Université de Montréal

Développement de méthodes bio-informatiques pour la découverte de variants codants et non codants dans le cadre des traits sanguins

par
Sébastian Méric de Bellefon

Département de biochimie et médecine moléculaire
Faculté de médecine

Mémoire présenté à la Faculté des études supérieures

Avril, 2020

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé:

Développement de méthodes bio-informatiques pour la découverte de variants codants et non codants dans le cadre des traits sanguins

présenté par:

Sébastian Méric de Bellefon

a été évalué par un jury composé des personnes suivantes:

Guillaume Lettre	Directeur de recherche
Stephen Michnick	Membre du jury
Sylvie Hamel	Présidente du jury

RÉSUMÉ

La santé cardiovasculaire, la fonction immunitaire, l'hémostase et la réponse à d'autres maladies dépendent de l'abondance et des caractéristiques spécifiques des cellules sanguines. Au fil des années, un effort considérable a été fait pour trouver les variants génétiques, les gènes et les mécanismes de régulation impliqués dans la création de ces cellules.

L'inactivation d'un allèle, appelée "perte de fonction" (LoF), est un type de variant codant que nous aimerions associer aux phénotypes sanguins. Comme ces mutations ne peuvent pas être artificiellement induites chez l'humain, pour des raisons éthiques évidentes, nous observons les occurences naturelles de ces pertes de fonction et espérons que la taille des cohortes sera suffisante pour trouver des associations statistiquement significatives.

L'inactivation des deux allèles, appelée "knockout" (KO), peut avoir des conséquences plus fortes qu'une simple perte de fonction. Nous espérons également trouver des KO d'origine naturelle grâce à la taille des cohortes. La combinaison de deux variants LoF différents sur les deux allèles est appelée knockout hétérozygote composé.

Nous nous intéressons également aux variants non codants qui affectent l'expression des gènes impliqués dans l'hématopoïèse. Certains de ces variants créent ou perturbent des sites de liaison des facteurs de transcription (TF), ces protéines qui se lient à des séquences d'ADN spécifiques et régulent l'expression des gènes. Les sites de liaison (TFBS) des facteurs de transcription se trouvent dans les promoteurs des gènes et dans les amplificateurs spécifiques au type cellulaire.

Alors que certaines de ces mutations peuvent être bénignes ou même bénéfiques, la présence d'un LoF ou d'un KO peut être trop nuisible à la survie de l'individu. Les résultats de cette étude sont limités par le biais de survie.

Comparée à une étude d'association pangénomique, cette étude se concentre sur un plus petit nombre de variants génétiques pour augmenter la puissance statistique et offrir une interprétation pour les résultats statistiquement significatifs.

Le programme Trans-Omics for Precision Medicine (TOPMed) recueille et garantit la qualité des 45 000 séquences du génome entier que nous avons utilisées dans cette étude, ainsi que les bilans sanguins correspondants. Grâce à ces données, nous avons pu trouver plusieurs associations connues et nouvelles entre des variants rares et des phénotypes sanguins.

Mots-clés : Association pangénomique, Méta-analyse, Expression génétique, Hématopoïèse, Facteur de transcription, Promoteur, SNP

ABSTRACT

Cardiovascular health, immune function, hemostasis and the response to other illnesses depend on the abundance and specific features of blood cells. Over the years, a considerable effort has been made to find which genetic variants, genes and regulatory mechanisms are involved in the creation of these cells.

The inactivation of an allele, called a loss-of-function (LoF), is a type of coding variant we would like to associate with blood phenotypes. For obvious ethical reasons, these mutations cannot be artificially induced in human, so we fall back on natural occurrences and hope that large cohorts will provide enough samples to find statistically significant associations.

The inactivation of both alleles, called a knockout (KO), may have stronger consequences than a simple loss-of-function. We also hope to find naturally occurring knockouts thanks to the size of a large cohort. The combination of two different LoF variants is called a compound heterozygote knockout.

We are also interested in non-coding variants that affect the expression of genes that are involved in hematopoiesis. Some of these variants create or disrupt the binding sites of transcription factors (TF), the proteins that bind to specific DNA sequences and regulate gene expression. Transcription factors binding sites (TFBS) are found in gene promoters and cell type specific enhancers.

While some of these mutations can be benign or even beneficial, the presence of a LoF or KO may be too detrimental for the individual to survive. The results of this study are limited by survival bias.

Compared to a genome-wide association study, this study focuses on a smaller number of genetic variants to increase statistical power and give an interpretation to the statistically significant findings.

The Trans-Omics for Precision Medicine (TOPMed) program collects and ensures the quality of the 45,000 whole-genome sequences we used in this study, as well as the corresponding complete blood counts. Thanks to this raw data, we were able to find several known and novel associations between rare variants and blood phenotypes.

Keywords : GWAS, SNP, Meta-analysis, Gene expression, Hematopoiesis, Transcription factor, Promoter

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ABBREVIATIONS

ATAC-seq : Assay for Transposase-Accessible Chromatin using sequencing
ChIP-seq : Chromatin Immunoprecipitation Sequencing
CMP : Common myeloid progenitor
eQTL : Expression quantitative trait locus
LoF : Loss of function variant
FWER : Family-wise error rate
FDR : False discovery rate
GMP : Granulocyte-monocyte progenitor
GWAS : Genome-wide association study
HSC : Hematopoietic stem cells
KO : Knockout gene
LMPP : Lymphoid-primed multipotential progenitor
MEP : Megakaryocyte-erythrocyte progenitor
MPP : Multipotential progenitors
pKO : Predicted knockout gene
pLoF : Predicted loss of function variant
PWM : Position weight matrix
SNV : Single-nucleotide variant
SNP : Single-nucleotide polymorphism
TF : Transcription factor
TFBS : Transcription factor binding site
TSS : Transcription start site
WGS : Whole genome sequence

DÉDICACE

Pour Katrine, la meilleure compagnie imaginable pendant ces mois d'isolation.

REMERCIEMENTS

Je tiens d'abord à remercier mon directeur de maîtrise, Guillaume Lettre, qui m'a accueilli dès mon année de propédeutique. Merci pour ses encouragements, sa surprenante patience et l'environnement de travail positif qu'il a créé à l'Institut de Cardiologie. Je n'en espérais pas autant.

Merci à toute l'équipe du laboratoire, en particulier Mélissa pour la qualité de sa critique, Ken pour son aide avec les outils bio-informatiques, Florian pour son flot d'information ininterrompu sur les technologies single-cell, et Simon pour sa gentillesse pendant notre brève collaboration. Une mention spéciale pour Yann et ses photos de famille.

Je remercie la Faculté des études supérieures et postdoctorales pour son soutien financier, ainsi que les Instituts de recherche en santé du Canada pour le financement du laboratoire.

Pour conclure, je remercie les membres du jury qui ont accepté de prendre le temps pour évaluer ce mémoire.

CHAPITRE 1

INTRODUCTION

Les études d'association pan-génomiques (GWAS) ont identifié des milliers de variants associés avec des traits complexes chez les humains, et le catalogue GWAS (NHGRI, 2018) recense aujourd'hui 78161 associations issues de 3640 publications. Ces études découvrent des corrélations statistiques entre des variations génétiques et des variations phénotypiques, qui suggèrent l'existence de liens causaux et nous invitent à valider ces liens de cause à effet à travers d'autres expériences. La méthode GWAS classique sera décrite plus en détail dans la section 3.4.

Plus de 90% des variants découverts se trouvent dans des régions non codantes et possèdent généralement un rôle dans la régulation de l'expression génique (MAURANO et al., 2012).

L'étude présente se focalise en premier sur les variants des régions codantes (SNPs et indels courts) dont on prédit qu'ils causent une perte de fonction (pLoF). Ce choix est justifié par l'augmentation de la puissance statistique qui résulte d'un nombre de tests plus faible et par l'a priori que ces variants sont plus susceptibles d'influencer fortement des phénotypes. De plus, la recherche de "knockout", c'est-à-dire de participants chez lesquels les deux copies d'un gène sont dysfonctionnelles, permet de découvrir de nouvelles associations.

Dans une deuxième partie, nous cherchons des variants dans les régions non-codantes qui sont susceptibles d'affecter le niveau d'expression d'un ou plusieurs gènes en modifiant les sites de liaison des facteurs de transcription (TFBS) d'une manière qui augmente ou réduit leur affinité avec l'ADN. Un nouveau logiciel appelé find-tfbs a été créé pour nous assister dans cette recherche et a été publié. Le rôle de ce type de variant a déjà été validé expérimentalement pour des gènes spécifiques (BAUER et al., 2013 ; CLAUSSNITZER et al., 2015 ; LESSARD et al., 2017 ; MUSUNURU et al., 2010).

Dans ces deux cas, le type de variant fournit une hypothèse sur la cause de l'association mais ne la garantit pas. La cause est plus claire pour les pLoF et KO, alors que les variants TFBS peuvent altérer l'expression d'un (ou plusieurs) gène(s) en cis qui ne sont pas les plus proches du variant, et occasionnellement cette interaction se fait en trans. Une analyse RNA-seq pourrait confirmer ces hypothèses.

Cette méthode est ici appliquée à quinze phénotypes sanguins et à 44,709 participants du projet TOPMed [1] issus de quatre groupes ethniques.

1. Site web : https://nhlbiwgs.org

1.1 Résultats antérieurs des études pan-génomiques

Le rythme d'apparition de nouvelles études d'association est soutenu et leur complexité augmente au fil des années (Figure 1.1). Cependant, la diversité ethnique telle que mesurée par le GWAS diversity monitor (THE LEVERHULME CENTRE FOR DEMOGRAPHIC SCIENCE, 2020) reste faible (91.25% des participants sont d'origine Européenne et seulement 0.94% sont d'origine Africaine), ce qui entrave la découverte de variants rares qui nous aideraient à comprendre certains mécanismes moléculaires à la fois pour la population concernée et pour les autres populations.

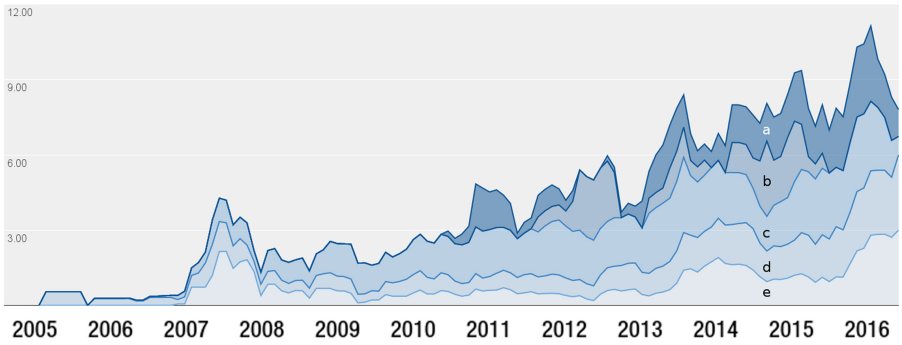


FIGURE 1.1 – Complexité des études pan-génomiques au fil du temps (A), Nombre d'études d'interaction SNP-environment, (B) Nombre d'études d'interaction SNP-SNP, (C) Nombre de traits par étude, (D) Nombre d'ethnicités par étude, (E) Nombre de GWAS par étude. Les valeurs sont normalisées pour accorder un poids égal à chaque catégorie. Source : J. MACARTHUR et al., 2017, D1

1.1.1 Historique des résultats cliniques

L'agrégation en 2016 de 24 études pan-génomiques sur des populations européennes, asiatiques et africaines, ainsi que sur des populations malades fournit une liste de 145 loci associés avec des traits hématologiques. Ces loci sont principalement situés dans les régions non-codantes. "Les gènes proches de ces loci sont enrichis pour des gènes qui régulent des fonctions hématologiques et des gènes impliqués dans des maladies Mendeliennes du sang, telles que l'anémie falciforme, l'anémie hémolytique, la thrombocytopénie, la leucémie et l'insuffisance de la moelle osseuse)" (VASQUEZ, MANN, CHEN & SORANZO, 2016, Suppl 1).

Deux des études mentionnées ont permis de découvrir 66 nouveaux gènes associés avec les érythrocytes et les thrombocytes, qui ont fait l'objet d'une validation fonctionnelle dans des organismes modèles (GIEGER et al., 2011 ; van der HARST et al., 2012).

