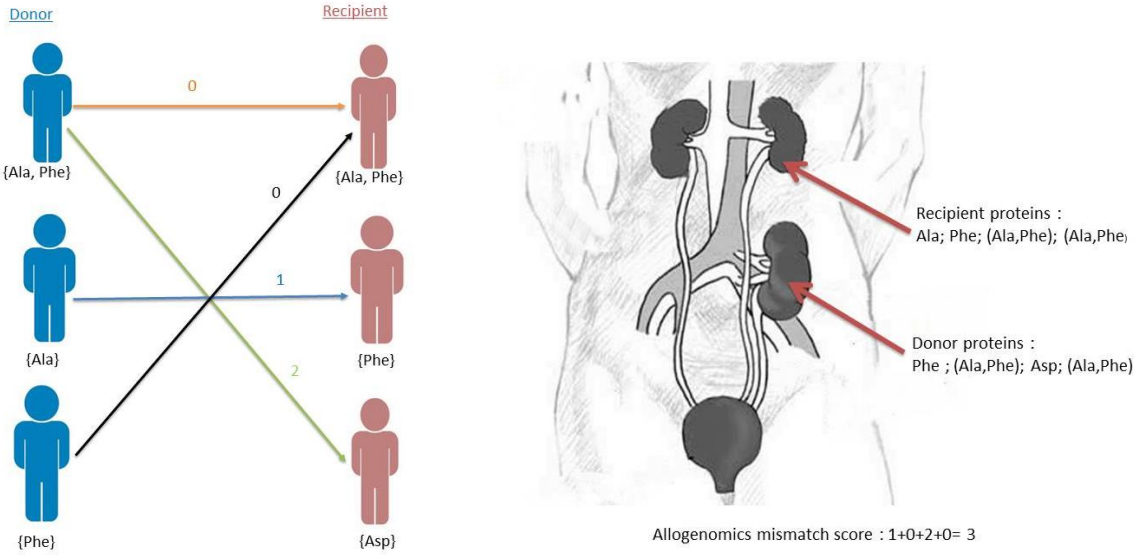


Figure 6 : Nombre d'incompatibilités pour des paires d'acides aminés [6]

- Si le donneur et le receveur sont homozygotes, et n'ont aucun élément en commun, la contribution est de 1 sinon de 0.
- Si le donneur est homozygote mais le receveur hétérozygote, et s'ils n'ont aucun élément en commun, la contribution du mismatch est de 0 sinon de 1.
- Si le donneur est hétérozygote mais le receveur homozygote, la contribution du mismatch est de 2 s'ils n'ont aucun élément en commun sinon de 1.

- Si le donneur et le receveur sont hétérozygote, et qu'ils n'ont aucun élément en commun, la contribution est de 2 ; sinon de 1 s'ils ont un élément en commun, sinon de 0 s'ils ont 2 éléments en commun.

L'Alloscore correspond donc à la somme des incompatibilités entre un donneur et un receveur. Ce score suggère que plus il est élevé, plus le donneur et le receveur sont différents, plus il y a de chance que le receveur possède un antigène immunogène ce qui augmente le risque de rejet. Un score faible permettrait de déterminer les meilleurs couples donneur-receveur.

Le code permet le calcul de plusieurs scores différents en comparant ces fichiers pour chaque variant selon leur localisation. Tout d'abord le score dit "global" qui correspond à l'Alloscore de tous les variants quelle que soit leur localisation cellulaire. Puis le score dit "secretion" qui correspond à l'Alloscore de tous les variants avec pour localisation cellulaire "secreted". Le score dit "Transmembrannaire protein atlas," correspond à l'Alloscore de tous les variants qui ont pour localisation cellulaire "membrane" dans The human protein athlas. Enfin, le score dit "Transmembrannaire Ensembl " qui correspond à l'Alloscore de tous les variants dont les protéines codées ont pour localisation cellulaire "membrane" et qui possède une ou plusieurs hélices transmembranaires selon l'outil de prédiction "Tmhmm".

2) Les scores IBD et IBS

L'identité par descendance (Identity by descent, IBD) est un score qui retranscrit la proportion de segment d'ADN identique partagé par deux personnes qui ont un ancêtre commun. Un segment est identique si tous les allèles d'un chromosome sont identiques. Grâce à ce score, il est possible d'identifier les individus appartenant à une même famille ainsi que leurs liens de parenté [28].

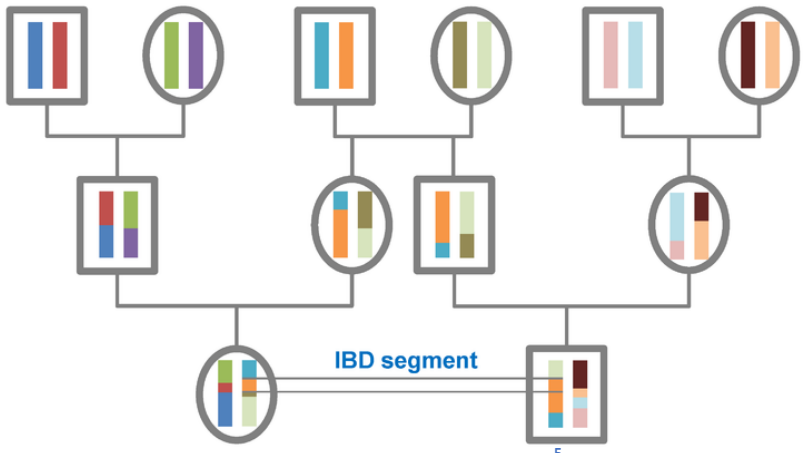


Figure 7 : Principe du score IBD [5]

Pour expliquer ce score, un pedigree (Figure 7) de 12 individus permet de montrer l'origine des segments d'ADN. Chaque case (homme) et chaque cercle (femme) représente un individu avec deux chromosomes homologues. Suite à des croisements, les individus de la rangée 2 héritent des chromosomes recombinés de leurs parents parmi les 3 couples à l'origine de ce pedigree. Les cousins du premier ordre (représentés au 3eme rang) partagent un segment IBD délimité par les lignes grises. Tous deux ont hérité de ce segment IBD du même individu, leur grand-père (chromosome de couleur orange) et ils partagent le même allèle en raison de leur descendance. Selon la génétique des populations, tous les individus ont eu une ascendance commune dans un passé lointain et nous avons tous en commun de vieux et courts segments d'ADN. En général, plus le segment est grand, plus la relation est étroite.

L'identité par État (Identity by State, IBS[(29)]) se base sur le même résonnement que le score IBD sauf que ce score permet de décrire le nombre de différences entre 2 individus qui ne partagent pas un ancêtre commun récent [6]. En d'autres termes, il fait référence au fait que deux individus, même s'ils ne sont pas liés l'un à l'autre, présentent le même allèle à un locus spécifique. En raison de leur manque de parenté, cette similitude est probablement due à un événement mutationnel similaire.

Les scores IBS et le score IBD sont calculés grâce à l'outil PLINK[(9)] et la création du fichier genome par la commande --genome[(30)]. Ce fichier se divise en colonnes distinctes :

- FID1 : Identifiant de la famille pour le premier individu de la combinaison
- IID1 : Identifiant du premier individu de la combinaison
- FID2 : Identifiant de la famille pour le second individu de la combinaison
- IID2 : Identifiant du second individu de la combinaison
- RT : Relation entre les deux individus (information présente dans le fichier fam et ped)
- EZ : Valeur pour vérifier qu'il ne s'agit pas de la même personne
- 3 colonnes Z0, Z1, Z2 : Permettent de déterminer le nombre d'hétérozygotes (deux allèles différents) et le nombre d'allèles homozygotes (deux allèles identiques) avec différentes probabilités de valeur de l'IBD, à Z0 : P(IBD=0), Z1: P(IBD=1) et Z2: P(IBD=2)
- PI_HAT : IBD
- PHE : Le code phénotype de la paire d'individus : si les deux allèles sont affectés la valeur retournée est 1, si l'un des allèles est affecté la valeur retournée est 0, sinon si aucun allèle n'est affecté la valeur retournée est -1.
- DST : IBS
- PPC : P-value pour le test binomial

Le score IBD est donc calculé à partir de la formule suite : PI_HAT = Z2 + 0.5*Z1 soit PI_HAT = P(IBD=2) + 0.5*P(IBD=1). Le score IBS est calculé par DST = (IBS2 + 0.5*IBS1)/ (IBS0 + IBS1 + IBS2) où IBS0, 1, 2 sont les scores calculés selon si les individus ont 0, 1 ou 2 allèles en commun pour chaque variant.

Développement méthodologique (voir Figure 4)

1) Contrôle qualité pré-imputation

Avant de réaliser l'imputation des données, il est nécessaire de réaliser un contrôle qualité des données génomiques et donc des variants mais aussi des individus.

Contrôle qualité sur les SNPs :

- Suppression des chromosomes extra-chromosomiques, mitochondriaux et sexuels : Seuls les chromosomes autosomes sont pris en compte dans les analyses génomiques ainsi que pour l'imputation de SNPs. C'est pourquoi nous avons enlevé 42 784 SNPs extra-chromosomiques (GL000), 543 SNPs mitochondriaux et 475 812 SNPs sexuels (457 931 X et 17 881 Y).
- Vérification des données manquantes sur les génotypes : Les SNPs ayant un pourcentage élevé de données manquantes peuvent indiquer un biais technique. Nous avons donc exclu tous les SNPs avec plus de 2% de données manquantes : 185 689 (option geno) SNPs ont ainsi été exclus.
- Test de fréquence de l'Allèle Mineur (MAF) : Suppression des variants absents ou très rares dans le HRC et qui ne peuvent pas être imputés. Donc, 14 370 886 SNPs avec une MAF<0,01% (SNPs monomorphiques) ont été exclus.
- Test de l'Equilibre d'Hardy-Weinberg (HWE) : En cas de stratification de la population ou d'erreurs systématiques de génotypage, il est possible d'observer une déviation de l'HWE. Pour cela, on utilise l'option hwe de PLINK pour supprimer les SNPs qui ont une p-value inférieure à une p-value seuil (ici 0,001). Il a donc été possible de supprimer 60 223 SNPs.
- Exclusion des indels : Les indels ne pouvant pas être imputés, 23 071 SNPs représentants des indels ont été exclus de l'étude.

Au final, 334 993 SNPs ont passé le contrôle qualité sur les 15 494 001 SNPs du fichier de départ.

Les données de départ ont été difficiles à contrôler et cela a pris beaucoup de temps. Il a été nécessaire de tester différents seuils pour chaque filtre afin de réussir à filtrer les variants en évitant d'en effacer trop.

2) Annotation des rsID des SNPs

Nous avons tout d'abord essayé d'annoter les SNPs avec l'outil GATK, mais il ne ressortait qu'une centaine de SNPs annotés. Nous avons donc décidé d'utiliser un autre outil qui est Annovar. Nous avons mis un certain temps pour comprendre comment fonctionne cet outil, les divers scripts à utiliser

et les différentes manipulations à exécuter. Cependant avec Annovar, beaucoup plus de SNPs ont été annotés. Les SNPs de nos jeux de données n'ont pas d'identifiant. Or, pour le fichier de référence 1 000 Genomes et pour utiliser la plupart des logiciels d'analyse, les SNPs sont désignés par un identifiant universel international : le rsID (par exemple rs1234). Il faut tout d'abord attribuer un numéro d'identification factice pour chaque SNP (par exemple, une série d'identifiant de 1 à 334 993). Il faut ensuite utiliser l'outil Annovar pour donner à chaque SNP son véritable identifiant rsID. Annovar permet l'annotation d'rsID à partir d'un fichier de référence qui contient les informations de position de début et de fin sur les chromosomes ainsi que l'allèle de référence.