MOUSAS et al., 2017 ont découvert 56 nouvelles associations entre des variants codants rares et des traits hématologiques, et POLFUS et al., 2016 ont découvert deux loci (dont un site d'épissage)

liés à hématopoïèse à partir de séquences complètes d'éxomes et les ont validé grâce à une expérience CRISPR/Cas9 sur des cellules progénitrices. Par rapport à un GWAS classique, qui analyse les variants génétiques du génome entier, l'analyse des éxomes réalisée par ces deux études offre une puissance statistique supérieure car le nombre d'hypothèses testées est nettement plus petite.

1.1.1.1 Pertes de fonction et knockout

Nous pouvons aller plus loin dans cette direction et augmenter notre puissance statistique en analysant uniquement les variants codants qui semblent causer une perte de fonction (section 2.3.1), via une analyse classique (variant par variant) et via une approche knockout. Pour une taille de population donnée, les approches perte-de-fonction/knockout mettent en valeur des variants plus rares et dont l'effet sur le phénotype est plus faible.

Une originalité de nos travaux est la recherche d'association entre ces knockouts prédits (pKO) et des phénotypes sanguins. Un knockout est l'invalidation totale d'un gène due à la présence de deux allèles perte-de-fonction (ou d'un seul sur le chromosome X chez un homme).

NARASIMHAN, XUE et TYLER-SMITH, 2016 recense des knockouts dans ~500 gènes distincts au sein d'une population de 40000 participants non-consanguins (knockouts causés par deux copies du même variant perte-de-fonction). Nos travaux incluent les knockouts hétérozygotes composés de deux variants pLoF différents.

Ce nombre augmente significativement dans une population consanguine : par exemple SALEHEEN et al., 2017 trouve 1317 pKO distincts (causés par 49138 mutations pLoF rares) dans une cohorte de 10503 participants d'origine Pakistanaise et observe des différences phénotypiques marquées. Au sein d'une population similaire, NARASIMHAN, HUNT et al., 2016 estime que chaque individu est porteur de 1.6 variants pLoF récessifs qui sont mortels lorsqu'ils forment un knockout ; leur étude est cependant limitée aux knockouts homozygotes.

NARASIMHAN, HUNT et al., 2016 indique aussi que le nombre de knockouts est proportionnel à la proportion d'autozygotie dans le génome. 94.9% des knockouts homozygotes se trouvent dans les régions autozygotes qui représentent en moyenne 5.6% du génome de cette population.

On estime que chaque individu dans le "1000 Genomes Project" porte environ 100 perte-de-fonctions dont environ 20 knockouts. Ce nombre élevé suggère que la sévérité des knockouts peut être réduite par des "mécanismes de redondance inattendus", tout du moins chez les individus qui ont atteint l'âge adulte (Daniel G. MACARTHUR et al., 2012). Dans d'autres cas, un knockout peut ne causer aucun phénotype clinique (MCGREGOR et al., 2020) chez les rares individus knockouts connus.

De manière générale, les associations statistiques sont nettement plus fortes pour les variants ho-

mozygotes que pour les variants hétérozygotes (odds ratio de 1.5-1.6 contre seulement 1.1, (FRAZER, MURRAY, SCHORK & TOPOL, 2009), et les expériences fonctionnelles suggèrent que l'effet d'un knockout est plus fort que celui d'un allèle dominant perte-de-fonction (D. G. MACARTHUR et al., 2014). Il est donc plausible que l'étude présente découvre des effets phénotypiques forts causés par des knockouts.

Étant donné qu'un gène peut être inactivé par des variants différents sur chaque allèle (hétérozygotes composé), nous utilisons la phase des génotypes TOPMed. Ainsi, nous étudions à la fois les pKO homozygotes et les pKO hétérozygotes composés, au contraire de DEBOEVER et al., 2017 et NARASIMHAN, HUNT et al., 2016 qui se focalisent sur les homozygotes, ce qui augmente notre puissance statistique.

1.1.1.2 Variant non-codants

Pour des variants rares non-codants, des méthodes d'analyses groupées de plusieurs variants par fenêtre glissantes (e.g SCANG) ont été proposées (LI et al., 2019 ; MORRISON et al., 2017 ; MORRISON et al., 2013 ; NATARAJAN et al., 2018). Mais ces méthodes ne prennent pas en compte notre connaissance des mécanismes de régulation de l'expression, et en particulier le rôle des facteurs de transcription (TF). Notre logiciel présenté dans l'article en annexe, find-tfbs, s'appuie sur cette connaissance pour prioriser l'étude des variants qui créent ou détruisent des sites de liaison des facteurs de transcription situés dans des régions de chromatine ouverte, ou plus généralement dans des régions d'intérêt. Les données extraites par find-tfbs peuvent être utilisées par les logiciels de tests d'association modernes.

1.1.2 Reproductibilité

Une étude de HIRSCHHORN, LOHMUELLER, BYRNE et HIRSCHHORN (2002) sur les traits complexes critique la robustesse des associations découvertes par les premières études d'association pangénomiques. Parmi 600 associations positives, seules 166 ont été étudiées "au moins trois fois". Parmi les autres, plus de la moitié ont été répliquées au moins une fois, mais "seulement 6 ont été répliquées à chaque fois".

Elle décrit plusieurs causes qui expliquent la faiblesse de ces résultats. Dès que possible, la présente étude tente de limiter chacune de ces causes.

— "Un biais de publication". Les résultats négatifs sont plus rarement publiées
— "Une stratification de population" qui provoque une inflation des tests statistiques (si une population est affectée par une maladie, tous ses variants caractéristiques paraissent associés)
— "Le déséquilibre de liaison" qui génère de fausses associations

— "Des interactions gène-gène et gène-environnement" ignorées par les études

Ces constats justifient plusieurs choix méthodologiques pour les études pan-génomiques modernes : un contrôle de la stratification et du déséquilibre de liaison, des études de réplication, une validation fonctionnelle et une publication systématique.

Il faut noter que HIRSCHHORN et al. (2002) se limite aux associations binaires génotype-pathologie. Il est possible que leurs résultats aient été différents s'ils avaient inclus des traits complexes continus tels que les traits sanguins de l'étude présente.

1.1.3 Contrôle de qualité

D'après TURNER et al. (2011) et SERRE et al. (2008), les problèmes de qualités suivants ont été observés dans les études d'association pan-génomiques :

— La duplication d'un participant (inter-cohortes ou intra-cohorte)
— Une déviation de l'équilibre de Hardy-Weinberg qui peut indiquer un problème de génotypage ou de séquençage
— Une erreur de sexe
— Une couverture de séquençage insuffisante pour un variant
— Les liens de parenté cachés qui peuvent provoquer des faux positifs (des associations qui ne correspondent pas à la réalité) et faux négatifs (absence d'associations réelles)
— Une stratification de population (problème similaire aux liens de parentés, mais au niveau d'une population entière)
— Les groupes éthniques auto-déclarés peuvent être imprécis

Les groupes éthniques auto-déclarés peuvent être validés en visualisant le PCA du génotype des participants (section 4.4). Les autres erreurs potentielles ont été prises en compte. En particulier, nous discuterons en détail la méthodologie utilisée pour prendre en compte les liens de parentés cachés et la structure de population (section 3.4.4).

1.2 Réplication et validation fonctionnelle

Idéalement, les associations statistiquement significatives issues de TOPMed pourront être répliquées dans d'autres jeux de données, tel que le UK Biobank (SUDLOW et al., 2015). Les associations répliquées pourront faire l'objet d'une validation fonctionnelle. En fonction du trait en question, cette expérience pourra être réalisée sur une lignée de cellules humaines hématopoïétiques éditées. Afin de réduire les coûts, les résultats non répliqués ne seront probablement pas validés de cette manière.

La technique utilisée pour l'édition du génome se base sur CRISPR/Cas9 (BOETTCHER & MCMANUS, 2015) et sur des lentivirus. Il sera donc nécessaire de concevoir une librairie de guides sgRNA spécifique à chaque locus que nous souhaitons valider.

PARNAS et al., 2015 ont utilisé une technique similaire pour identifier les gènes qui contrôlent l'induction de TNFα par les endotoxines, un mécanisme de la réponse immunitaire.

1.3 Description de l'article en annexe

L'article en annexe, intitulé "find-tfbs : a tool to identify functional non-coding variants associated with complex human traits using open chromatin maps and phased whole-genome sequences", va prochainement être soumis pour publication au journal Bioinformatics. Il présente un nouvel outil informatique, find-tfbs, qui recherche les sites de liaison des facteurs de transcription (TFBS) dans les régions de chromatine ouverte de nombreux participants, et nous aide à découvrir des associations statistiques entre les variations du nombre de TFBS et des variations phénotypiques. Nous avons utilisé cet outil avec le même jeu de données que l'étude des perte-de-fonction et des knockouts, pour découvrir des variants qui influencent des traits sanguins en créant ou détruisant des TFBS.

Les concepts biologiques et bio-informatiques utilisés dans cet article seront introduits dans les chapitres suivants, et nous examinerons le fonctionnement et l'usage du logiciel.

CHAPITRE 2

NOTIONS BIOLOGIQUES

2.1 Concept généraux

Le génome humain contient environ 20000 gènes qui codent pour des protéines et est composé de 3 milliards de paires de nucléotides. Les gènes, qui occupent moins de 2% du génome, sont transcrits en ARN messager grâce à l'aide de plusieurs protéines dont les facteurs de transcription qui déterminent la fréquence de cette transcription. Les transcrits connaissent ensuite une phase de maturation et sont traduits en protéines.

Nous connaissons plus de 150 millions de variations génétiques entre individus, dont certaines modifient les séquences codantes (donc les protéines) et dont certaines changent la fréquence de transcription des gènes.

Ces variations expliquent une partie des différences phénotypiques entre individus, et la découverte d'un nombre croissant de corrélations entre les variations génétiques et phénotypiques ouvre la porte à d'autres expériences qui améliorent notre connaissance des mécanismes moléculaires.

L'identité cellulaire à l'âge des technologies "single cell" est devenue difficile à définir (MORRIS, 2019). Cependant l'étude des lignées et de la différentiation cellulaire (JASON D. BUENROSTRO, 2018) nous montre quels mécanismes moléculaires participent à la formation de cette identité. Parmi ces mécanismes, l'ouverture sélective de la chromatine restreint l'ensemble des gènes qui peuvent être transcrits, et différents marqueurs épigénétiques (marques d'histone, méthylation de l'ADN..) altèrent l'activité des protéines qui régulent l'expression des gènes.

Nous avons accès à des séquences génomiques complètes et "phasées" (i.e nous savons sur quels allèles se trouvent les variants), ainsi qu'à plusieurs phénotypes sanguins, et nous avons accès à des cartes de chromatine ouverte pour plusieurs types cellulaires. Les travaux présentés ici se basent sur ces informations et tentent de découvrir des liens entre un sous-ensemble des variations génétiques et des variations phénotypiques.

2.2 Hématopoïèse, cellules progénitrices et différentiation

Afin de guider la conception de notre expérience sur les facteurs de transcription (TF) et nous aider à interpréter ses résultats, nous allons ici décrire les mécanismes fondamentaux de l'hématopoïèse, qui est le processus de création de cellules sanguines à partir de cellules souches et de cellules partiellement différentiées ("cellules progénitrices").

L'hématopoïèse humaine comporte successivement une phase transitoire dite "primitive" chez l'embryon, suivie par la phase adulte qui est plus complète (JAGANNATHAN-BOGDAN & ZON, 2013). Toutes les cellules du sang sont issues de ce processus et proviennent d'un nombre réduit de cellules progénitrices.