Cette manipulation de fichiers demande beaucoup de ressources de calcul. Le fichier de référence étant un fichier de plus de 10 GigaBases, il faut donc à ce stade utiliser le serveur de calcul BirdCluster[31] (serveur de la plateforme de bioinformatique de Nantes (BiRD, pf-bird.univ-nantes.fr/ressources/cluster-de-calcul/birdcluster-1313050.kjsp) composé de nombreux outils dédiés à la bioinformatique, plus particulièrement aux analyses de données NGS). A partir de la version snp138 du génome de référence Hg19, 290 329 SNPs sur les 334 993 SNPs récupérés précédemment, ont pu être annotés et se voir attribuer un rsID.

3) Imputation et contrôle qualité post-imputation

Après ce contrôle qualité et cette annotation, il est possible d'imputer des SNPs additionnels afin d'augmenter la couverture d'étude du génome et la densité de marqueurs. Pour l'imputation, nous avons utilisé le serveur Sanger mais celui-ci nous ressortait une erreur récurrente sur la probabilité de distribution de génotype qui ne correspond pas à celle du panel de référence (en précisant que cela viendrait du filtre HWE ou de l'allèle de référence qui est incorrect). C'est pourquoi, tout d'abord, nous avons lancé l'imputation avec différents filtres HWE mais toujours la même erreur. Nous avons donc ensuite testé l'hypothèse de l'allèle de référence par le biais de l'outil BCFtool[15] qui ne trouve aucun problème avec nos allèles de référence mais l'imputation nous renvoie toujours la même erreur. Ne trouvant pas d'où peut provenir l'erreur nous avons décidé d'envoyer un mail au server Sanger[20] pour de plus amples explications. En attendant leur réponse, nous avons choisi d'utiliser le serveur d'imputation Michigan Server[21] pour pouvoir continuer nos recherches et nos réflexions. L'imputation a permis d'obtenir 47 073 495 SNPs dont 10 440 601 SNPs annotés avec une post-probabilité supérieure à 0,8 alors que seul 290 329 SNPs annotés ont passé le contrôle. Ci-dessous un schéma récapitulatif du nombre de SNPs pour chaque étape (Figure 8) :

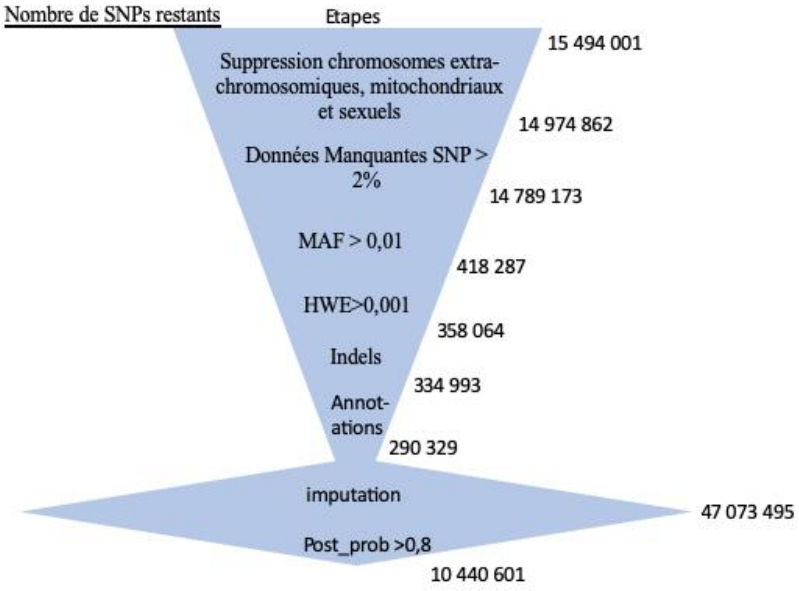


Figure 8 : Nombre restant de SNPs pour chaque étape réalisée

4) Annotation des conséquences génétiques, des noms de gènes et de transcrits

Le fichier sorti de l'imputation possède une colonne INFO qui contient les informations statistiques liées à l'imputation (Figure 9). Or pour pouvoir appliquer le score à nos données imputées, il est essentiel d'avoir pour chaque variant le nom du gène, des transcrits et leur localisation qui lui correspondent.

```
#CHROM  POS ID  REF ALT QUAL    FILTER  INFO    FORMAT  AELTXQC-alloval3-11-P6-Re-6 AVEIWCJ-alloval3-24-P12-Do-12   AWCMWEX-alloval3-19-P10-Re-10   BADARZN-
1   865625  rs146327803 G   A   .   PASS    "RefPanelAF=0.000338774;AN=80;AC=0;INFO=1"  GT  0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0
1   865662  rs140751899 G   A   .   PASS    "RefPanelAF=0.000292578;AN=80;AC=0;INFO=1"  GT  0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0
1   865700  rs116730894 C   T   .   PASS    "RefPanelAF=0.000323375;AN=80;AC=0;INFO=1"  GT  0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0
1   866438  rs144010167 G   A   .   PASS    "RefPanelAF=0.000277179;AN=80;AC=0;INFO=1"  GT  0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0
1   874456  rs149677938 G   A   .   PASS    "RefPanelAF=7.69941e-05;AN=80;AC=0;INFO=1"  GT  0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0
1   877523  rs200195897 C   G   .   PASS    "RefPanelAF=0.00052356;AN=80;AC=0;INFO=1"   GT  0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0
1   877556  rs201313326 G   A   .   PASS    "RefPanelAF=0.000169387;AN=80;AC=0;INFO=1"  GT  0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0
1   878331  rs148327885 C   T   .   PASS    "TYPED;RefPanelAF=0.0110102;AN=80;AC=1;INFO=1"  GT  0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0
1   878653  rs112422281 A   G   .   PASS    "RefPanelAF=9.2393e-05;AN=80;AC=0;INFO=1"   GT  0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0
1   879142  rs148506094 C   G   .   PASS    "RefPanelAF=9.2393e-05;AN=80;AC=0;INFO=1"   GT  0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0
1   879409  rs200048444 C   G   .   PASS    "RefPanelAF=0.00103172;AN=80;AC=0;INFO=1"   GT  0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0
1   880942  rs112163984 A   G   .   PASS    "RefPanelAF=0.000107792;AN=80;AC=0;INFO=1"  GT  0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0
1   880943  rs113650656 G   A   .   PASS    "RefPanelAF=9.2393e-05;AN=80;AC=0;INFO=1"   GT  0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0
1   881602  rs148962722 C   T   .   PASS    "RefPanelAF=0.000123191;AN=80;AC=0;INFO=1"  GT  0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0
1   881918  rs35471880  G   A   .   PASS    "TYPED;RefPanelAF=0.0513705;AN=80;AC=2;INFO=1"  GT  0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|1 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0
1   883883  rs143278708 T   C   .   PASS    "RefPanelAF=0.000384971;AN=80;AC=0;INFO=1"  GT  0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0
1   887812  rs142181102 C   T   .   PASS    "RefPanelAF=7.69941e-05;AN=80;AC=0;INFO=1"  GT  0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0
1   888563  rs150127608 C   T   .   PASS    "RefPanelAF=0.000153988;AN=80;AC=0;INFO=1"  GT  0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0
1   888659  rs3748597   T   C   .   PASS    "TYPED;RefPanelAF=0.946212;AN=80;AC=75;INFO=1"  GT  1|1 0|1 1|1 1|1 1|1 1|1 1|1 0|1 1|1 1|1 1|1 1|1 1|1 1|1 1|1 1|1 1|1 1|1 1|1 1|1 1|1 1|1 1|1 1|1 1|1 1|1
1   889238  rs3828049   G   A   .   PASS    "TYPED;RefPanelAF=0.0515707;AN=80;AC=3;INFO=1"  GT  0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 1|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0
1   889451  rs150111944 G   A   .   PASS    "RefPanelAF=7.69941e-05;AN=80;AC=0;INFO=1"  GT  0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0
1   891309  rs148852075 C   T   .   PASS    "RefPanelAF=7.69941e-05;AN=80;AC=0;INFO=1"  GT  0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0
```

Figure 9 : Fichier VCF avant la recherche de gène et de transcrits

La récupération de ces données a été réalisée avec l'outil BioMart[24] de Ensembl[26] via le langage R. Pour utiliser BioMart[24], nous avons choisi d'utiliser les données présentes dans la base de données "ENSEMBL_MART_SNP" qui permet de retrouver les SNPs en précisant l'espèce qui est "hsapiens_snp". Pour cela, il faut donner la liste des rsID que l'on souhaite mais l'outil a une limite de 500 rsID par requête. Pour nos 10 440 601 variants, nous avons tout d'abord découpé notre vcf en plusieurs sous fichiers contenant chacun les rsIDs correspondants à 1 chromosome. Puis, chaque sous fichier a été divisé en de nombreux sous fichiers contenant tous 500 rsID. Grâce à une boucle, il a été possible de lancer plusieurs requêtes les unes après les autres sans avoir à les lancer manuellement (ce qui a été un gain de temps considérable même si cette étape a pris plusieurs jours à être réalisée). Ensuite, nous avons utilisé des scripts AWK qui nous ont permis de reformater le VCF[8] imputé (Figure 10). Une adaptation du script a également été nécessaire.

```
##fileformat=VCFv4.1
##Goby=development (20150227145140)
##INFO=<ID=BIOMART_COORDS,Number=1,Type=String,Description="Coordinates for use with Biomart.">
##INFO=<ID=INDEL,Number=1,Type=Flag,Description="Indicates that the variation is an indel.">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=BC,Number=5,Type=String,Description="Base counts in format A=?;T=?;C=?;G=?;N=?.">
##FORMAT=<ID=GB,Number=1,Type=String,Description="Number of bases that pass base filters in this sample, or ignore string.">
##FORMAT=<ID=FB,Number=1,Type=String,Description="Number of bases that failed base filters in this sample, or ignore string.">
##FORMAT=<ID=Zygosity,Number=1,Type=String,Description="Zygosity">
##INFO=<ID=VariantEffectPrediction,Number=1,Type=String,Description="Variant Effect Predictions">
##source_20150128.1=vcf-annotate(r797) -a output.tsv.gz -d /scratchLocal/gobyweb/tmp/395843.1.rascals.q/attributes.lst -c CHROM,FROM,TO,INFO/VariantEffectPrediction
##INFO=<ID=GENE,Number=1,Type=String,Description="Ensembl gene identifier">
##INFO=<ID=GENE_NAME,Number=1,Type=String,Description="Gene name">
##source_20150128.2=vcf-annotate(r797) -a /scratchLocal/gobyweb/ARTIFACT_REPOSITORY-dev/artifacts/ENSEMBL_ANNOTATIONS/ANNOTATIONS/2.1.3/HOMO_SAPIENS/1000GENOMES/37/ref-start-end-gene-hgnc-sorted.tsv.gz -
#CHROM  POS ID  REF ALT QUAL    FILTER  INFO    FORMAT  AELTXQC-alloval3-11-P6-Re-6 AVEIWCJ-alloval3-24-P12-Do-12   AWCMWEX-alloval3-19-P10-Re-10   BADARZN-exome-S8-Recipient-ACR  BDWIERD-alloval3-36-P18
1   865625  rs146327803 G   A   .   PASS    ENSG00000187634;ENST00000420190|ENST00000437963|ENST00000342066|ENST00000618181|ENST00000622503|ENST00000618323|ENST00000616016|ENST00000618779|ENST00000616125
1   865662  rs140751899 G   A   .   PASS    ENSG00000187634;ENST00000420190|ENST00000437963|ENST00000342066|ENST00000618181|ENST00000622503|ENST00000618323|ENST00000616016|ENST00000618779|ENST00000616125
1   865700  rs116730894 C   T   .   PASS    ENSG00000187634;ENST00000420190|ENST00000437963|ENST00000342066|ENST00000341065|ENST00000618181|ENST00000622503|ENST00000618323|ENST00000616016|ENST00000618779|ENST00000616125
1   866438  rs144010167 G   A   .   PASS    ENSG00000187634;ENST00000420190|ENST00000437963|ENST00000342066|ENST00000341065|ENST00000622503|ENST00000618323|ENST00000616016|ENST00000618779|ENST00000616125
1   874456  rs149677938 G   A   .   PASS    ENSG00000187634;ENST00000420190|ENST00000342066|ENST00000341065|ENST00000618181|ENST00000622503|ENST00000616016|ENST00000618779|ENST00000616125|ENST00000617307
1   877523  rs200195897 C   G   .   PASS    ENSG00000187634;ENST00000342066|ENST00000341065|ENST00000455979|ENST00000622503|ENST00000618323|ENST00000616016|ENST00000618779 GT  C   C   C   C   C   C
1   877556  rs201313326 G   A   .   PASS    ENSG00000187634;ENST00000342066|ENST00000341065|ENST00000455979|ENST00000622503|ENST00000618323|ENST00000616016|ENST00000618779 GT  G   G   G   G   G   G
1   878331  rs148327885 C   T   .   PASS    ENSG00000187634;ENST00000342066|ENST00000341065|ENST00000455979|ENST00000618181|ENST00000622503|ENST00000618779|ENST00000616125|ENST00000617307 GT  C   C   C
1   878653  rs112422281 A   G   .   PASS    ENSG00000187634;ENST00000342066|ENST00000341065|ENST00000455979|ENST00000618181|ENST00000622503|ENST00000618779|ENST00000616125|ENST00000617307 GT  A   A   A
1   879142  rs148506094 C   G   .   PASS    ENSG00000187634;ENST00000342066|ENST00000341065|ENST00000455979|ENST00000618181|ENST00000622503|ENST00000618779|ENST00000616125|ENST00000617307 GT  C   C   C
1   879409  rs200048444 C   G   .   PASS    ENSG00000187634;ENST00000342066|ENST00000341065|ENST00000455979|ENST00000618181|ENST00000622503|ENST00000618779|ENST00000616125|ENST00000617307 GT  C   C   C
1   880942  rs112163984 A   G   .   PASS    ENSG00000188976;ENST00000327044 GT  A   A   A   A   A   A   A   A   A   A   A   A   A   A   A   A   A   A   A   A   A   A   A   A   A   A   A   A   A
1   880943  rs113650656 G   A   .   PASS    ENSG00000188976;ENST00000327044 GT  G   G   G   G   G   G   G   G   G   G   G   G   G   G   G   G   G   G   G   G   G   G   G   G   G   G   G   G   G
1   881602  rs148962722 C   T   .   PASS    ENSG00000188976;ENST00000327044 GT  C   C   C   C   C   C   C   C   C   C   C   C   C   C   C   C   C   C   C   C   C   C   C   C   C   C   C   C   C
1   881918  rs35471880  G   A   .   PASS    ENSG00000188976;ENST00000327044 GT  G   G   G   G   G   G   G   G|A G   G   G   G   G   G   G   G|A G   G   G   G   G   G   G   G   G   G   G   G   G
1   883883  rs143278708 T   C   .   PASS    ENSG00000188976;ENST00000327044 GT  T   T   T   T   T   T   T   T   T   T   T   T   T   T   T   T   T   T   T   T   T   T   T   T   T   T   T   T   T
```

Figure 10 : Fichier VCF après la recherche de gène et de transcrits et l'exécution de scripts awk

5) Sélection des variants codants et séparation des variants selon la fréquence

Les variants codants ont été sélectionnés après l'annotation des conséquences de chaque variant selon s'ils sont stop_gained ou stop_lost ou initiator_codon ou inframe_insertion ou inframe_deletion ou missense_variant ou splice_donor_variant ou frameshift_variant ou splice_acceptor_variant.

La séparation des variants selon leur fréquence dans le projet « 1000 genomes » s'est réalisée avec l'outil VEP[22] selon une fréquence de 0,01 dans la population européenne (EUR).

6) Calcul de l'Alloscore sur nos diverses populations

Après la formation des divers fichiers de données et l'adaptation du script de l'Alloscore, il a été possible de calculer ce score sur :

- Les variants natifs non imputés avec une séparation entre les rares, les communs, les rares+communs avec et sans HLA.
- Les variants codants imputés avec une séparation entre les rares, les communs, les rares+communs avec et sans HLA.
- Les variants imputés avec une séparation entre les rares, les communs, les rares+communs avec et sans HLA.

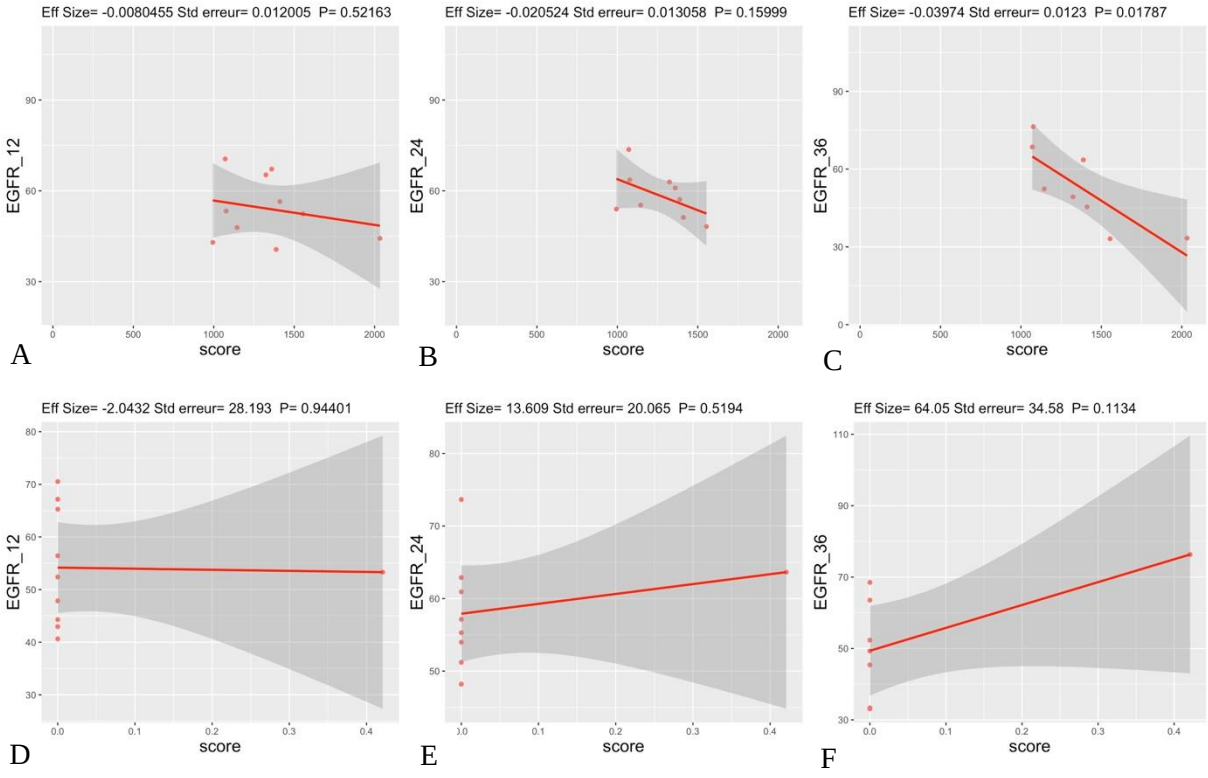
Il a été aussi possible de comparer l'Alloscore aux scores IBD et IBS et récupérer les valeurs de PI_HAT et de DST.

Résultats obtenus : corrélation de l'Alloscore avec le taux d'GFR

Dans un premier temps, les corrélations entre les scores calculés et les taux d'eGFR, à 12, 24, 36 mois post-greffe ont, été réalisées avec les modèles de régressions linéaires simples suivants : lm(EGFR_12~score), lm(EGFR_24~score), lm(EGFR_36~score) où seul le score expliquerait le taux d'eGFR. Cependant, les résultats étant peu informatifs, ces corrélations ont donc été réalisées grâce à des modèles de régressions linéaires multiples proposés par le Pr. Laurent Mesnard dans son article. Ces nouveaux modèles d'étude de la corrélation entre les scores et les taux d'eGFR prennent en compte l'âge du donneur ainsi que le score qui expliqueraient le taux d'eGFR : lm(EGFR_12~score+Age_donneur), lm(EGFR_24~score+Age_donneur), lm(EGFR_36~score+Age_donneur).

1) Comparaison aux scores IBD et IBS

La définition de l'Alloscore est proche de celle de l'IBS. C'est pourquoi il est intéressant de comparer l'Alloscore de la cohorte Discovery à l'IBS ainsi qu'à l'IBD de cette même cohorte. Ces scores ont été calculés grâce à l'outil PLINK et récupérés par la variable PHI_HAT (pour IBD) et DST (pour IBS).



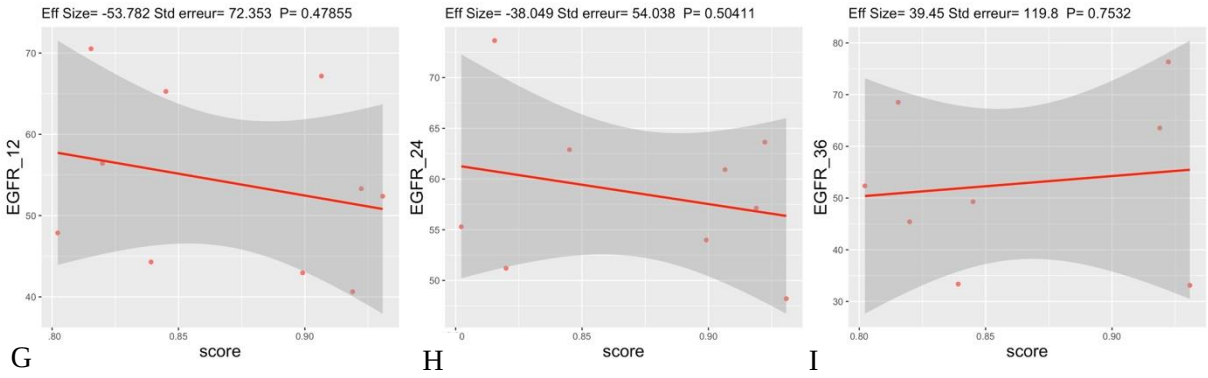


Figure 11 : Corrélation pour la population Discovery du score de l'article (A, B, C), de l'IBD (D, E, F) et de l'IBS (G, H, I) avec les taux d'eGFR à 12, 24, 36 mois post greffe : il n'y a pas de corrélation entre l'Alloscore, l'IBD et l'IBS quel que soit le mois

L'IBD et l'IBS ont pu être calculés sur les données non imputées avec les rares ou communs, rares+communs avec et sans HLA. Cependant, ces scores ne sont pas informatifs (3 paires de donneur-receveur semblent être identiques (IBS>0.8) ce qui se traduit par le fait que les deux individus de la paire seraient des jumeaux ce qui paraît étrange) et les données non imputées ne semblent pas adaptées au calcul de l'IBS et de l'IBD.

Cette observation se vérifie lorsque l'on recherche la corrélation qui peut exister entre l'Alloscore de la cohorte Discovery et les scores IBS, IBD correspondants.