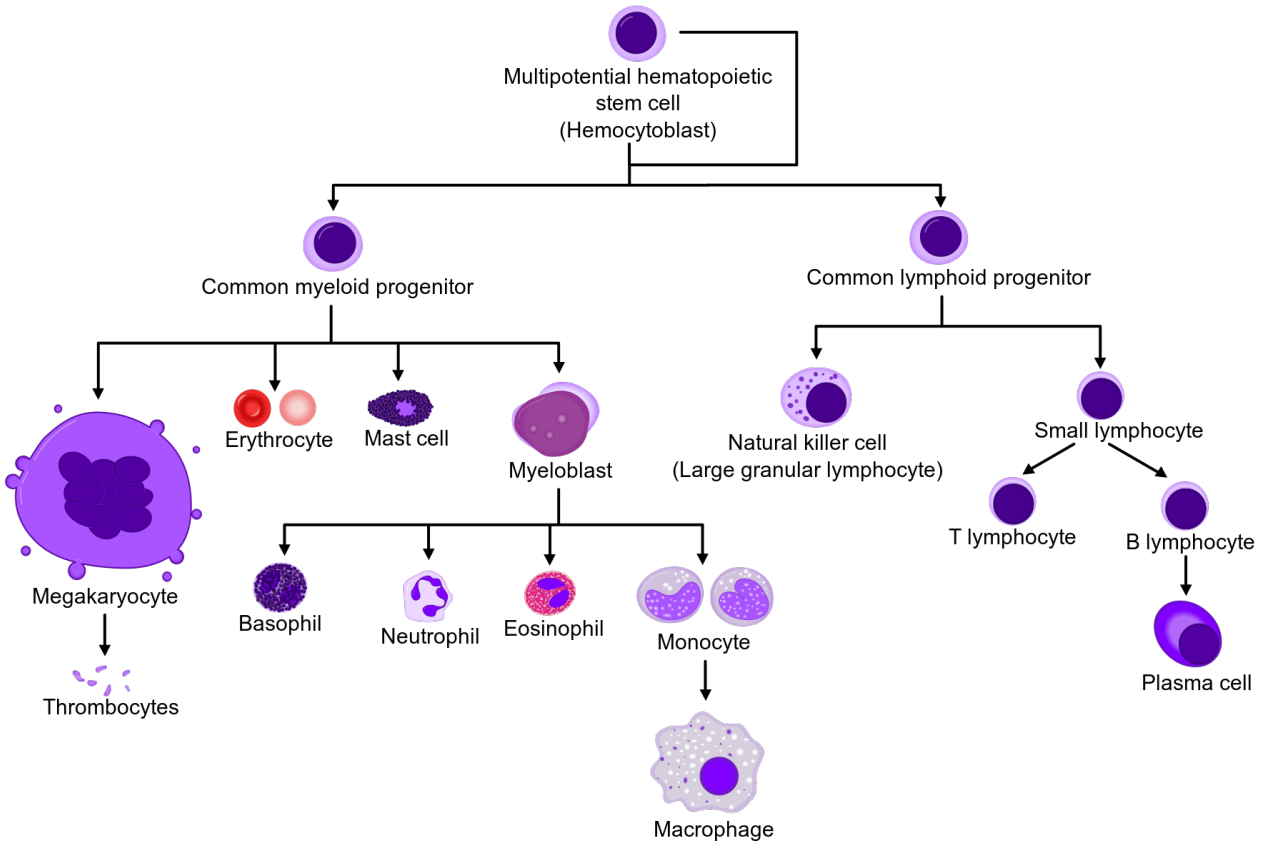


FIGURE 2.1 – Schéma simplifié de l'hématopoïese (A. Rad et M. Häggström. CC-BY-SA 3.0)

Les cellules hématopoïetiques se différencient étape par étape à partir des cellules souches multipotentes. Chaque type cellulaire possède un profil d'expression, des régions de chromatine ouvertes et fermées, un profil de méthylation et un ensemble de facteur de transcription actifs.

Un variant pLoF ou TFBS peut avoir un effet pendant certaines étapes de l'hématopoïèse : si par chance ce variant est situé dans une région de chromatine qui est ouverte spécifiquement dans un seul type de cellule progénitrice, nous pouvons supposer que ce variant intervient dans un processus moléculaire spécifique à ce type cellulaire.

Il est possible d'éditer en laboratoire le génome de cellules progénitrices humaines, et d'étudier par exemple le taux d'hémoglobine des érythrocytes produits. Ce projet vise à découvrir quels variants perte-de-fonction et TFBS induisent une anomalie dans la composition du sang.

2.3 Variants génétiques codants

2.3.1 Variants perte-de-fonction

On prédit qu'un variant génétique est une perte-de-fonction (pLoF) s'il rend inactif au moins l'un des transcrits associés. L'inactivation peut être le résultat de l'insertion d'un codon stop, le déplacement du cadre de lecture par une insertion ou une délétion, ou la modification d'un site d'épissage essentiel (site donneur ou accepteur). Comme expliqué dans la section 4.6.2, les variants qui se trouvent aux extrémités (5%) du transcrit sont ignorés.

Une perte de fonction peut affecter un phénotype de manière positive ou négative.

2.3.2 Variants knockouts

Un knockout (KO) est un gène rendu inactif par l'inactivation de ses deux allèles. Dans le cadre de cette étude, on définit les gènes pKO comme étant les gènes dont les deux allèles sont pLoF. On remarque que la plupart des pKO sont homozygotes pour un variant pLoF, alors que certains sont des hétérozygotes composites. Nous faisons l'hypothèse que la présence d'un knockout peut expliquer une partie du phénotype.

2.4 Transcription et chromatine ouverte

2.4.1 Facteurs de transcription

Les facteurs de transcription sont des protéines qui régulent l'expression des gènes en se liant à des séquences spécifiques dans le génome. On trouve ces séquences dans les promoteurs et dans les amplificateurs (enhancers). Certaines marques d'histones favorisent ou empêchent ces interactions.

Le promoteur est une séquence d'ADN, située en amont du gène et à proximité du site de démarrage de la transcription, à laquelle se lient des protéines et qui démarre le processus de transcription du gène voisin. Un amplificateur est une séquence d'ADN distante à laquelle se lient des protéines et qui augmente la probabilité de transcription d'un gène.

Certains variants affectent les sites de liaison des facteurs de transcription (Transcription Factor Binding Site). La modification ou la création d'un TFBS peut altérer l'expression de gènes en cis, et

parfois en trans. Nous faisons l'hypothèse que la variation du nombre de TFBS dans les régions de chromatine ouverte peut expliquer une partie du phénotype.

2.4.2 Ouverture de chromatine et identité cellulaire

Nous nous intéressons ici aux régions de chromatine ouvertes. Ces régions font partie des élements qui déterminent l'identité cellulaire en sélectionnant l'ensemble des gènes qui peuvent être exprimés dans une cellule. Bien qu'il existe une variabilité dans l'ouverture de la chromatine pour les cellules d'un même type, la connaissance des régions ouvertes est suffisante pour déterminer l'identité des cellules hématopoïetiques (JASON D. BUENROSTRO, 2018). Cette connaissance nous permet aussi de restreindre une recherche de TFBS à des régions pertinentes, où les facteurs de transcription sont physiquement capables de se lier à la chromatine, ce qui réduit le nombre d'hypothèses testées.

Ceci exclut les facteurs de transcription dits "pionniers", qui peuvent interagir avec la chromatine condensée, recruter d'autres facteurs de transcription, des enzymes de méthylation de l'ADN et des enzymes de modification des histones.

CHAPITRE 3

NOTIONS BIO-INFORMATIQUES

Plusieurs notions de bio-informatique, ainsi que les outils informatiques qui permettent de les mettre en oeuvre, sont présentés ici afin de clarifier la description de notre méthode.

3.1 Normalisation des phénotypes

Les valeurs phénotypiques extrêmes peuvent être dues à des erreurs de mesure, des erreurs de saisie ou correspondre à de vrais phénotypes extrêmes (par exemple l'effet d'un knockout, ou une infection qui augmente temporairement le nombre de leucocytes). Si nous ignorons entièrement ces valeurs, l'information des vraies mesures extrêmes sera perdue.

Une alternative, appelée winsorisation, consiste à remplacer les valeurs extrêmes par une valeur fixe correspondant par exemple à la valeur du 95e percentile. Cette alternative corrige partiellement les erreurs de mesure et de saisie, tout en conservant de l'information sur les vraies valeurs extrêmes (FERNÁNDEZ et al., 2002 ; SHETE et al., 2004).

Cependant, l'algorithme d'association EMMAX (section 3.4.4) exige que les phénotypes suivent une distribution normale, et nous avons donc choisi de respecter cette assomption en normalisant les phénotypes avec une transformation inverse de rang (Rank-Based Inverse Normal Transformation) qui prend aussi soin des valeurs extrêmes.

La transformation inverse de rang ordonne les phénotypes par ordre croissant et crée une distribution normale à partir de cet ordre. Comme nos distributions brutes (après correction des covariables) suivent approximativement une distribution normale, la transformation inverse de rang affecte peu les données.

Suite à la normalisation des phénotypes, les effets des variants génétiques seront exprimés en nombre de déviations standards par rapport à la moyenne du phénotype dans la cohorte.

3.2 Biais des cohortes

Les différences entre laboratoires peuvent introduire des biais et augmenter le risque de faux positifs. Par exemple, un laboratoire pourrait générer des valeurs d'hématocrite plus élevées que la réalité. Si les participants dont les échantillons ont été analysés par ce laboratoire sont à majorité hispaniques, une analyse naïve nous laisserait croire que tous les variants communs qui sont spécifiques à la population hispanique sont liés au taux d'hématocrite.

Pour éliminer ce biais, nous corrigeons puis normalisons les phénotypes séparément pour chaque cohorte. Les données issues de cette normalisation sont les données d'entré du test d'association.

Pour des raisons similaires, les différences entre ethnicités introduisent des biais. Nous définissons quatre blocs continentaux (Africains, Asiatiques, Européens et Hispaniques) et séparons les individus de chaque blocs. Les associations statistiques seront données par bloc continental, et nous fournirons une méta-analyse.

3.3 Puissance statistique et contrôle du taux d'erreur

3.3.1 Taille d'échantillons

Comme expliqué par PRICE et al., 2006, "la grande majorité des facteurs de risque pour les traits complexes ont, individuellement, un faible effet sur le phénotype. Les études de populations ont donc besoin de nombreux échantillons pour détecter des différences génotypiques entre individus affectés et non affectés".

De même, des simulations montrent que les études pan-génomiques ont besoin d'un grand nombre d'échantillons pour découvrir de nouvelles associations entre variants et phénotypes (HONG & PARK, 2012). Les variants les plus faciles à découvrir, qui ont probablement déjà été identifiés, sont des variants communs qui ont un effet phénotypique fort.

3.3.2 Contrôle du taux d'erreur dans le cadre d'hypothèses multiples

À cause du grand nombre d'hypothèses testées (chaque variant est une hypothèse statistique), nous nous attendons à observer un certain nombre de faux positifs dans la liste des résultats. Nous préférons contrôler le FWER (Family-wise error rate, probabilité d'avoir au moins un faux positif parmi plusieurs hypothèses) plutôt que le FDR (False discovery rate, proportion de faux positifs) pour chaque phénotype et ethnicité grâce à un ajustement de Bonferroni (section 3.3.2). L'ajustement de Bonferroni nous indique quel seuil de p-valeur rend un résultat crédible.

Dans un modèle statistique fréquentiste, la vraisemblance d'une corrélation peut être quantifiée à l'aide d'une p-valeur, et la tradition recommande un seuil de $p = 0.05$. Le risque de faux positif augmente avec le nombre d'hypothèses testées simultanément, et on s'attendrait à voir par malchance 500 faux positifs pour 10000 tests indépendants et nuls.

Les variations génétiques de cette étude (perte-de-fonction, knockouts et TFBS) sont moins nombreuses que le nombre de SNPs dans les populations étudiées, ce qui augmente notre puissance statistique.

Nous pouvons corriger le seul de p-valeur avec la procédure de Bonferroni $p = 0.05/N$ avec N le nombre de variations génétiques. Cette procédure est conservative, car le déséquilibre de liaison dans le génome rend certaines hypothèses non-indépendantes (BUSH, 2012).

Notons que plusieurs traits sanguins sont corrélés par définition, et que plusieurs TFBS sont en pratique identiques (e.g GATA4 et GATA6). La procédure de Bonferroni est donc trop exigeante.