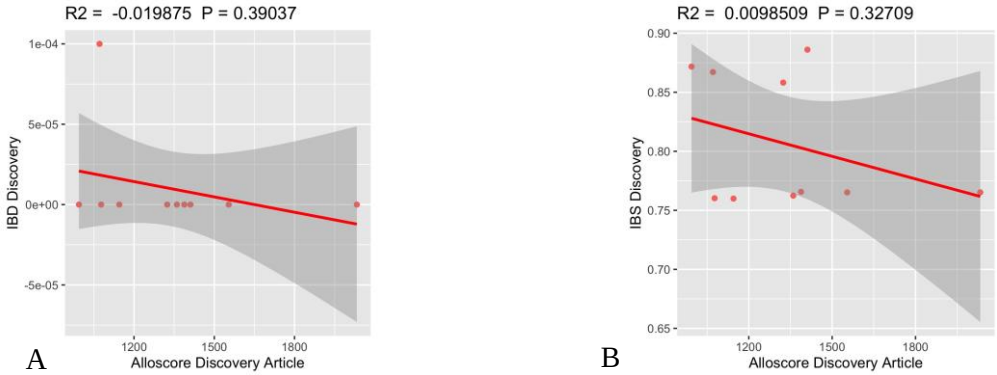
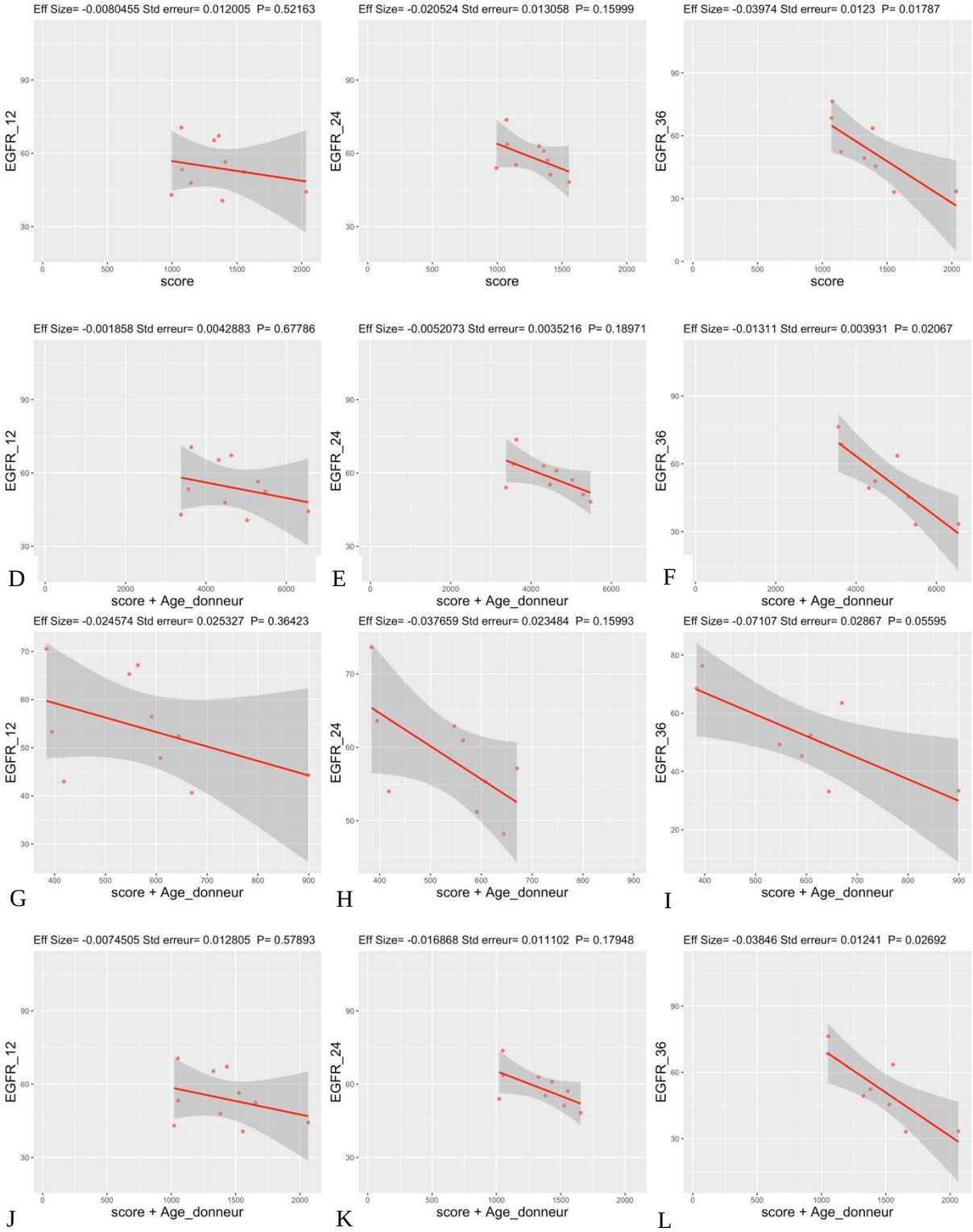


Figure 12 : Corrélation entre l'Alloscore, l'IBD (A) et l'IBS (B) : ces corrélations confirment qu'il n'y a pas de corrélation entre l'Alloscore, l'IBD et l'IBS

Il n'existe pas de corrélation entre l'Alloscore, l'IBD et l'IBS (avec un modèle mieux adapté à nos données au vu du faible R2). Plusieurs hypothèses peuvent être émises sur le fait que ces deux scores (pourtant utilisés comme référence) ne sont pas informatifs dans notre cas. On peut se demander si nos cohortes ont un nombre suffisant de paires de donneur-receveur et si nos données sont adaptées aux calculs de ces scores car PLINK ne permet pas de calculer l'IBS et l'IBD sur des variants liés entre eux avec des déséquilibres de liaison. Il serait donc intéressant de contrôler et de recalculer ces deux scores sur nos données.

2) Détermination de l'Alloscore parmi les quatre scores calculés

La cohorte Discovery (composée de 10 paires de donneur-receveur) est la population de référence pour l'établissement de l'Alloscore. Grâce à l'exécution du script pour les calculs des scores pour la cohorte Discovery sur les données non imputées et par comparaison aux résultats proposés dans l'article, il sera possible de déterminer lequel des quatre scores calculés par le script (score global, score secretion, score transmembranaire ensembl et score transmembranaire protein atlas) correspond à l'Alloscore représenté dans l'article.


Eff Size= -0.0080455 Std erreur= 0.012005 P= 0.52163
Eff Size= -0.020524 Std erreur= 0.013058 P= 0.15999
Eff Size= -0.03974 Std erreur= 0.0123 P= 0.01787
EGFR_12
EGFR_24
EGFR_36
score
Eff Size= -0.001858 Std erreur= 0.0042883 P= 0.67786
Eff Size= -0.0052073 Std erreur= 0.0035216 P= 0.18971
Eff Size= -0.01311 Std erreur= 0.003931 P= 0.02067
D
E
F
score + Age_donneur
Eff Size= -0.024574 Std erreur= 0.025327 P= 0.36423
Eff Size= -0.037659 Std erreur= 0.023484 P= 0.15993
Eff Size= -0.07107 Std erreur= 0.02867 P= 0.05595
G
H
I
Eff Size= -0.0074505 Std erreur= 0.012805 P= 0.57893
Eff Size= -0.016868 Std erreur= 0.011102 P= 0.17948
Eff Size= -0.03846 Std erreur= 0.01241 P= 0.02692
J
K
L

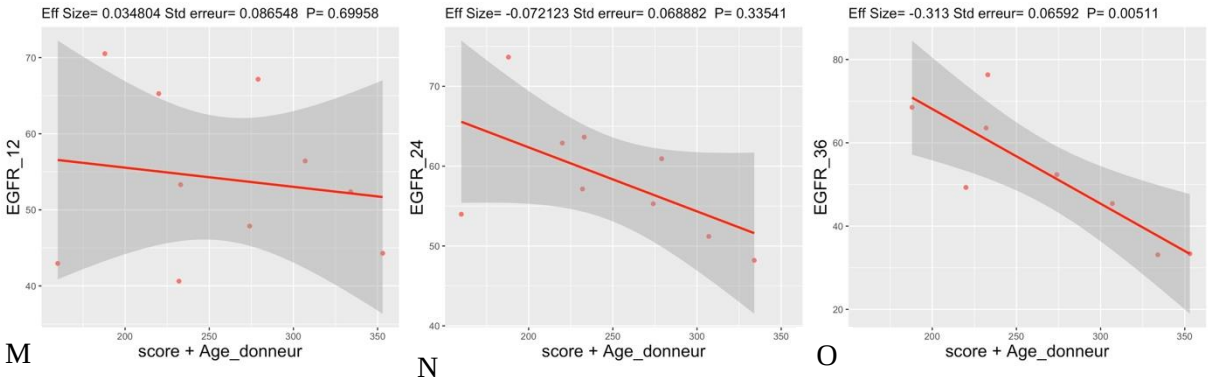


Figure 13 : Corrélation pour la population Discovery du score de l'article (A, B, C), du score global (D, E, F), du score secretion (G, H, I), du score transmembranaire Ensembl (J, K, L) et du score transmembranaire protein atlas avec les taux d'eGFR à 12, 24, 36 mois post greffe : le score transmembranaire Ensembl représente l'Alloscore présenté dans l'article. (Eff Size = Effect Size, Std erreur=standard d'erreur, P=P-value)

En comparant ces premiers résultats au niveau de la courbe et de l'échelle utilisée, il est possible de dire que l'Alloscore définit par le Pr. Laurent Mesnard (Size Effect de -0.0397 à 36 mois) correspond au score transmembranaire Ensembl (Size Effect de -0.03845 à 36 mois) et que le taux d'eGFR diminue de 0.039 mL/min/1,73m^2 par unité de score à 36 mois. Cependant, le score global ressemble fortement aux résultats de l'article ce qui montre que la localisation des variants n'aurait pas d'effet sur le calcul de l'Alloscore. Le reste de cette étude se concentrera donc sur le score transmembranaire Ensembl ainsi que sur les trois cohortes en notre possession.

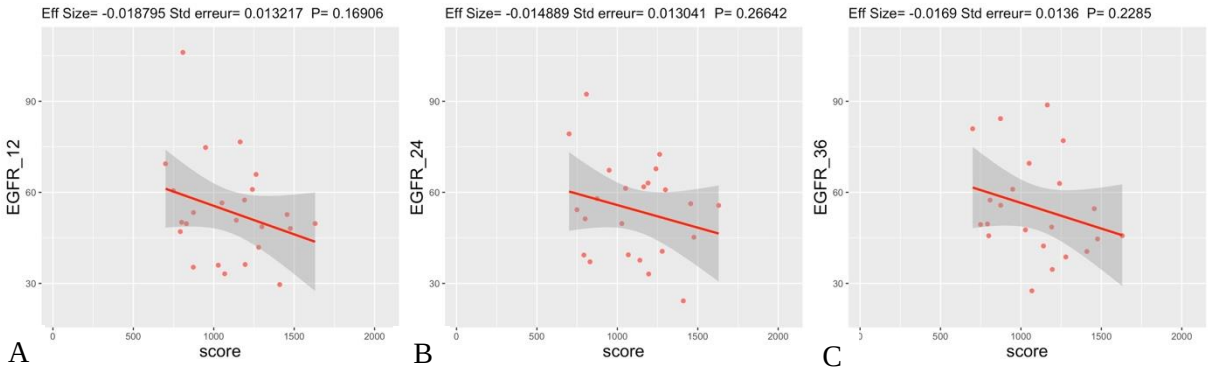
3) Réalisation de l'Alloscore sur les données non imputées

a. Population Discovery

Le score transmembranaire Ensembl a été calculé pour la population Discovery (10 paires de donneur-receveur) sur un fichier de plus de 15 millions de SNPs. Ce score a été calculé au graphique J, K, L de la figure 11 et comparé à ceux obtenus aux graphiques A, B, C de la figure 11. Pour la cohorte Discovery, on retrouve les mêmes courbes que pour le score de l'article, néanmoins, on ne retrouve pas le même coefficient Effect Size ni la même p-value. Cette différence de coefficients peut être due aux données, au script qui ne serait pas identique à celui utilisé dans l'article ou au choix de la régression pour représenter les corrélations.

b. Population Cornell

Le score transmembranaire Ensembl a été calculé pour la population Cornell (Cornell Validation Cohort (New York Medical Center), composée de 24 paires de donneur-receveur) sur un fichier de plus de 15 millions de SNPs.