On conseille d'employer un test de permutations pour définir un seuil de p-valeur empirique qui prendra en compte le déséquilibre de liaison (i.e la corrélation entre chaque variable), tout en ne perdant que peu de puissance statistique. Les tests de permutation estiment la distribution d'une statistique en supposant l'hypothèse nulle. Ici nous pourrions estimer le χ^2, qui teste l'indépendance entre un variant et un phénotype. Cependant, la précision de ce seuil de p-valeur dépend du nombre de permutations testées, et le volume de données traitées rend ce calcul prohibitif au niveau computationnel.

3.4 Tests d'association

3.4.1 Une approche simple : Le modèle linéaire

Les traits complexes, par opposition aux traits mendéliens, peuvent varier de manière continue et être la conséquence de nombreuses variations génétiques. Les traits mendéliens étant plus faciles à découvrir, cette étude se focalise sur des traits polygéniques connus.

En première approximation, on peut supposer que les contributions de chaque gène sont indépendantes et additives (HILL, GODDARD & VISSCHER, 2008), et que l'effet d'un variant est proportionnel au nombre de copies de ce variant (0, 1 ou 2 copies).

Le modèle additif classique se formalise ainsi :

$$Pheno_i = \sum_{v=1}^{V} \beta_v N_{iv} + \varepsilon_i$$

Où i représente l'individu, V le nombre de variants, $N_{iv} \in [0,1,2]$ le nombre de copies du variant, et β_v l'effet du variant. Dans ce modèle, nous cherchons à estimer β_v aussi précisément que possible pour chaque variant v.

ε_i représente ce qui n'est pas expliqué par le modèle et doit être minimisé. Ceci inclut les contributions de l'environnement, et les interactions entre gènes (épistasie) telle que la dominance d'un allèle.

Grâce à l'hypothèse d'additivité, les β_v peuvent être estimés séparément et chaque variant représente une hypothèse statistique. En contraste, les modèles non-additifs doivent estimer un nombre de

paramètres plus important et sont sujets à un plus grand risque de faux positifs.

Dans le cadre des traits polygéniques dans des populations non consanguines, les effets des variants rares et épistatiques (i.e non additifs) se présentent en grande partie sous la forme d'effets additifs (MÄKI-TANILA & HILL, 2014), même si en théorie la condition d'additionalité n'est pas respectée. Un modèle additif est donc souvent suffisant, même si parfois des modèles dits complets (additifs et non-additifs) peuvent expliquer une fraction plus importante de l'héritabilité (MONIR et ZHU, 2017, 6500 personnes). En outre, la détection d'effets épistatiques est imparfaite (SHANG et al., 2011) et demande des ressources computationnelles importantes (WEEKS et al., 2018).

3.4.2 Difficultés communes et concept d'inflation statistique

Le modèle linéaire présenté dans la section 3.4.1 est susceptible de générer des faux positifs, c'est-à-dire des associations statistiques qui ne correspondent pas à de vraies relations causales.

Prenons un exemple : si une douzaine de personnes aux yeux verts ont un taux d'hématocrite nettement supérieur à la moyenne, un test d'association naïf pourrait suggérer que les gènes qui codent pour la couleur des yeux influencent aussi l'hématopoïèse. De manière générale, tous les liens de parenté, qu'ils soient directs (famille) ou lointains (ethnicité) augmentent le risque de faux positifs.

Pour cette raison, les tests d'association ont été effectués séparément pour les quatre blocs continentaux : Africain, Asiatique, Européen et Hispanique. Le bloc hispanique est une agrégation des participants d'origine Mexicaine, Colombienne, Portoricaine etc.

Il existe aussi une structure de population au sein des blocs continentaux. Les dix premières composantes principales du génome sont utilisées comme covariables dans une régression linéaire, et le résidu de cette régression est ensuite corrélé au génotype. Les composantes principales sont calculées avec 150k variants qui ont été choisis pour être en équilibre de liaison, tout en ayant une fréquence allélique suffisante ($MAF \geq 0.01$) et un taux de génotypage supérieur à 95%. Après cette régression linéaire, les résidus expriment la variation du phénotype par rapport au sous-groupe ethnique du participant.

Les liens de parentés sont un défi supplémentaire. Comme nous n'avons pas d'information sur les liens de parentés entre les participants, nous devons inférer ce lien et améliorer le modèle statistique pour le prendre en compte afin d'éliminer cette source de faux positifs. Le modèle choisi est présenté dans la section 3.4.4.

Chaque laboratoire peut introduire un biais de mesure lors du bilan sanguin. Pour cette raison, les phénotypes sont corrigés (sexe, âge, fumeur) et normalisés indépendamment dans chaque cohorte. Le phénotype normalisé représente donc une déviation par rapport à la moyenne de la cohorte.

3.4.3 QQ-plot

Le QQ-plot est une figure qui nous aide à visualiser les résultats d'un test d'association à hypothèses multiples. Elle compare la distribution de nos p-valeurs à une distribution théorique qui correspond à l'absence d'association réelle. Cette distribution théorique contient des p-valeurs faibles si le nombre d'hypothèses testées est élevé ; nous voulons savoir si notre distribution contient davantage de p-valeurs faibles que la distribution théorique du nul ("expected" vs "observed" dans la figure 3.1).

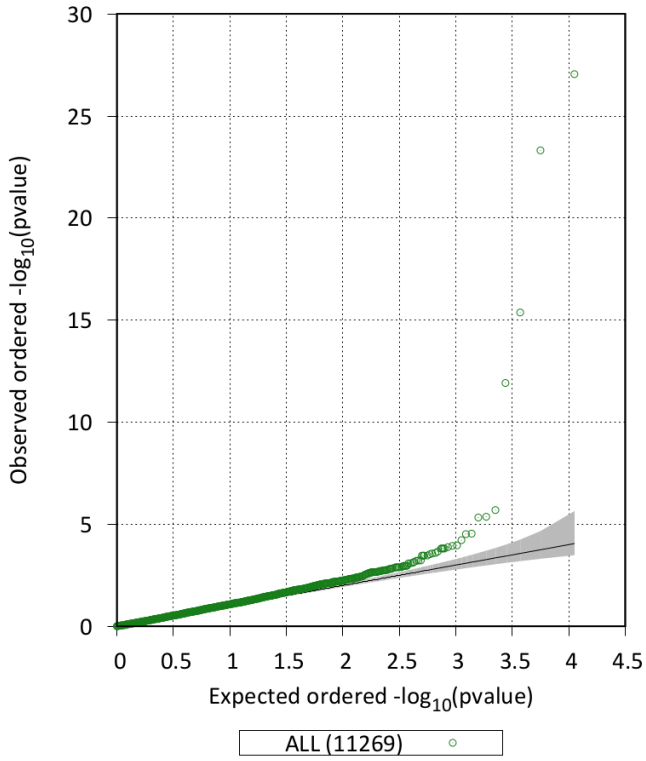


FIGURE 3.1 – Exemple de QQ-plot : Associations entre variants pLoF et le nombre de neutrophiles dans la population Africaine. Un point représente un variant génétique. Les résultats nuls (pas d'association détectée) sont dans la bande grise.

Dans cet exemple, quelques variants sont associés au nombre de neutrophiles, et nous voyons que nos procédures de contrôle de l'inflation statistique sont suffisantes car la plupart des p-valeurs suivent précisement la distribution théorique du nul.

3.4.4 Modèle linéaire mixte

Comme discuté dans la section 3.4.2, la stratification de population et les liens de parenté cachés sont "deux facettes d'un même facteur confondant". Il est possible de contrôler ces deux problèmes avec une procédure unique d'analyse par composantes principales (ASTLE & BALDING, 2010 ; PRICE et al., 2006) et/ou un modèle linéaire mixte (vu plus haut) et qui est considéré supérieur (KANG et al., 2010).

Ce modèle est dit "mixte" car il contient à la fois des paramètres liés à la population dans son ensemble (par exemple : l'effet β_v d'un variant v sur le phénotype) et des paramètres liés aux individus.

L'équation de base du modèle linéaire mixte est la suivante :

$$Pheno_i = \sum_{v=1}^{V} \beta_v N_{iv} + g + \varepsilon_i$$

Le nouveau terme à déterminer g, qui a été ajouté au modèle linéaire décrit plus haut, modélise les effets des relations de parenté. Sa moyenne est nulle et sa variance est $\sigma_g^2 K$ où K est la matrice de parenté.

Pour chaque ethnicité, nous devons créer une matrice de parenté ("kinship matrix") qui contient pour toutes les paires d'individus leur degré de similarité génotypique, exprimé par un nombre à virgule flottante. La taille de cette matrice est donc quadratique avec la taille de la cohorte, ce qui rend les calculs moins rapides qu'un modèle linéaire. L'implémentation EMMAX du modèle linéaire mixte charge cette matrice de parenté et la décompose (complexité algorithmique $O(N_{cohorte}^3)$) pour chaque test d'association, ce qui est souvent l'étape la plus longue pour l'ordinateur. En théorie, cette décomposition pourrait être réutilisée entre chaque test mais l'interface de EPACTS ne le permet pas.

3.5 Variants causaux et proxys

D'un point de vue mécanique, certains variants ont un lien causal réel avec le phénotype. Cependant, les études d'associations découvrent souvent des liens statistiques avec des variants qui ne sont pas causaux mais qui sont présents chez les mêmes individus que les variants causaux.

Ces variant non-causaux sont appelés "proxys". Ils sont hérités en même temps que les variants causaux, et se trouvent donc en déséquilibre de liaison (i.e la co-occurrence du variant causal et de son proxy ne sont pas statistiquement indépendantes).

Pour l'étude présente, des outils ont prédit que les variants analysés ont des conséquences réelles sur la fonction d'un transcrit où sur son expression (pLoF, knockout, TFBS). Ces prédictions sont, évidemment, imparfaites.

3.6 Méta-analyse

La méta-analyse des tests d'associations réalisés dans plusieurs populations permet d'améliorer le degré de certitude pour certains associations lorsque les variants génétiques sont présents chez plusieurs populations et ont un effet similaire sur le phénotype. À l'inverse, la méta-analyse n'offre pas d'information supplémentaire ni pour les variants spécifiques à une population, qui sont nombreux parmi les variants rares, ni pour les variants qui ont des effets opposés chez plusieurs populations. La procédure utilisée pour notre méta-analyse est décrite dans la section 4.9.

3.7 Outils

Les phénotypes ont été collectés, normalisés et présentés avec les logiciels Python (VAN ROSSUM & DRAKE JR, 1995), R (R CORE TEAM, 2013) et les paquets tidyverse (WICKHAM et al., 2019) et ggplot2 (WICKHAM, 2016).

Les logiciels PLINK1.9 (PURCELL et al., p. d.) et FlashPCA (ABRAHAM & INOUYE, 2016) nous ont permis d'extraire les composantes principales des génomes afin de modéliser les structures de population.

Les fichiers de génotypes au format VCF et BCF ont été lus avec vcftools 0.1.15 (DANECEK et al., 2011) et bcftools 1.7 (HENG LI, BOB HANDSAKER, PETR DANECEK, SHANE MCCARTHY & JOHN MARSHALL, 2019).

La méta-analyse des quatre populations a été réalisée pour les perte-de-fonctions et les knockouts grâce à METAL (WILLER, LI & ABECASIS, 2010).

CHAPITRE 4

MATÉRIELS ET MÉTHODES

4.1 Problème et objectif

Trois types de variations génotypiques ont été extraits : des allèles perte-de-fonction, des gènes knockout, et des variations dans les sites de liaison des facteurs de transcription. Le but de l'étude présente est de découvrir des corrélations statistiques entre ces variations et des phénotypes sanguins, ce qui nous aiderait à formuler des hypothèses de recherche pour des études futures.

4.2 Population étudiée

Le projet TOPMed [1] nous offre l'opportunité d'étudier les génomes complets d'un grand nombre de participants. TOPMed a pour objectif de d'accélérer les progrès de la médecine de précision en étudiant les mécanismes fondamentaux des maladies du coeur, du sang, des poumons et du sommeil. Ce mémoire se concentre sur les données sanguines.

Les participants de TOPMed sont divisés en plusieurs cohortes [2], telles que WHI, Amish et CARDIA, qui ont chacune des spécificités. Le tableau 4.I indique le nombre de participants de chaque cohorte, ainsi qu'un détail par ethnicité et par sexe. Toutes les populations hispaniques (Porto Rico, Mexique ..) ont été groupées ensemble.

1. Site web : https://nhlbiwgs.org
2. Page web : https://nhlbiwgs.org/group/project-studies

	Africaine	Asiatique	Européenne	Hispanique	Hommes	Femmes	Total
Amish	0	0	1104	0	559	545	1104
ARIC	1897	0	6210	0	3576	4531	8107
CARDIA	1364	0	1685	0	1324	1725	3049
CHS	691	2*	2789	0	1450	2032	3482
FHS	3*	1*	2686	0	1256	1434	2690
GeneSTAR	575	0	997	0	663	909	1572
HCHS_SOL	0	0	0	7366	3047	4319	7366
JHS	3256	0	0	0	1212	2044	3256
MESA	652	25	1110	732	1219	1300	2519
SAFS	0	0	0	1478	596	882	1478
WHI	1432	203	8988	181	0	10804	10804
Total	9870	231	25569	9757	14902	30525	45427

TABLE 4.I – Nombre de participants dans chaque cohorte TOPMed, et détail par ethnicité et par sexe. Dans certaines cohortes, les individus d'une certaine ethnicité (marqués par un astérisque) ont été retirés de l'étude car ils n'étaient pas assez nombreux pour établir une distribution statistique fiable et pour effectuer un contrôle de qualité.

Comme nous le voyons dans le tableau 4.I, les participants d'origine Européenne sont les plus nombreux, et nous avons une majorité de femmes grâce à la cohorte WHI (Women Health Initiative). Au contraire, les participants d'origine Asiatiques sont peu nombreux, ce qui limite notre puissance statistique (section 3.4). Certaines cohortes sont spécifiques à une ethnicité (Amish, HCHS_SOL, SAFS et JHS).

Pour les études longitudinales (i.e le suivi dans le temps des mêmes participants), TOPMed possède les phénotypes de la même personne à plusieurs années d'intervalle. Nous n'avons utilisé que les données les plus récentes pour chaque participant, car nos méthodes de recherche d'association ne peuvent utiliser qu'un seul jeu de données par personne.

Les données du projet TOPMed sont publiées progressivement. Chaque version est appelée un "Freeze" et contient les données génétiques et phénotypiques de nouveaux participants. Nous utilisons la huitième version des données [3], appelée le Freeze 8.

La relative diversité ethnique de TOPMed est favorable pour la découverte de variants rares et spécifiques aux populations d'origine Africaine et Hispanique. Chaque population possède en outre une structure de déséquilibre de liaison propre, ce qui nous aide à distinguer plus précisément certains

3. Page web : https://www.nhlbiwgs.org/topmed-whole-genome-sequencing-methods-freeze-8

variants causaux de leurs proxys quand ils sont présents dans plusieurs populations. Les populations d'origine africaine sont particulièrement diverses et leurs génomes contiennent moins de déséquilibre de liaison que les autres populations (CAMPBELL & TISHKOFF, 2008). Ceci est dû à l'effet fondateur lors de la migration hors de l'Afrique.

Les participants des cohortes TOPMed et UK Biobank sont en majorité d'origine européenne. Hors les effets phénotypiques forts dans une population minoritaire risqueraient d'être "dilués" par une analyse globale si l'effet phénotypique est faible ou nul dans la population majoritaire (CARLSON et al., 2013). Il est donc nécessaire de faire des études d'association indépendantes pour chaque bloc continental.

4.3 Jeux de données

4.3.1 Génotypes

TOPMed nous a fourni les séquences génomiques complètes et phasées sous la forme de différences par rapport au génome de référence GRCh38 : une liste de polymorphismes mononucléotidiques (SNV), d'insertions et de délétions simples pour chaque participant.

Les variants structuraux (i.e réarrangements de plus de 50 nucléotides) ne sont pas mesurés par TOPMed car les méthodes de détection actuelles manquent de fiabilité et donnent des résultats incomplets, et parce que ces variants ne représentent que 1% des variations du génome (TATTINI, D'AURIZIO & MAGI, 2015). Comme suggéré par TATTINI et al., 2015, les nouveaux outils qui séquencent de long segments d'ADN (e.g Oxford Nanopore) pourraient nous aider à détecter ces variants structuraux.

La couverture moyenne des génomes est 30x, et TOPMed a exclu les variants qui ne passent pas un contrôle de qualité :

— **Seuil de qualité** Seuls les variants dont le score de qualité phred est $QUAL \geq 20$ sont étudiés. En dessous de ce seuil, la probabilité que l'allèle alternatif soit incorrect vaut $10^{-QUAL/10} = 1\%$.

— **Seuil de fréquence** En addition aux contrôle de qualité de TOPMed, nous filtrons les variants trop rares : les pLoF et pKO doivent apparaître au moins trois fois dans la population étudiée et la précision de la dénomination (fraction de génotypes non-absents) doit être supérieur à 0.95. Les pseudo-variants TFBS doivent apparaître au moins 5 fois.

— **Anomalies chromosomiques** Les individus porteurs d'une anomalie sur les chromosomes X et Y ont été exclus.

4.3.2 Phénotypes

Nous étudions ici quinze phénotypes issus du bilan sanguin de plusieurs milliers de volontaires. Ces phénotypes sont des quantités et propriétés des principaux types de cellules sanguines.

Les traits sanguins observés dans cette étude sont des propriétés des cellules produites par l'hématopoïese dans sa phase définitive :

Phénotypes d'érythrocytes :
— HCT : Fraction du volume sanguin composé d'érythrocytes (hématocrite)
— HGB : Concentration d'hémoglobine dans le sang (grammes/dL)
— RBC : Concentration d'érythrocytes ($million/mm^3$)
— RDW : Variation de la taille des érythrocytes ($100 * \sigma_{MCV}/MCV$)
— MCH : Quantité moyenne d'hémoglobine dans un érythrocyte, (picogrammes)
— MCHC : Quantité d'hémoglobine dans la phase d'érythrocytes (grammes/dL)
— MCV : Volume moyen des érythrocytes (femtolitres)

Phénotypes de thrombocytes :
— MPV : Volume moyen des plaquettes (femtolitres)
— PLT : Nombre de plaquettes (milliers par microlitre)

Phénotypes de leucocytes :
— BASO : Nombre de basophiles (milliers par microlitres)
— EOSIN : Nombre d'éosinophiles (milliers par microlitres)
— LYMPH : Nombre de lymphocytes(milliers par microlitres)
— MONO : Nombre de monocytes (milliers par microlitres)
— NEUTRO : Nombre de neutrophiles (milliers par microlitres)
— WBC : Nombre de leucocytes (milliers par microlitres)

Certaines mesures sont corrélées par définition. De plus, nous nous attendons à observer de la pléiotropie (un gène influençant plusieurs phénotypes), comme souvent dans les études pangénomiques (VISSCHER et al., 2017), car l'abondance et d'autres caractéristiques d'une cellule progénitrice peuvent influencer plusieurs types de cellules filles (e.g d'après WENDLING, 1999 la thrombopoïétine stimule la production de mégacaryocytes et de cellules progénitrices CD34+ et CD38-). Exemples de corrélations :
— $MCH \propto HGB/RBC$
— WBC est la somme de tous les leucocytes

Certains phénotypes sont mesurés pour les participants de toutes les cohortes, alors que d'autres phénotypes ne sont mesurés que plus rarement (Figure 4.1). Cette abondance relative impacte la puissance statistique.

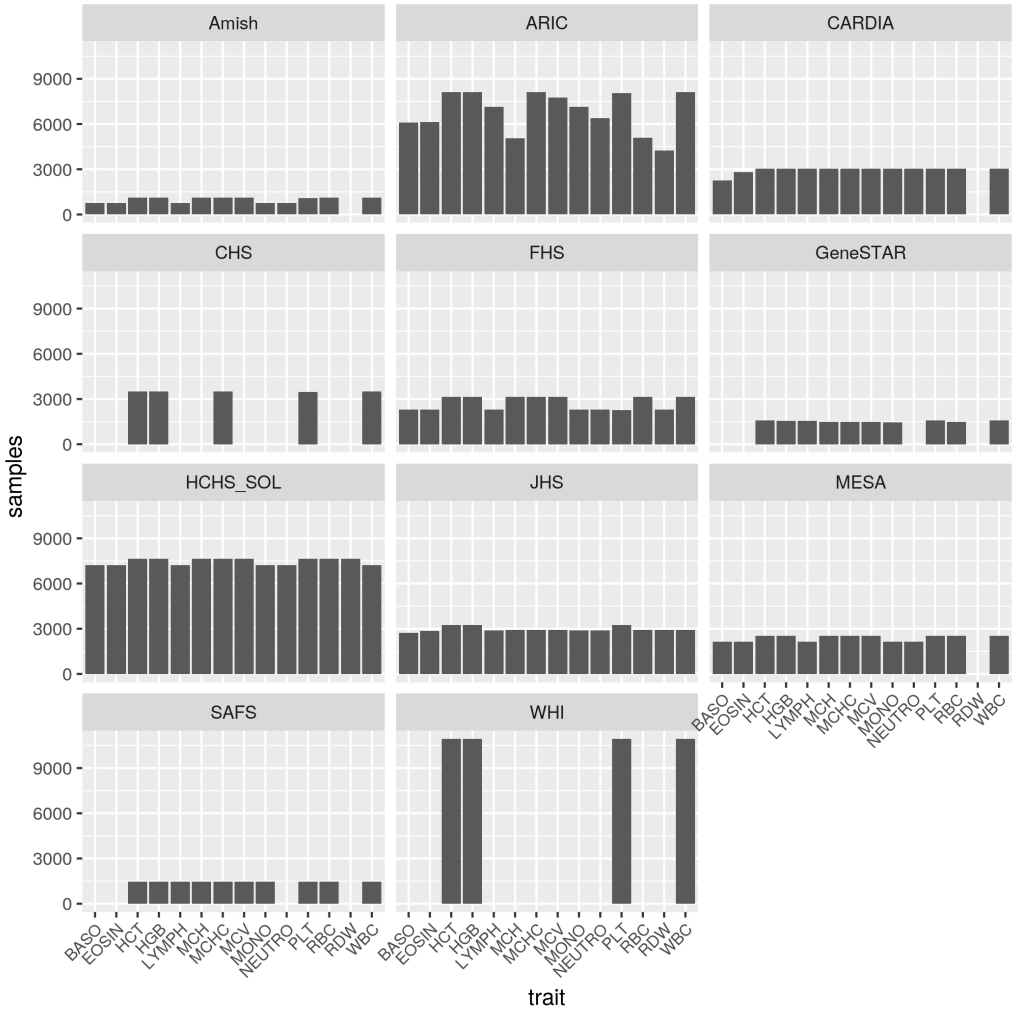


FIGURE 4.1 – Nombre d'échantillons par cohorte (e.g Amish, ARIC..) et par phénotype (e.g BASO, EOSIN..)

Les participants sont en majorité des femmes. Toutes les cohortes sont mixtes, à l'exception du Women Health Initiative (Figure 4.2).