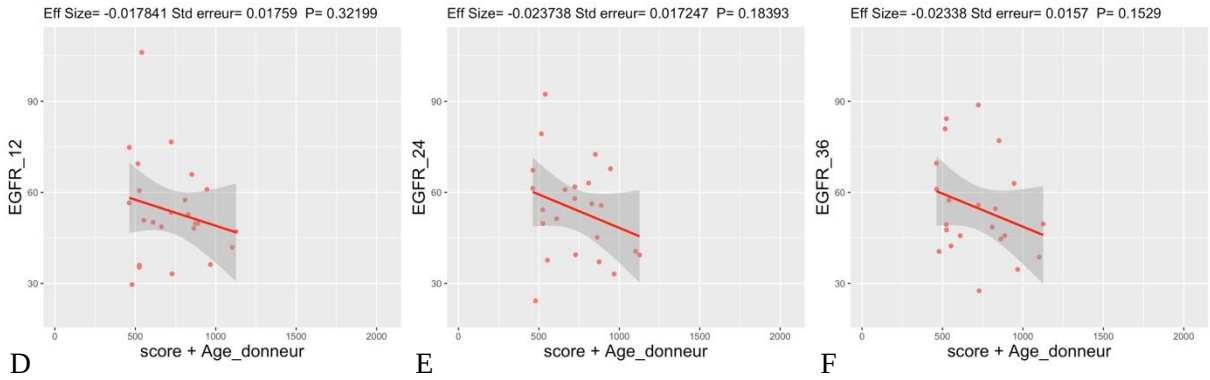


Figure 14 : Corrélation pour la population Cornell du score de l'article (A, B, C) et du score transmembranaire Ensembl (D, E, F) avec les taux d'eGFR à 12, 24, 36 mois post greffe : la tendance de la courbe est retrouvée. (Eff Size = Effect Size, Std erreur=standard d'erreur, P=P-value)

On retrouve les mêmes courbes que pour le score de l'article, néanmoins, on ne retrouve pas le même coefficient d'Effect Size (-0.0169 pour l'article contre -0.02338 pour le script à 36 mois) ni la mêmes p-value lors de la comparaison des données de référence avec une exécution du script. En observant les échelles, on remarque que les données du Pr. Laurent Mesnard ont une amplitude plus importante comparé à l'exécution du script ce qui peut s'expliquer par un biais dans le choix du modèle de régression. Néanmoins, on peut dire que dans cette population, le taux d'eGFR diminue en moyenne de 0.02 mL/min/1,73m^2 par unité de score à 36 mois.

c. Population Paris

Le score transmembranaire Ensembl a été calculé pour la population Paris (French Validation Cohort, composée de 19 paires de donneur-receveur) sur un fichier de plus de 15 millions de SNPs.

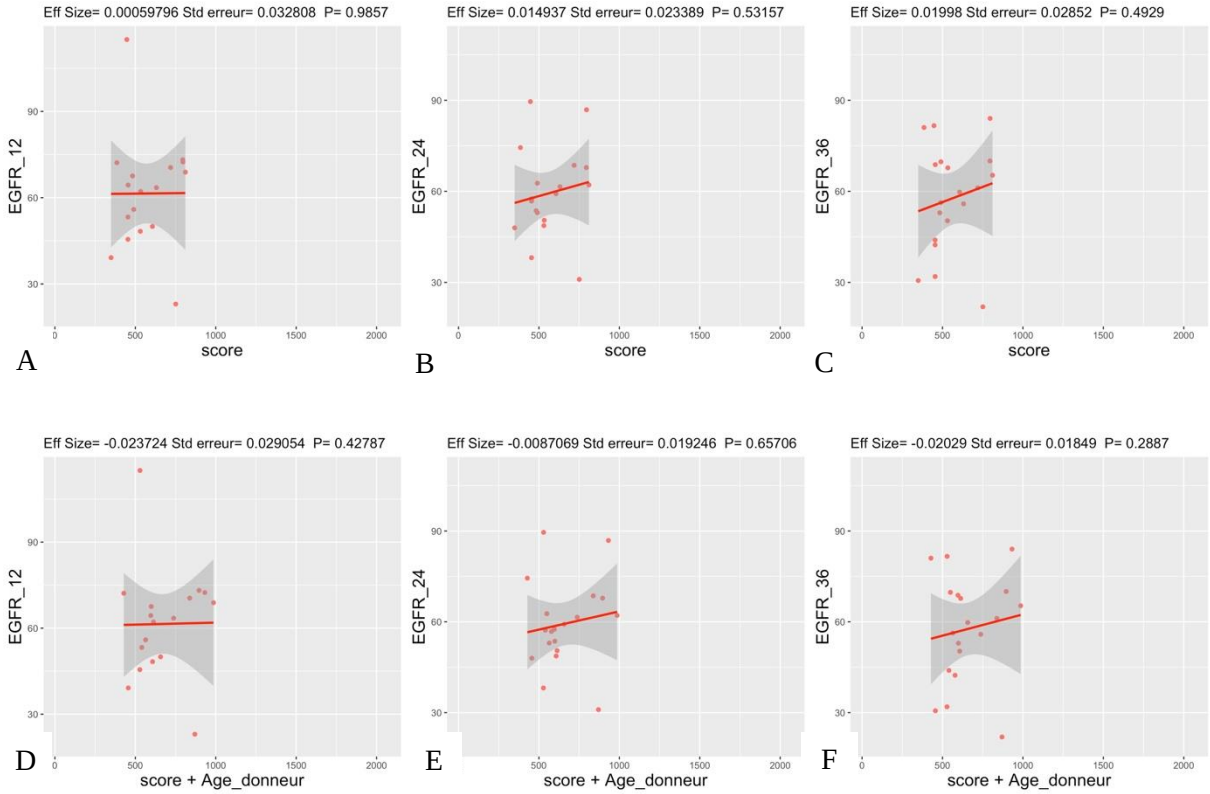


Figure 15 : Corrélation pour la population Paris du score de l'article (A, B, C) et du score transmembranaire Ensembl (A, B, C) avec les taux d'eGFR à 12, 24, 36 mois post greffe : la tendance de la courbe est retrouvée. (Eff Size = Effect Size, Std erreur=standard d'erreur, P=P-value)

Pour la cohorte Paris, on retrouve les mêmes courbes ainsi qu'un coefficient Effect Size similaires mais une p-value différente, que ce soit avec les données de référence ou après une nouvelle exécution du script. Le taux d'eGFR à 36 mois augmente de 0.02 mL/min/1,73m^2 par unité de score contrairement à la population Cornell. La petite différence pour le coefficient d'Effect Size peut être due à une différence de choix pour le modèle de régression. Cependant, la courbe de la population Paris ne suit pas les courbes qui sont présentées dans l'article comme applicable à toutes les populations (graphique A, B, C de la figure 9) et il n'est pas possible de dire pour cette population qu'un Alloscore élevé correspond à un faible taux d'eGFR, alors que dans l'article cette conclusion est généralisée aux trois cohortes envoyées.

4) Contribution de la fréquence des variants

Pour voir la contribution des fréquences d'apparition des variants sur la corrélation entre l'Alloscore et les taux d'eGFR, les variants rares et communs ont été séparés selon leur fréquence dans la population européenne de 1 000 Genomes avec un seuil de fréquence à 1%. La séparation des variants a été effectuée à partir des 290 329 SNPs ayant passés le contrôle qualité ainsi que l'annotation des rsIDs. Pour les données non imputées, le fichier contenant les variants rares se compose de 114 058 SNPs, le fichier contenant les variants communs se compose de 176 271 SNPs et le fichier contenant les variants rares et communs se compose de 290 329 SNPs.

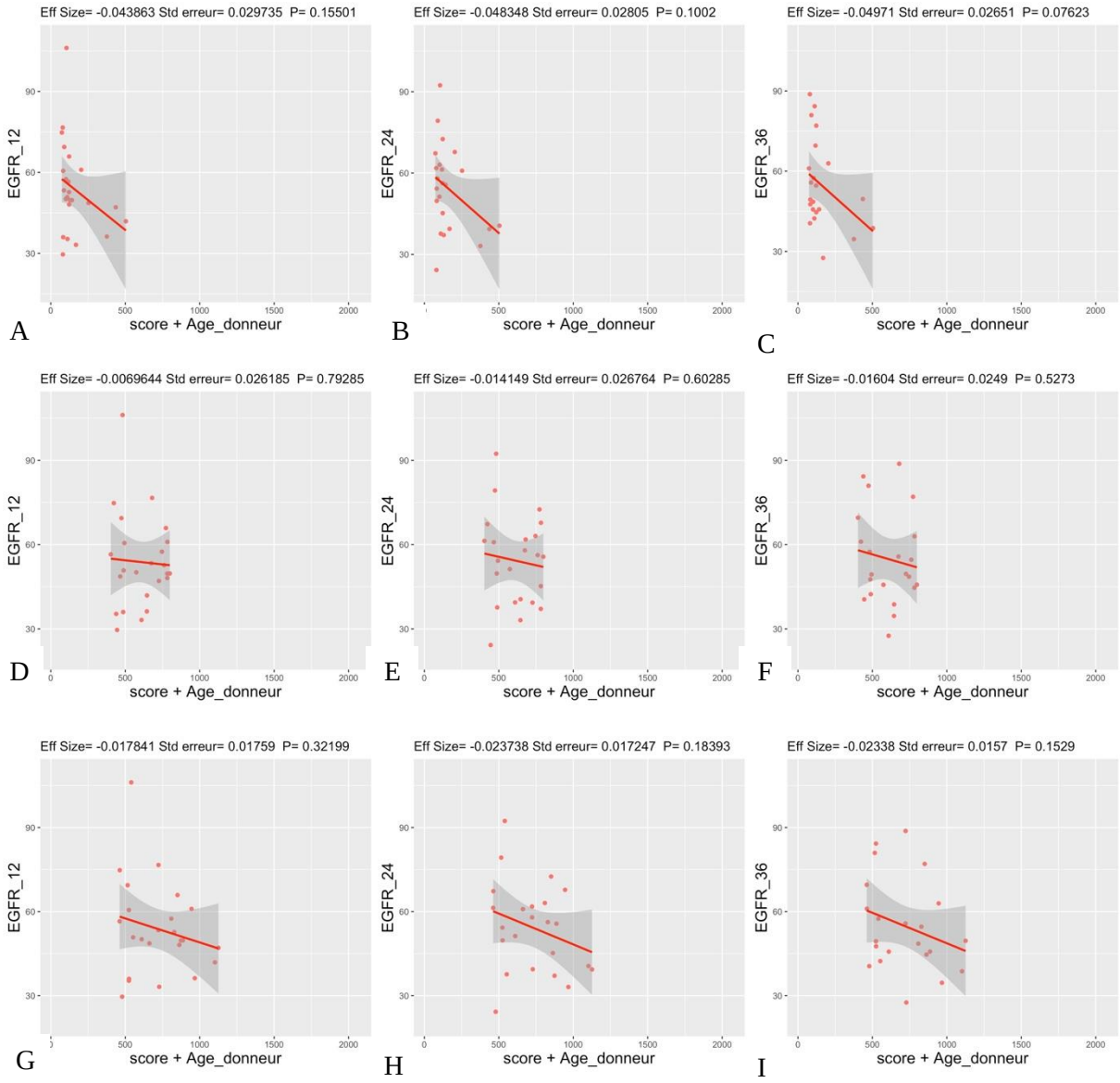


Figure 16 : Corrélation pour la population Cornel du score des variants rares (A, B, C), des variants communs (D, E, F) et des variants rares+communs (G, H, I) avec les taux d'eGFR à 12, 24, 36 mois post greffe : les variants rares, fort Effect Size, contribuent le plus dans le calcul de l'Alloscore. (Eff Size = Effect Size, Std erreur=standard d'erreur, P=P-value)

Les variants rares ont un Effect Size plus important que celui des variant communs. Pour les variants rares, le taux d'eGFR diminue en moyenne de 0.05 mL/min/1,73m^2 par unité de score à 36 mois alors que cette diminution est de 0.016 mL/min/1,73m^2 quand les données ne sont que des variants communs. Grâce à ces résultats, il est possible de voir que les variants rares utilisés pour le calcul de l'Alloscore ont une plus grande importance que les variants communs pour la corrélation qui existe entre l'Alloscore.

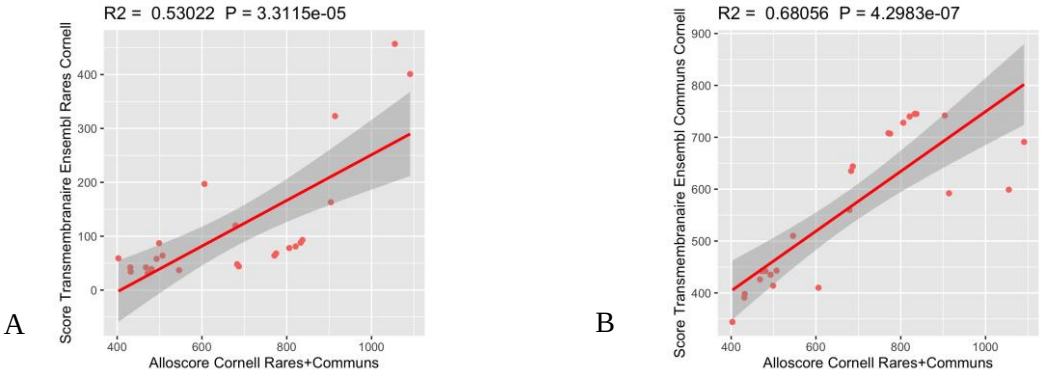


Figure 17 : Importance des variants rares (A) et des variants communs (B) non imputés dans le calcul de l'Alloscore : les variants rares ont une plus grande importance dans le calcul du score au vu d'un modèle plus fiable.

Les corrélations entre l'Alloscore calculé (uniquement avec les variants rares et communs, communs et rares) montrent que l'Alloscore avec les variants rares a une corrélation plus faible qu'avec celui calculé avec les variants rares et communs ; ce qui montre une importance des variants rares sur le calcul du score.

5) Contribution du HLA

Pour voir la contribution du HLA sur la corrélation entre l'Alloscore et les taux d'eGFR, il est nécessaire d'enlever 3 500 variants (se trouvant dans la région du HLA sur le chromosome 6) au fichier contenant les variants rares et communs qui se compose de 290 329 SNPs.

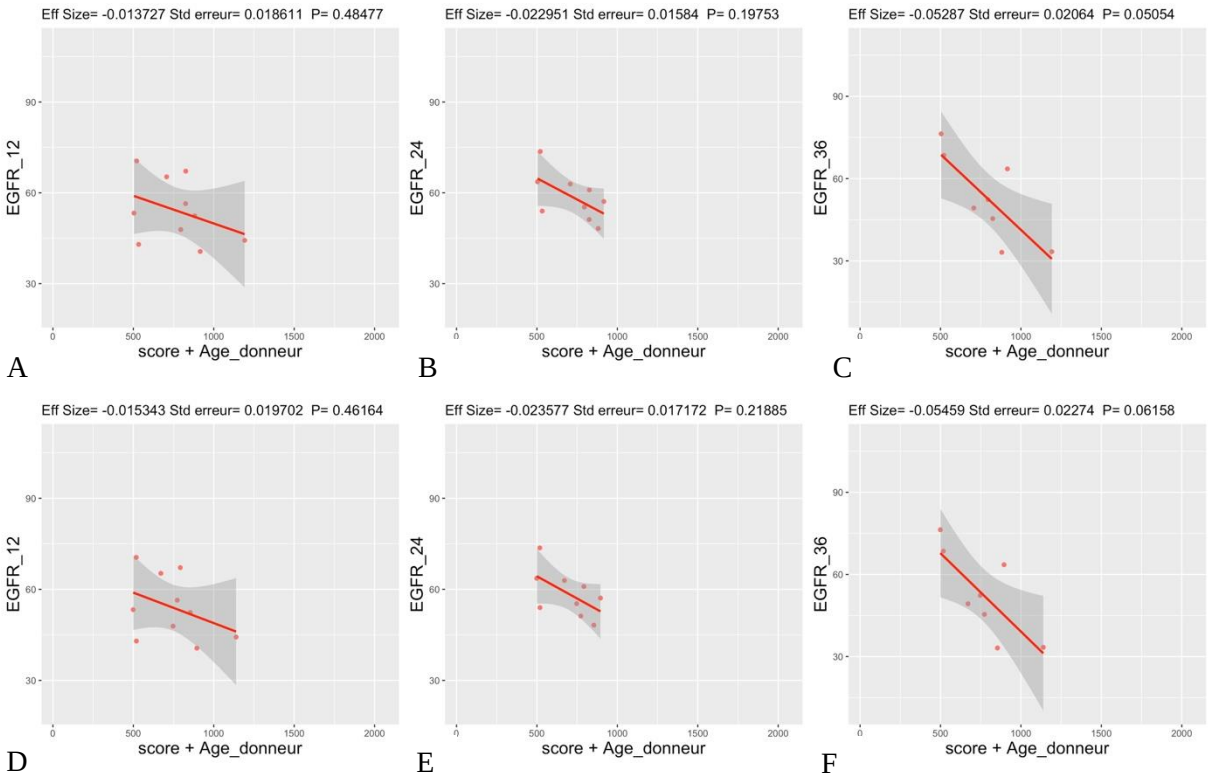


Figure 18 : Corrélation pour la population Discovery du score avec HLA (A, B, C) et sans HLA (D, E, F) avec les taux d'eGFR à 12, 24, 36 mois post greffe : Le HLA n'a aucune incidence sur le calcul de l'Alloscore. (Eff Size = Effect Size, Std erreur=standard d'erreur, P=P-value)

Ces résultats montrent que les corrélations sont identiques que le HLA soit présent ou non. Il est donc possible de conclure que le HLA n'a pas d'incidence sur la corrélation entre l'Alloscore et le taux d'eGFR. Le taux d'eGFR à 36 mois diminue de 0.05 mL/min/1,73m^2 par unité de score.Cette étape a été également répétée sur les variants rares et les variants communs et il est possible de faire la même conclusion.

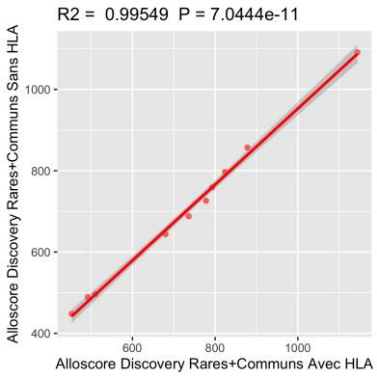


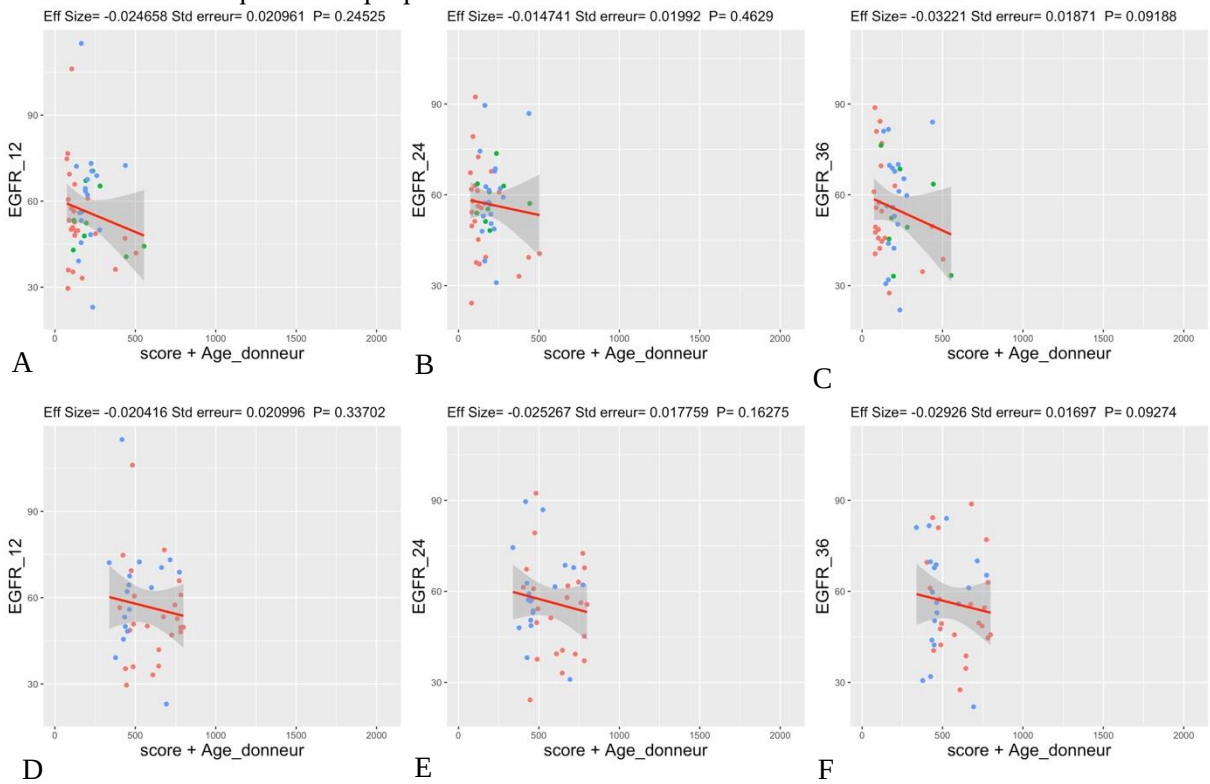
Figure 19 : Corrélation pour déterminer l'incidence du HLA

La corrélation calculée entre l'Alloscore obtenu avec et sans HLA montre qu'il y a une très forte corrélation entre l'Alloscore avec HLA et l'Alloscore sans HLA, ce qui confirme que le HLA n'a pas d'incidence sur le calcul de l'Alloscore.

6) Analyses sur les trois populations regroupées en une seule

a. Analyse sur les données non imputées

Lorsque les trois cohortes sont réunies, la courbe de régression que l'on retrouve est sensiblement identique à celle proposée dans l'article du Pr. Laurent Mesnard.