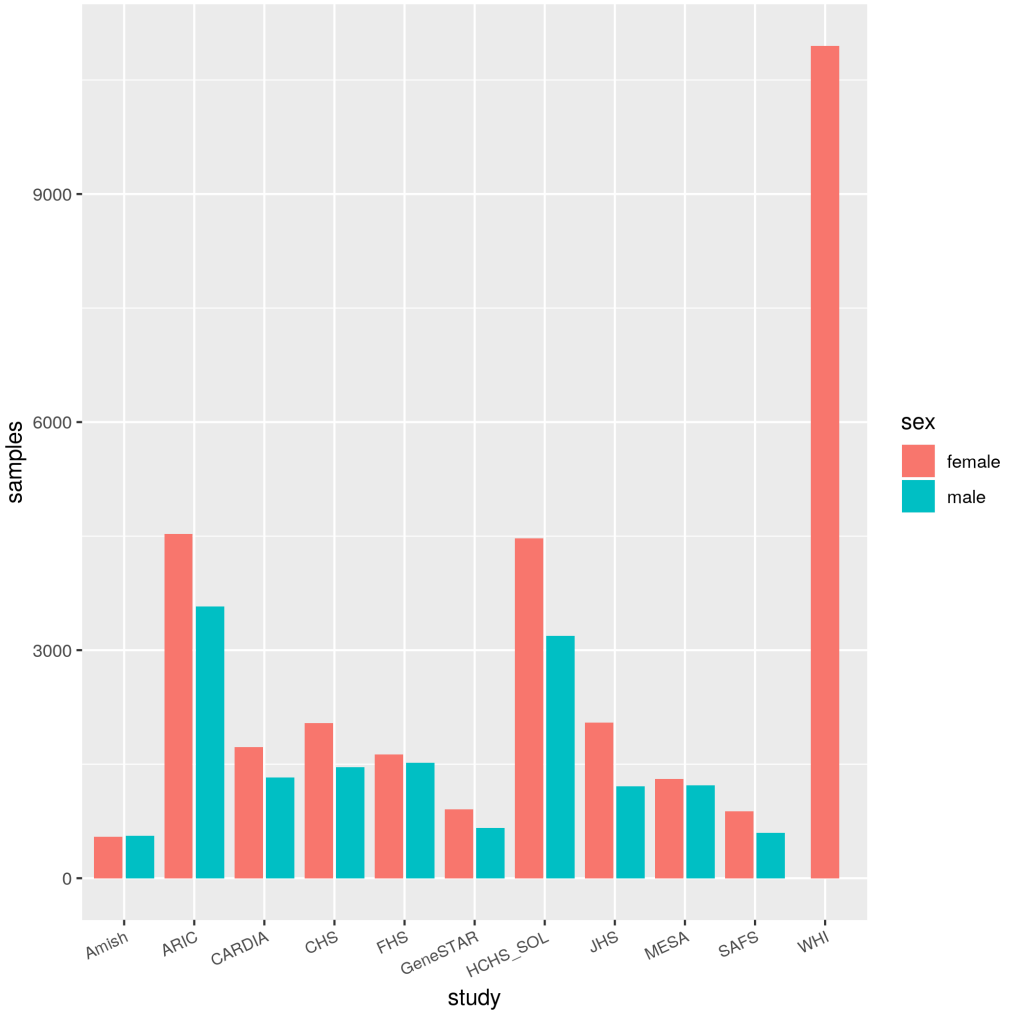


FIGURE 4.2 – Nombre d'échantillons par cohorte et par sexe

4.3.3 Covariables

L'âge des participants au moment de la mesure est disponible. Pour certaines personnes, différents phénotypes ont été mesurés à des moments différents et nous avons les âges correspondant à chaque mesure. Étant donné que les phénotypes sont étudiés indépendamment, ce possible décalage ne cause aucun biais.

Le sexe des participants est disponible pour toutes les cohortes, et la cohorte du Women Health Initiative (WHI) est entièrement féminine.

Pour toutes cohortes à l'exception de Amish, nous savons quels participants sont des fumeurs au moment du bilan sanguin ou dans le passé.

4.3.4 Régions de chromatine ouverte

Les cellules hématopoïetiques sont caractérisées par différentes régions de chromatine ouverte, dont certaines sont communes entre plusieurs types cellulaires ou se chevauchent. Nous avons utilisé les coordonnées de CORCES et al., 2016 et sélectionné les régions qui sont clairement ouvertes (seuil arbitraire de qualité de 0.8).

Ces régions correspondent aux cellules suivantes :
— Leucocytes : Bcell, CD4, CD8, Mono, Nkcell, mDC, pDC
— Cellules progénitrices : HSC, LMPP, CLP, CMP, MPP, GMP, MEP
— Erythroblastes
— Mégacaryocytes

Les chevauchements des régions de plusieurs types cellulaires facilite leur analyse simultanée. Ceci est particulièrement important car la lecture des fichiers BCF consomme des ressources importantes. L'ensemble du jeu de données génotypiques de TOPMed pèse 781 gigabytes après compression.

Type cellulaire	Nombre de régions de chromatine ouverte	Total (bp)
Bcell	225570	81478558
CD4	149519	60474945
CD8	131202	57396762
CLP	247623	80911653
CMP	207782	94673626
Erythro	272723	75238197
GMP	211739	86727810
HSC	173029	83410327
LMPP	171477	72310475
mDC	156239	60491817
MEGA1	172180	61893614
MEGA2	185308	62487000
MEP	203875	90897562
Mono	209782	64261765
MPP	179870	84764702
Nkcell	149636	61499433
pDC	189111	53942122
Unifiés	1003602	368.588.354

TABLE 4.II – Nombre de régions de chromatine ouverte, par type cellulaire et après unification

4.3.5 Définitions des motifs des TFBS

Les TFBS sont peuvent être imparfaitement prédits in silico par une recherche de motifs d'ADN. Les facteurs de transcription sont aussi sensibles à plusieurs autres élements (méthylation de l'ADN, marques d'histones) auxquels nous n'avons pas accès pour les participants de TOPMed.

Les PWMs (Position Weight Matrix) sont une méthode courante de représentation de ces motifs d'ADN. Un PWM attribue un score à chaque nucléotide tout au long d'une séquence. Si la somme de ces scores dépasse un certain seuil (calculé pour ce PWM spécifique), nous déclarons que la séquence analysée est un site de liaison potentiel.

Les définitions des PWM et leurs seuils respectifs sont collectés dans plusieurs bases de données. Nous avons utilisé les définitions de HOCOMOCO [4] version 11 (KULAKOVSKIY et al., 2018, D1).

Comme ces scores de nucléotides sont indépendants les uns des autres, le PWM n'est pas capable d'encoder des informations de séquences complexes. D'autres formes de motifs (diPWM, chaînes de Markov, réseau neuronaux) offrent des encodages plus riches et une reconnaissance plus précise des TFBS (AVSEC et al., 2020 ; SIEBERT & SÖDING, 2016). Nous n'avons pas utilisé les méthodes les plus avancées car il n'existe pas encore une base de données suffisamment exhaustive pour nos besoins.

(a)

(b)

FIGURE 4.3 – Les logos (a) et (b) représentent les PWMs des facteurs de transcription GATA1 et GATA2. Cet exemple a été choisi pour illustrer la ressemblance de certains motifs. À cause de ces ressemblances, plusieurs facteurs de transcription peuvent entrer en compétition pour un site donné.

4.3.6 Liste des facteurs de transcription

Une liste de facteurs de transcription liés à l'hématopoïese a été créée à l'aide d'une revue de littérature, et nous avons collecté les motifs PWM correspondants dans la base de données HOCOMOCO (KULAKOVSKIY et al., 2018, D1). Par construction, cette liste est restreinte aux facteurs

4. Site web : https://hocomoco11.autosome.ru

de transcription qui possède un domaine qui se lie à la chromatine et qui sont présents dans HOCOMOCO. Voici la liste des ressources utilisées : LETTRE et al., 2008, SIU, WURSTER, LIPSICK et HEDRICK, 1992, KATSUMURA, DEVILBISS, POPE, JOHNSON et BRESNICK, 2013, BAO et al., 2016, LE SAOUT et al., 2017, ZHU, THOMAS et HEDRICK, 2016, VAGAPOVA, SPIRIN, LEBEDEV et PRASSOLOV, 2018, MEDVEDOVIC, EBERT, TAGOH et BUSSLINGER, 2011, ITOH-NAKADAI et al., 2017, SIMONE MESMAN et MARTEN P. SMIDT, 2017, KOHU et al., 2009, BATISTA, LI, XU, SOLOMON et DEKOTER, 2017, HANNA et al., 2011, GAUTAM et al., 2019, ZHUANG, CHENG et WEINTRAUB, 1996, GRUSDAT et al., 2014, NCBI - TCF12 gene [5].

Pour un phénotype et facteur de transcription donnés, la literature contient des informations plus ou moins précises. Parfois, nous savons précisement dans quel type cellulaire un mécanisme moléculaire peut prendre place, et les tests d'association peuvent donc se restreindre aux régions de chromatine ouverte de ce type cellulaire. Dans d'autre cas, le type cellulaire est inconnu, et nous devons effectuer des tests d'association dans les régions de chromatine ouverte de toutes les cellules progénitrices correspondantes.

Nous avons validé la liste de facteurs de transcription par type cellulaire grâce aux données de JASON D. BUENROSTRO, 2018. Ces données proviennent d'expériences ChIP-seq touchant un grand nombre de facteurs de transcription dans plusieurs types de cellule hématopoïetiques, et du calcul de deux scores :

— L'abondance du nombre de TFBS comparé à un bruit théorique aléatoire, appelé "score d'enrichissement", calculé par type cellulaire et par TF. Un score élevé suggère que le TF a une fonction dans ce type cellulaire, et un score faible suggère que l'expression de certains gènes doit être minimisée dans cette lignée cellulaire. Nous avons vérifié que la distribution de ce score est positive et significativement différente du nul (t.test$(scores, \mu = 0) \geq 10^{-7}$ and $\mu_{scores} > 0$) pour au moins un type de cellule hématopoïetique.

— La variance de ces scores entre type cellulaire. Une variance élevée suggère l'existence de mécanisme cellulaires spécifiques

La figure 4.4 illustre la distribution du score d'enrichissement pour quelques TF tels que GATA1. Les TFBS de GATA1 sont significativement enrichis (distribution décalée vers la droite) dans les cellules MEP, progénitrices des érythrocytes, ce qui est en accord avec la littérature. Au contraire, les TFBS de GATA1 sont rares dans les monocytes, ce qui suggère que certains gènes doivent être réprimés dans ce type cellulaire.

5. https://www.ncbi.nlm.nih.gov/gene/6938

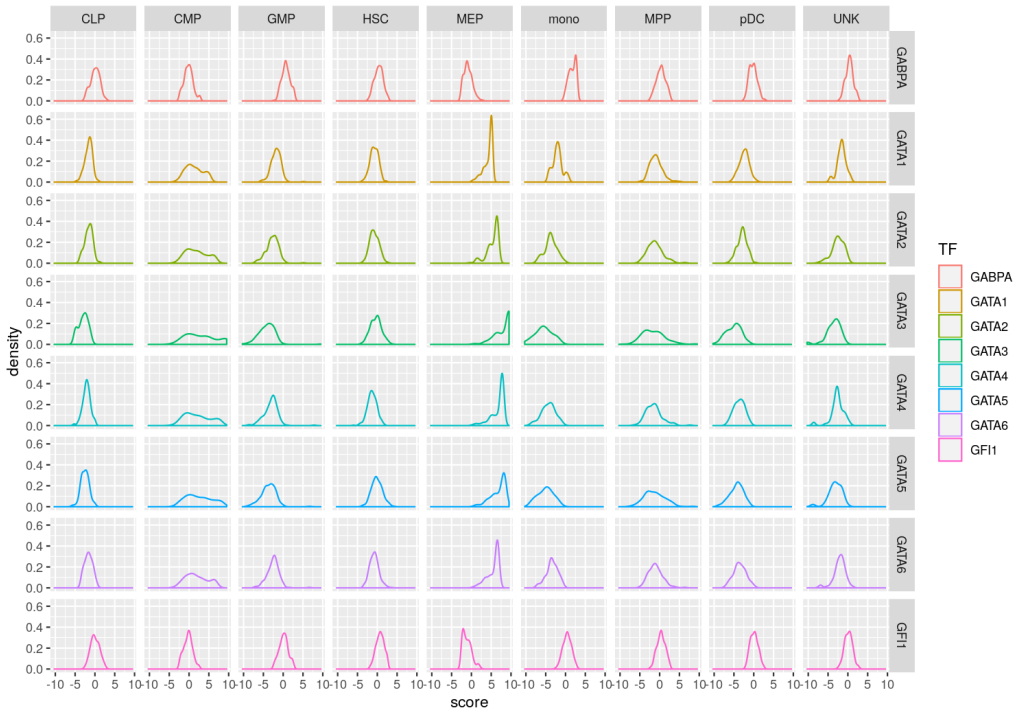


FIGURE 4.4 – Enrichissement de plusieurs TF par type cellulaire (adapté de Buenrostro et al.)

4.3.7 Sélection des phénotypes par type cellulaire

Pour découvrir des associations, nous avons besoin d'étudier les TFBS dans les régions de chromatine ouverte des cellules sanguines correspondantes et de toutes leurs cellules progénitrices. Les autres combinaisons (par exemple un test d'association entre le nombre d'érythrocytes et une région ouverte exclusivement dans les monocytes n'aurait pas de sens a priori et augmenterait inutilement le nombre d'hypothèses testées).

Phénotypes et cellules progénitrices des érythrocytes :
— Types cellulaires : Erythro, CMP, MEP, MPP, HSC
— Traits : HGB, HCT, MCH, MCHC, MCV, RBC, RDW

Phénotypes et cellules progénitrices des thrombocytes :
— Types cellulaires : MEGA1, MEGA2, MEP, CMP, MPP, HSC
— Traits : MPV, PLT

Phénotypes et cellules progénitrices des leucocytes :

	Basophiles	Éosinophiles	Lymphocytes	Monocytes	Neutrophiles	WBC
pDC						x
mDC						x
Mono				x		x
CD4			x			x
CD8			x			x
CLP			x			x
GMP	x	x		x	x	x
CMP	x	x		x	x	x
LMPP	x	x	x	x	x	x
MPP	x	x	x	x	x	x
HSC	x	x	x	x	x	x

TABLE 4.III – Phénotypes et cellules progénitrices des leucocytes

4.4 Validation des ethnicités déclarées

Les ethnicités des participants ont été validées à l'aide d'une analyse par composante principale (PCA) du génome. Les composantes principales ont été calculées à partir de 150k variants choisis pour être en équilibre de liaison et distribués sur tous les chromosomes homologues. La figure 4.5 présente les deux premières composantes principales (PC1 vs PC2) des participants de TOPMed, et les points sont colorés par bloc continental.

La population Hispanique est le résultat du mélange des peuples préhispaniques d'origine Asiatique avec des populations Européennes et Africaines, ce qui se traduit dans la figure 4.5 par des points intermédiaires entre ces trois blocs.

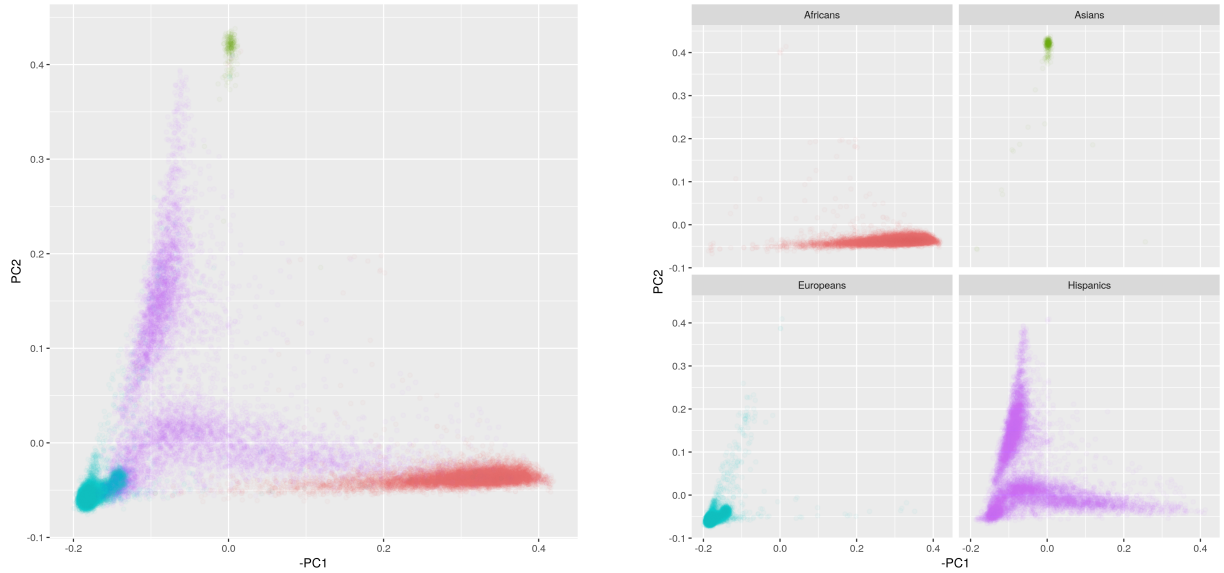


FIGURE 4.5 – PCA des génotypes, par ethnicité. Un point représente un participant, et les couleurs représentent les ethnicités

Cette distribution est familière et cohérente avec les ethnicités déclarés par les participants : nous pouvons comparer cette distribution avec celle du "1000 genomes" (Figure 4.6).

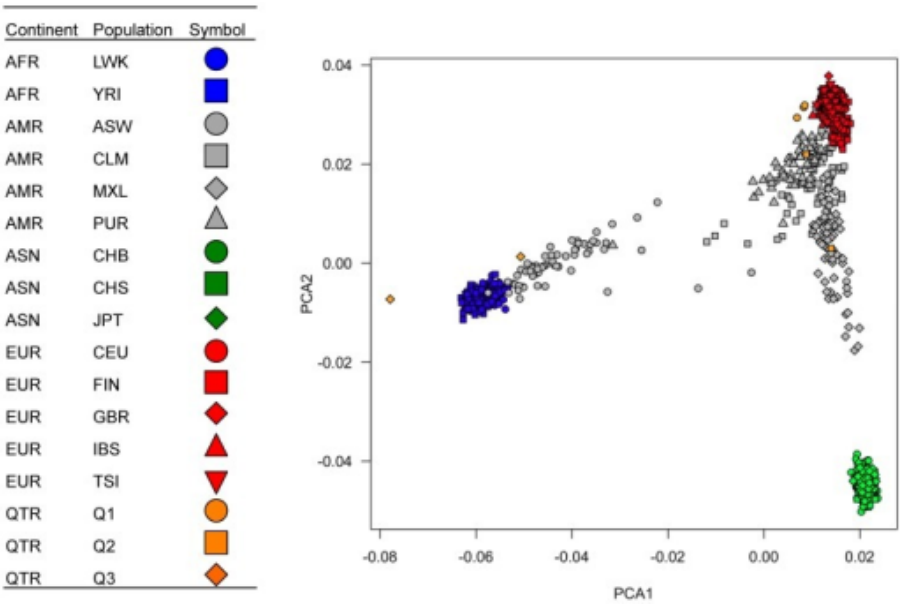


FIGURE 4.6 – PCA des génotypes, 1000 genome

Les dix premières composantes principales sont utilisées pour réduire le risque de faux positifs dû à la structure de population dans chaque bloc continental. Cette procédure est décrite dans la section 3.4.2.

4.5 Correction et normalisation des phénotypes

L'âge, le sexe et le statut de fumeur expliquent une partie du phénotype. Nous devons donc le corriger avant de commencer une étude d'association. La correction est réalisée avec une régression linéaire standard :

$$phenotype = \mu + a * age + b * age^2 + c * sexe + d * S + \varepsilon$$

Dans cette formule :

— μ est la valeur moyenne du phénotype brut dans la cohorte, par exemple la moyenne du nombre d'érythrocytes

— a,b,c et d sont des paramètres à ajuster. Par exemple, si $c = 0$ le phénotype est indépendant du sexe

— $S = 1$ pour les fumeurs et $S = 0$ pour les non-fumeurs

— ε est le phénotype corrigé, dont la distribution a une moyenne nulle, et qui sera étudié par la suite

Quelques échantillons des cohortes Framingham et CHS ont été ignorés, car le nombre de participants d'origine Asiatique était trop faible pour estimer l'effet des covariables.

Les figures 4.7 et 4.8 illustrent le rôle des covariables principales (age, sexe, fumeur) sur la variance des phénotypes. Ces trois covariables expliquent une proportion importante (~20%-50%) de la variance pour les phénotypes liés à l'abondance des érythrocytes (HCT, HGB et RBC, en rose) dans toutes les populations (Figure 4.7), alors qu'elles n'expliquent qu'une moindre proportion de la variance (environ 0%-15%) pour les autres traits :

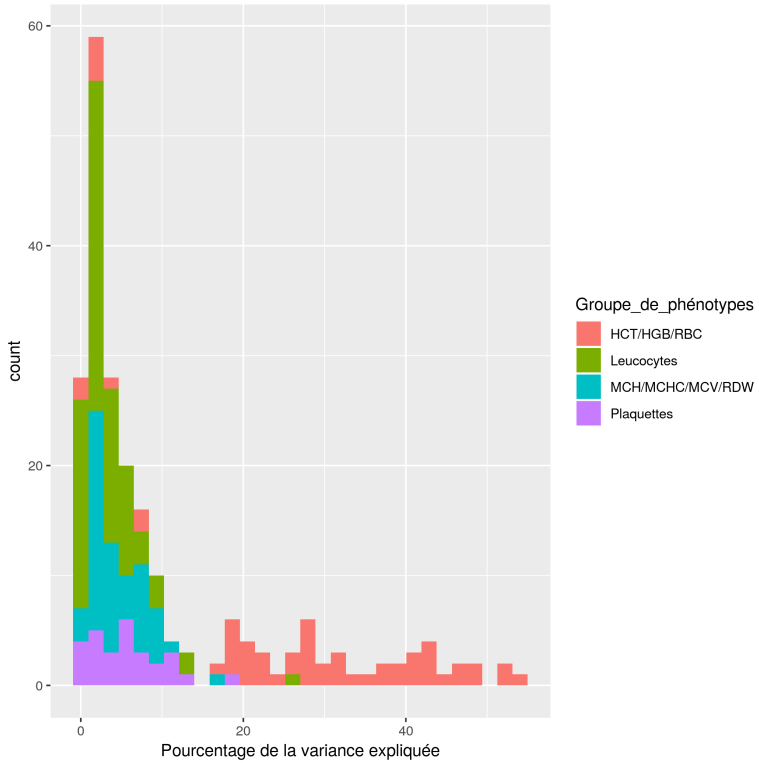


FIGURE 4.7 – Pourcentage de la variance du phénotype (par cohorte) qui est expliquée par sexe+age+fumeur. Les phénotypes liés à l'abondance d'érythrocytes sont colorés en rose et sont largement expliqués par ces trois covariables. Au contraire, les autres phénotypes d'érythrocytes, les phénotypes de plaquettes et de leucocytes sont relativement indépendants de ces covariables. Les cohortes ne sont pas visuellement différenciées dans cette figure.

La figure 4.8 présente les mêmes données sous un angle différent : au lieu de colorer les phénotypes nous colorons la cohorte. Cette représentation montre que les phénotypes de la cohorte WHI (en rose, composée uniquement de femmes) sont peu corrélés aux trois covariables, ce qui illustre que le sexe est la covariable principale.

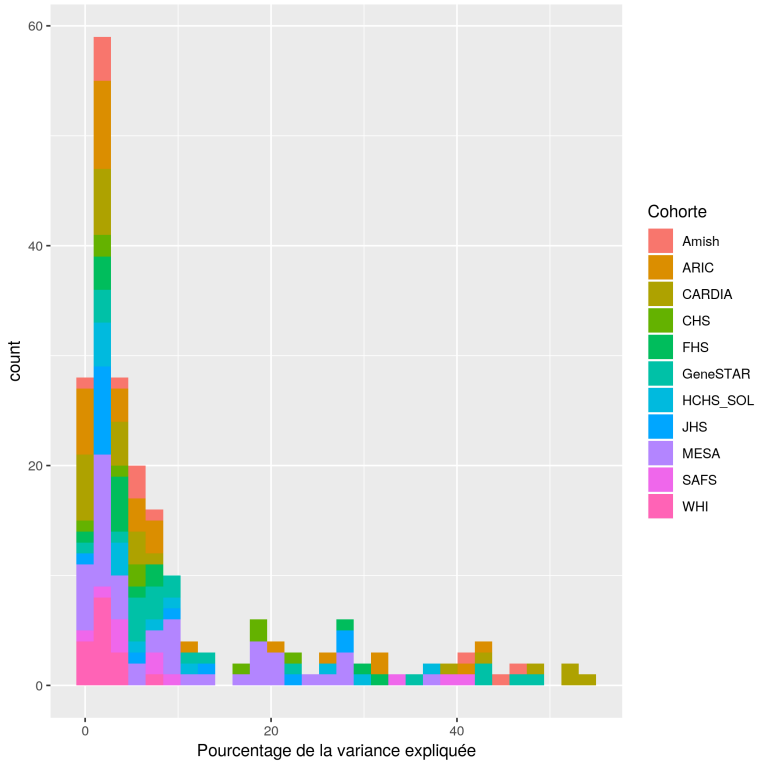


FIGURE 4.8 – Pourcentage de la variance du phénotype (par cohorte) qui est expliquée par sexe+age+fumeur. Les cohortes sont visuellement différenciées, et les phénotypes ne sont pas visuellement différenciés.