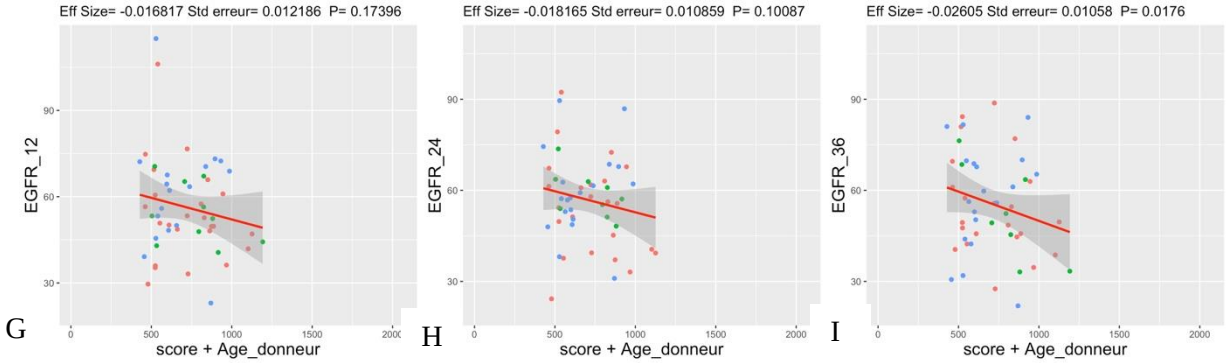


Figure 20 : Corrélation pour les 3 populations du score des variants rares non imputés (A, B, C), des variants communs non imputés (D, E, F) et des variants rares+communs (G, H, I) non imputés avec les taux d'eGFR à 12, 24, 36 mois post greffe : les variants rares, qui ont un fort Effect Size, contribuent le plus dans le calcul de l'Alloscore. (Eff Size = Effect Size, Std erreur=standard d'erreur, P=P-value)

Les variants rares ont plus d'importance que les variants communs pour les corrélations entre l'Alloscore et le taux d'eGFR au contraire du HLA qui n'a toujours pas d'incidence. Le taux d'eGFR diminue en moyenne de 0.02 mL/min/1,73m^2 par unité de score à 12, 24 ou 36 mois avec une importance des variants rares qui ont un Effect Size plus important.

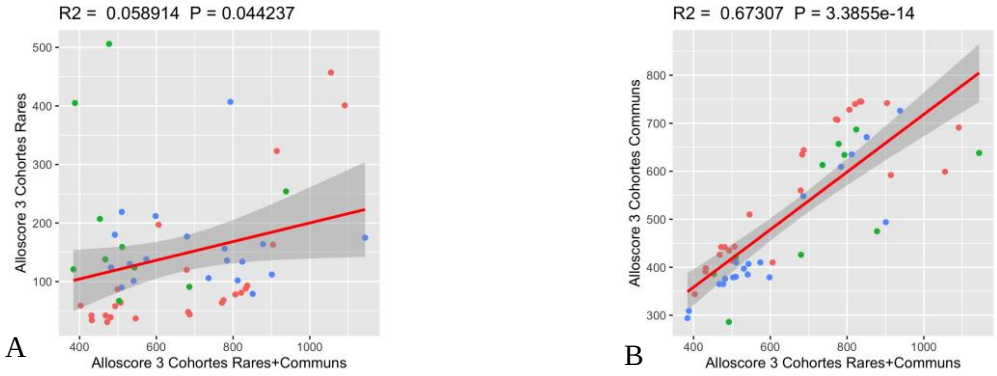


Figure 21 : Importance des variants rares (A) et des variants communs (B) non imputés

Ces corrélations montrent que les variants rares non imputés ont une très forte incidence sur le calcul de l'Alloscore due à une faible corrélation entre le score calculé à partir de variants rares non imputés (avec un modèle mieux adapté à nos données au vu du faible R2) et celui composé des variants rares et communs non imputés.

b. Réalisation de l'Alloscore sur les données imputées composées uniquement de variants codants.

Les fichiers de données génétiques ont été imputés par le serveur Michigan. Après l'imputation, les variants codants ont été sélectionnés s'ils sont missense_variant, stop_gained, stop_lost, frameshift_variant, splice_donor_variant, splice_acceptor_variant. Comme précédemment, à partir de ces nouveaux fichiers, il y a eu la séparation des variants communs et rares ainsi que la formation des fichiers sans HLA. Les analyses suivantes ont donc été réalisées sur un fichier contenant 148 054 variants rares codants, un fichier contenant 19 030 variants communs codants ainsi qu'un fichier de 167 084 variants regroupant les variants rares et communs codants auxquels il est nécessaire d'enlever la région du HLA dans un second temps.

Le HLA n'a une nouvelle fois, aucune d'incidence sur la corrélation entre l'Alloscore et le taux d'eGFR quel que soit le fichier de variants utilisés. L'IBD et L'IBS ne sont toujours pas informatifs.

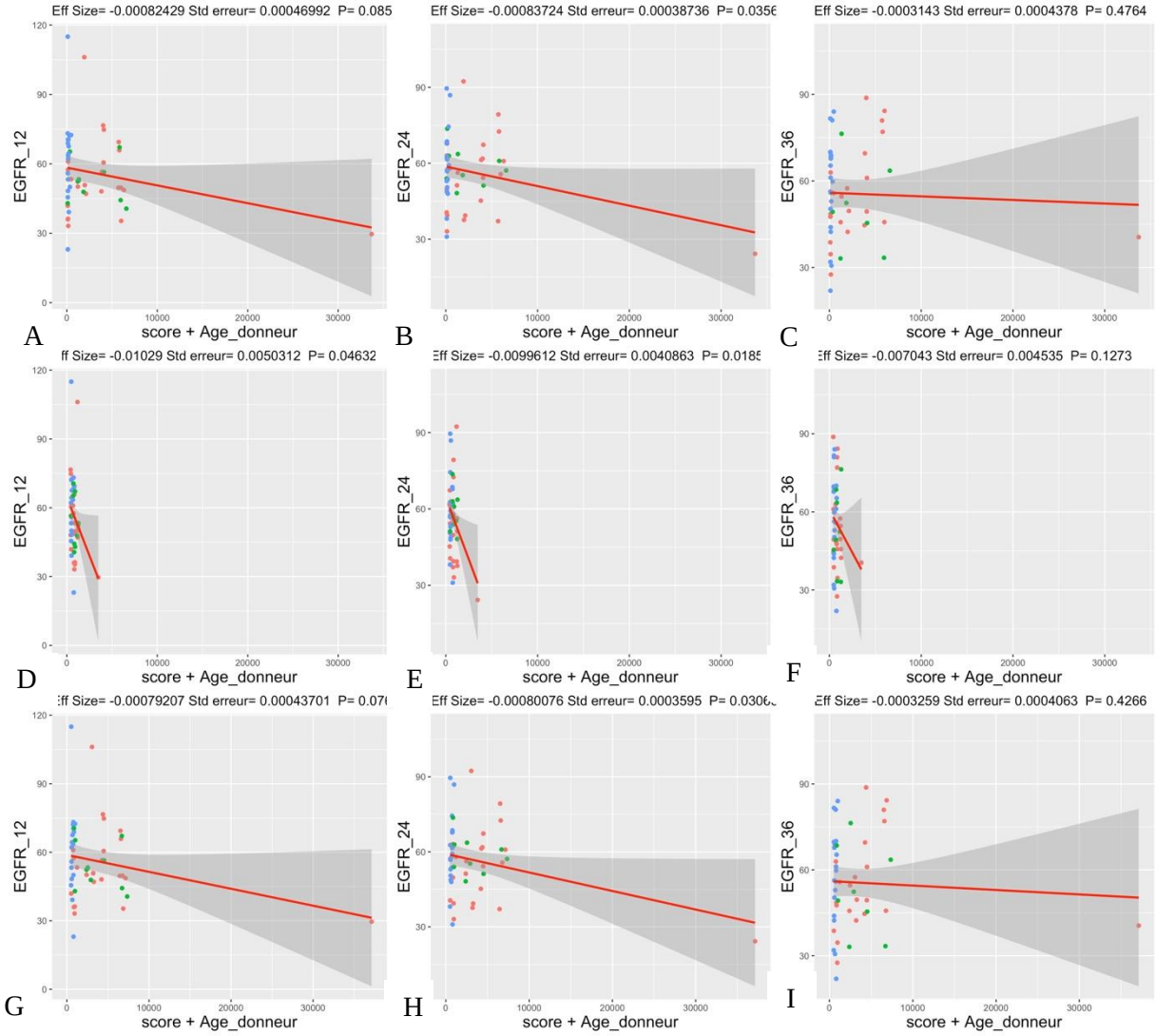


Figure 22 : Corrélation pour les 3 populations du score des variants rares codants imputés (A, B, C), des variants communs codants imputés (D, E, F) et des variants rares+communs codants imputés (G, H, I) avec les taux d'eGFR à 12, 24, 36 mois post greffe : les variants rares ont une importance dans le calcul du score. (Eff Size = Effect Size, Std erreur=standard d'erreur, P=P-value)

Ces résultats montrent une nouvelle fois l'importance des variants rares dans la corrélation entre l'Alloscore et les taux d'eGFR. En effet, ils ont un Effect Size plus important que celui des variants communs.

Un couple faisant partie de la cohorte Cornell se détache des autres couples et ce couple semble être l'élément essentiel à cette corrélation. La position des différents couples, l'échelle des scores, le nombre de variants obtenus après l'imputation ainsi que les valeurs de p-value et la position du couple éloigné des autres, nous suggèrent que l'imputation ne s'est pas bien réalisée ou que seuls les variants codants ne permettent pas de reproduire l'Alloscore sur des données GWAS.

c. Réalisation de l'Alloscore sur les données imputées composées des variants imputés codants et non codants.

Après l'imputation des données, les variants n'ont subi aucune sélection contrairement aux analyses précédentes. Ainsi, le fichier sur lequel les études ont été réalisées, est composé de 10 440 601 variants dont 9 647 607 variants rares et 792 994 variants communs. Pour pouvoir comparer nos différentes analyses, il est également nécessaire de créer de nouveaux fichiers en excluant la région du HLA sur le chromosome 6.

Comme précédemment, il est possible de faire les mêmes conclusions concernant le HLA. Il n'a aucune incidence et l'IBD et L'IBS ne sont toujours pas informatifs. Les résultats étant similaires, il n'est pas nécessaire de les représenter.

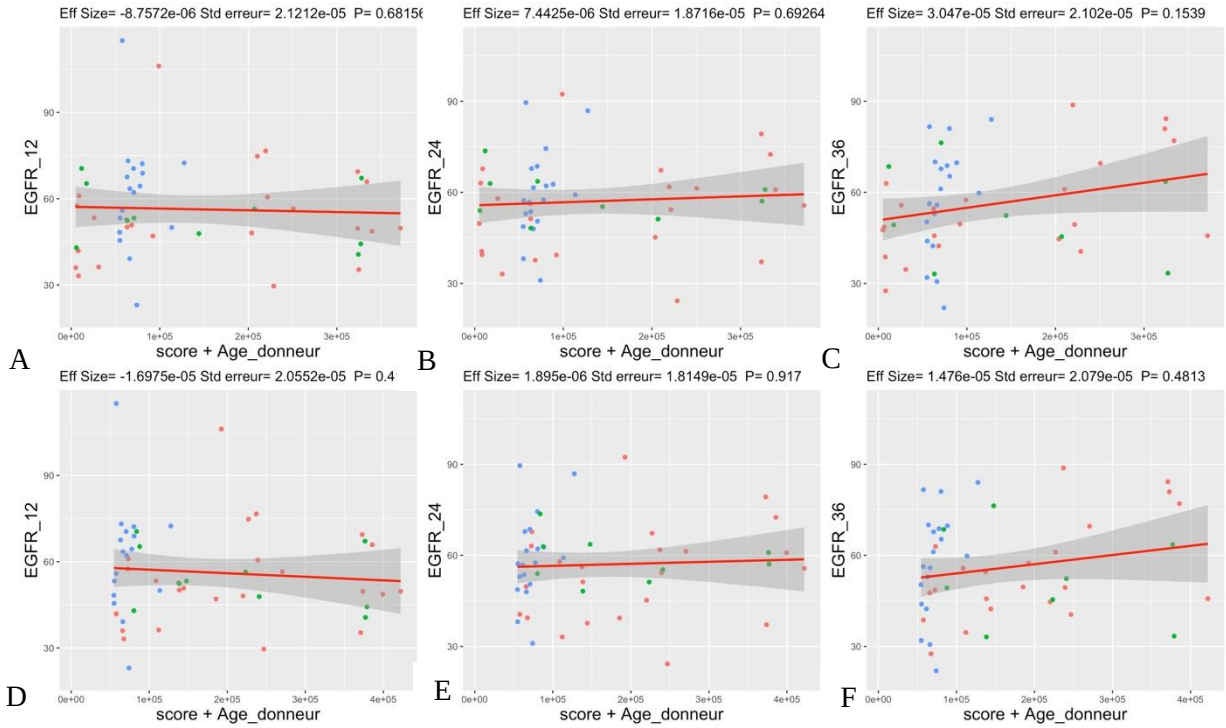


Figure 23 : Corrélation pour les 3 populations du score des variants rares imputés (A, B, C) et des variants rares+communs imputés (D, E, F) avec les taux d'eGFR à 12, 24, 36 mois post greffe : les variants rares ont une importance dans le calcul du score. (Eff Size = Effect Size, Std erreur=standard d'erreur, P=P-value)

Ces résultats confirment l'importance des variants rares dans la corrélation entre l'Alloscore et les taux d'eGFR. Les populations se positionnent en clusters et non de façon uniforme comme nous le pensions. Cela confirme donc que l'imputation ne s'est pas réalisée correctement. L'imputation des données exomiques est plus complexe que pour des données de GWAS et le server Michigan utilisé n'est peut-être pas adapté pour l'imputation de données exomiques.

Pour l'imputation, seuls les variants communs (les plus fréquents, 0,01) ont été sélectionnés selon la même technique utilisée pour des données GWAS. Or, on se retrouve avec 12 fois plus de variants rares que de variants communs ce qui induit un biais potentiel dans le calcul du score. Il serait donc nécessaire d'inclure les variants rares pour l'imputation, ceci pour forcer la sélection d'haplotypes contenant les variants rares connus grâce au WES.

Conclusions et Perspectives

1) Discussion et Conclusion

Mon stage portait sur la validation d'un Alloscore proposé par le Pr. Mesnard qui pourrait appuyer la compatibilité ABO (groupe sanguins) et la compatibilité tissulaire (HLA) lors de greffe de rein. Cette validation s'est effectuée sur les données séquencées d'exome entier (WES), sur des données séquencées d'exome entier imputé codant ou non.

La validation de ce score sur les données séquencées d'exome entier a tout d'abord montré que le score transmembranaire n'était pas le seul score corrélé à la fonction rénale à trois ans post-greffe contrairement à ce qui a été démontré dans l'article. Cela signifie que les protéines à la surface des cellules ne seraient pas les seules à intervenir lors de la compatibilité entre un donneur et un receveur pour une greffe. Lors de cette validation, il a été confirmé que ce score est indépendant du HLA qui ne serait pas le seul facteur à prendre en compte lors d'une greffe d'organe mais également d'autres facteur non immunogènes.

L'article de Reindl-Schwaighofer R et al a montré que, lors de l'exécution d'un Alloscore, les variants communs ont une importance significative. Or, notre étude a pu montrer que seuls les variants rares avaient une importance pour la corrélation entre l'Alloscore et la fonction rénale. Cette incapacité

à montrer l'importance des variants communs dans le calcul de ce score peut être du au faible nombre d'individus qui composent nos cohortes. En effet, cela peut être engendré par un problème de puissance de notre étude qui a été réalisée sur des données de séquençage d'un total de 53 couples de donneur-receveur alors que l'importance des variants communs a été observée sur des données de génotypage d'une cohorte de 477 couples de donneur-receveur. Les deux études n'ayant pas été effectuées sur le même type de données génétiques (respectivement en WES et en GWAS), l'information contenue dans ces données n'est pas la même et il n'est peut-être donc pas possible d'effectuer les mêmes observations ni les mêmes conclusions.

Toutes ces observations montrent que, dans l'optique d'une greffe, il serait intéressant de réaliser un séquençage en complément d'un génotypage des données génétiques des patients pour appuyer le choix du couple donneur-receveur déterminé par la compatibilité HLA et ainsi favoriser la survie du greffon grâce à une compatibilité immunogène et non-immunogène.

L'étude sur des données séquencées d'exome entier imputé codant ou non, montre qu'il y a une mauvaise couverture du génome lors de l'imputation. Les données de séquençage d'exome ont été imputées selon le même pipeline que pour des données GWAS et seuls les variants communs ont été sélectionnés ce qui nous donne des résultats aberrants. Il serait intéressant de voir si en incluant les variants rares lors de l'imputation, si la couverture du génome serait plus importante, si les résultats obtenus seraient semblables à des données GWAS. Ce qui nous permettrait de calculer l'Alloscore et ainsi le valider sur des données génomiques séquencées imputées.

Concernant la comparaison entre l'Alloscore, l'IBD et l'IBS, il n'est pas possible d'effectuer de bonne conclusion alors que les scores IBD et IBS pourraient être une bonne alternative à ce score. Pour pouvoir recalculer l'IBD et l'IBS, il serait intéressant de voir si les variants rares ne créent pas une erreur pour le calcul de l'IBD et l'IBS. Cette hypothèse est évoquée dans la documentation de PLINK. Nous pensions que le nombre de variants était suffisant pour minimiser l'effet des variants rares et réaliser le calcul de ces scores. Au vu de nos résultats, le calcul devra être refait sans prendre en compte les variants rares.

L'Alloscore est donc bien un score indépendant du HLA qui pourrait permettre de déterminer les meilleurs couples de donneur et de receveur pour une greffe de rein. Ce score serait donc bien un outil qui appuie la compatibilité HLA pour le choix des patients, mais dont l'importance des variants rares et des variants communs reste à préciser en augmentant la puissance de l'étude.

2) Perspectives

Dans un premier temps, il sera nécessaire de trouver le bon modèle de régression utilisé dans l'article pour valider son score et l'appliquer à de nouvelles données en transplantation. Il serait intéressant :

- de comprendre les différences qui existent entre une imputation de données WES et une imputation de données GWAS.
- d'adapter les données en notre possession pour le calcul de l'IBS et l'IBD et ainsi pouvoir comparer ces deux scores à l'Alloscore.
- d'identifier et d'exclure les couples dont les résultats sont douteux lors des différents calculs de scores pour voir le comportement de l'Alloscore.
- d'effectuer de nouvelles validations en augmentant le nombre d'individus dans la cohorte d'étude pour améliorer la puissance de l'étude.

3) Conclusion personnelle

Durant ce stage, de nombreuses compétences ont été acquises dans plusieurs domaines. J'ai tout d'abord fait de nombreux progrès au niveau de la programmation grâce au langage Bash surtout dans la manipulation de fichiers via AWK mais j'ai également pu découvrir de nombreux logiciel pour l'analyse NGS. La polyvalence et la liberté d'utilisation de langages différents m'a permis d'approfondir et d'améliorer mes connaissances sur le langage R. J'ai essayé d'être aussi indépendante que possible en faisant de nombreuses recherches sur Internet et plus particulièrement sur des forums (Biostar et Stack Overflow) pour résoudre mes problèmes de programmations. Enfin, j'ai eu l'occasion de présenter mon projet sous forme d'un poster à la conférence JOBIM 2019 où j'ai eu une première expérience très intéressante sur la communication scientifique et de me familiariser avec ce type d'exercice.

Références

1. Institut Nationnal du cancer (https://www.e-cancer.fr/Patients-et-proches/Les-cancers/Cancer-du-rein/Anatomie-du-rein)
2. HAS (Haute Autorité De Santé) Guide du parcours de soins. Maladie Rénale Chronique de l'adulte (Février 2012)
3. Couchoud, C., Lassalle, M. & Stenge, I. Registre de Réseau Epidémiologie Information en Néphrologie (REIN). (2015).
4. néphropôle Lyon,projet de transplantation rénale : https://www.nephrologie-lyon.com/projet-de-transplantation-renale.html
5. Le Manuel MSD Version pour professionnels de la santé, de Merck Sharp & Dohme Corp. : https://www.msdmanuals.com/fr/professional/immunologie-troubles-allergiques/biologie-du-syst%C3%A8me-immunitaire/syst%C3%A8me-de-l-human-leukocyte-antigen-hla
6. Mesnard L, Muthukumar T, Burbach M, Li C, Shang H, Dadhania D, et al. (2016) Exome Sequencing and Prediction of Long-Term Kidney Allograft Function. PLoS Comput Biol 12(9): e1005088. https://doi.org/10.1371/journal.pcbi.1005088
7. Reindl-Schwaighofer R, Heinzel A, et al. (2019) Contribution of non-HLA incompatibility between donor and recipient to kidney allograft survival: genome-wide analysis in a prospective cohort. Lancet 2019 ; 393: 910–17. https://doi.org/10.1016/S0140-6736(18)32473-5
8. IGSR: The International Genome Sample Resource : https://www.internationalgenome.org/wiki/Analysis/vcf4.0/
9. Dixon -Salazar et al. , 2012 ; Bonnefond et al. , 201
10. Site illumina https://www.illumina.com/techniques/sequencing/dna-sequencing/targeted-resequencing/exome-sequencing.html
11. Broad Institute of Harvard : tutoriel plink : http://zzz.bwh.harvard.edu/plink/index.shtml mis à jour le 25 janvier 2017
12. Site officiel de l'outil plink (http://www.cog-genomics.org/plink/1.9/formats#bed) mis à jour en 2019
13. Site de GWASPI (Genome Wide Association Study Pipeline) http://www.gwaspi.org/?page_id=213 mis à jour en 2019
14. Site du NIH, National Human Genome Research Institute https://www.genome.gov/genetics-glossary/Genome-Wide-Association-Studies
15. Site officiel : https://samtools.github.io/bcftools/bcftools.html
16. Site officiel : http://vcftools.sourceforge.net/
17. Site de McCarthy Group Tools pour les outils de pré et post-imputation : http://www.well.ox.ac.uk/~wrayner/tools/
18. Site de téléchargement du script CheckVCF.py et z-ses informations : https://genome.sph.umich.edu/wiki/CheckVCF.py
19. Site officiel du projet HRC (Haplotype Reference Consortium) : http://www.haplotype-reference-consortium.org/home
20. Site officiel du Site d'Imputation Sanger : https://imputation.sanger.ac.uk/
21. Site officiel du Site d'Imputation Michigan : https://imputationserver.sph.umich.edu/index.html#!pages/home
23. Site officiel : https://imputationserver.sph.umich.edu/index.html#!
23. Site officiel : http://annovar.openbioinformatics.org/en/latest/
24. Site officiel : https://www.ensembl.org/biomart/martview/b3c461de41126047d52476a0f9a55062
25. Site officiel : https://www.ensembl.org/info/docs/tools/vep/index.html
26. Site officiel : https://www.ensembl.org/index.html
27. Site officiel : https://www.proteinatlas.org/about
28. International Society of Genetic Genealogy https://isogg.org/wiki/Identical_by_descent
29. International Society of Genetic Genealogy https://isogg.org/wiki/Identical_by_state
30. Algorithme de la fonction genome de plink : https://github.com/GELOG/adam-ibs/wiki/Algorithm-for-Pairwise--IBS-IBD-computation-(--genome)
31. Site officiel du cluster Bird hébergé par l'Université de Nantes : https://pf-bird.univ-nantes.fr/ressources/cluster-de-calcul/birdcluster-1313050.kjsp Mis à jour le 16 juillet 2019