Notons que due à sa moindre taille, l'estimation des paramètres de la régression linéaire (a,b,c,d) est moins précise pour la population d'origine Asiatique, et que cette correction réduit la puissance statistique.

Comme illustré dans la figure 4.9, les données de chaque bloc continental sont analysées indépendamment et les données de chaque cohorte sont corrigées et normalisées indépendamment.

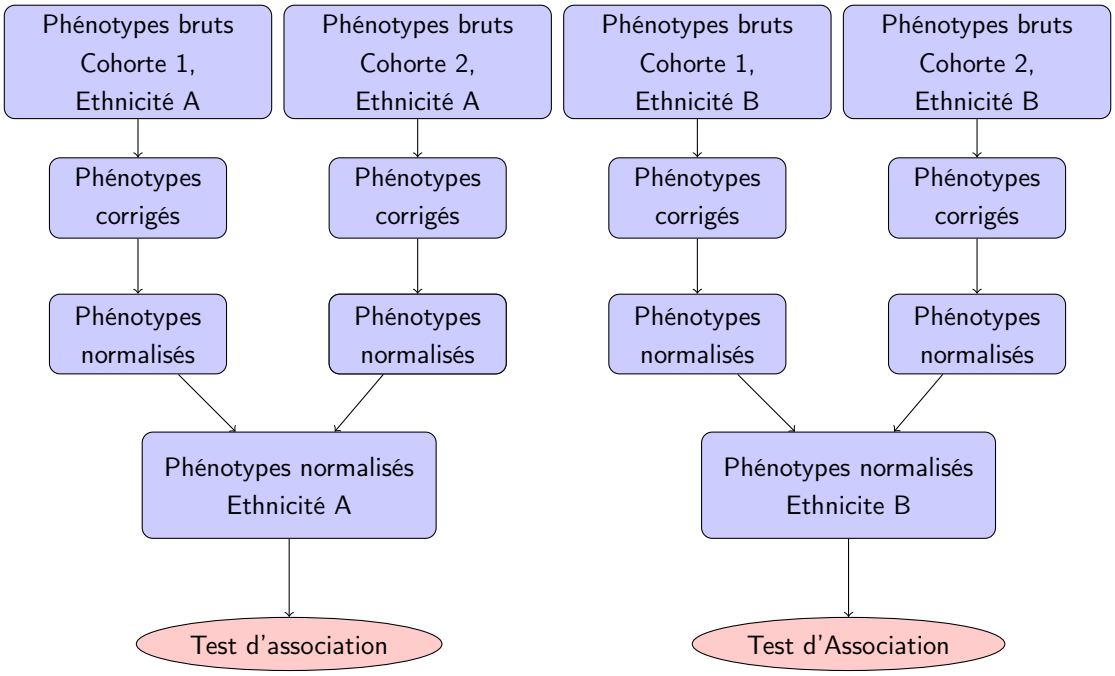


FIGURE 4.9 – Flot de données pour la preparation d'un phénotype

4.6 Recherche de variants perte-de-fonction

4.6.1 Annotations

Pour créer une liste de variants pLoF, nous avons utilisé les annotations de trois outils pour étudier les variants codants et les variants des sites d'épissage essentiels :

— VEP (Variant Effect Predictor, MCLAREN et al., 2016)
— SnpEff (CINGOLANI et al., 2012)
— ANNOVAR (WANG, LI & HAKONARSON, 2010)

La décision d'utiliser plusieurs outils d'annotation a été motivée par le manque d'exhaustivité de chaque outil pris séparemment et par leurs incohérences occasionnelles. VEP contient la base de données de transcrits et d'annotations pLoF la plus complète, et nous a permis de classifier la majorité des variants TOPMed. De plus, pour les variants/transcrits qui ne sont pas connus uniquement par VEP, les trois outils d'annotation étaient en général (mais pas toujours) cohérents à propos de la classification pLoF/non-pLoF.

Dans VEP et SnpEff, nous avons sélectionné les annotations suivantes dont on suppose qu'elles rendent la protéine inactive et causent une perte de fonction :

— *stop_gained* : Un variant crée un codon stop, qui tronque la protéine
— *splice_acceptor_variant* et *splice_donor_variant* : Un variant affecte l'épissage de la molécule d'ARN messager après la transcription
— $frameshift\_variant$: Une insertion ou une déletion décale le cadre de lecture et change tous les amino-acides dans le reste de la protéine

Dans ANNOVAR, nous avons sélectionné les annotations *stopgain* et $frameshift$. Les annotations de ANNOVAR sont moins précises et ne donnent pas d'information sur les variants des sites d'épissage essentiels.

Étant donné que "the choice of transcript set can have a large effect on the ultimate variant annotations obtained in a whole-genome sequencing study" (McCarthy et al., 2014), nous avons examiné certains variants pour comparer les prédictions des trois outils. Cet examen a confirmé l'observation de McCarthy et al., 2014 : certaines annotations sont erronnées, et dans certains cas ces erreurs affectent la classification pLoF/non-pLoF. Comme nous privilégions l'exhaustivité à la précision, nous avons sélectionné les variants pour lesquels au moins l'un des transcrits est un pLoF d'après au moins l'un des outils d'annotation.

L'ambiguité de certaines annotations n'est pas problématique : par exemple, deux outils différents peuvent classifier un variant donné comme $frameshift\_variant$ et *stop_gained* : les deux annotations sont correctes, et sont suffisantes pour nos besoins.

Les annotations des variants et les identifiants des transcrits d'ARN messager sont basés sur la base de données RefSeq (Pruitt, Tatusova & Maglott, 2005, suppl_1). Seuls les variants qui affectent des transcrits observés (préfixés par "NM") sont été pris en compte, à fin de réduire le nombre d'hypothèses testées : les transcrits prédits de manière informatique (préfixés par "XM") n'ont pas été utilisés.

Après avoir examiné les incohérences des trois outils, nous avons ajouté la règle suivante : les variants pLoF "faible confiance" de VEP ont été exclus, sauf si SnpEff ou ANNOVAR les classifient aussi comme pLoF. Le champ "Haute/Faible confiance" indique le degré de certitude de VEP au sujet d'un variant

4.6.2 Exclusion des variants situés aux extrémités des transcrits

On observe que les variants pLoF ne sont pas répartis de manière uniforme sur la longueur des transcrits qu'ils affectent. La figure 4.10 présente la distribution de cette position, entre 0 (début du transcrit) et 1.0 (fin du transcrit). Lorsqu'un variant affecte plusieurs transcrits, chaque position est incluse dans cette figure.

On constate un enrichissement de variants pLoF aux extrémités, ce qui suggère que ces variants sont moins délétères et subissent une sélection négative moins intense. Pour cette raison, les variants pLoF situés aux extrémités (5%) des transcrits sont ignorés.

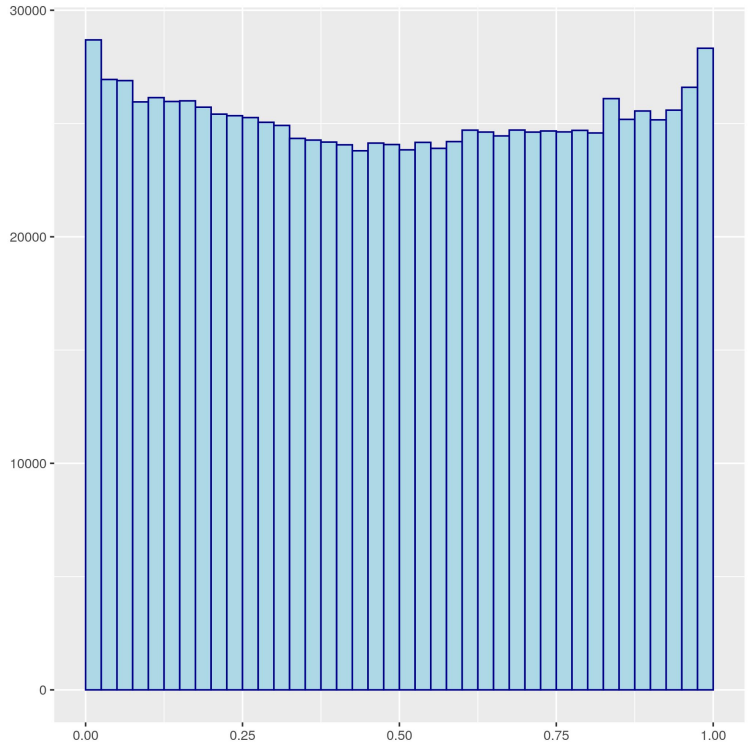


FIGURE 4.10 – Position relative des variants dans le transcript (0=début,1=fin)

4.6.3 Score de sévérité CADD

Nous avons aussi tenté d'utiliser les information du logiciel CADD (RENTZSCH, WITTEN, COOPER, SHENDURE & KIRCHER, 2018). CADD estime un score de sévérité pour chaque variant grâce à plusieurs heuristiques et bases de données (score de conservation, région codante vs région non codante..). Étonnamment, nous avons découvert que la sévérité des variants $frameshift$ et $stop\_gained$ suit une distribution bimodale (Figure 4.11) où les $frameshift$ (bleu) ont un score de sévérité plus faible que les $stop\_gained$ (rouge), alors que ces deux types devraient causer une inactivation du transcrit. Pour cette raison, le score CADD n'a pas été utilisé dans cette étude. Les variants des sites d'épissage sont plus rares et ne sont pas visibles dans cette figure.

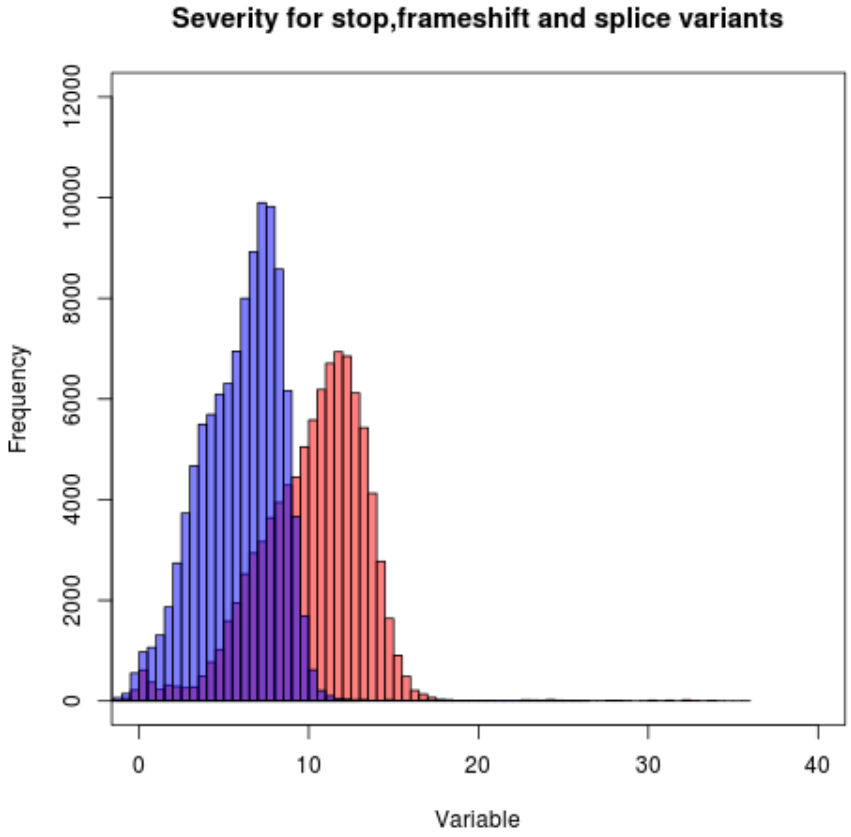


FIGURE 4.11 – Score de sévérité CADD par type de variant. Rouge=codon stop, Bleu=décalage du cadre de lecture

4.6.4 Nombre de variants pLoF retenus

Le nombre de variants pLoF retenus et leur origine sont illustrés par la figure 4.12 :

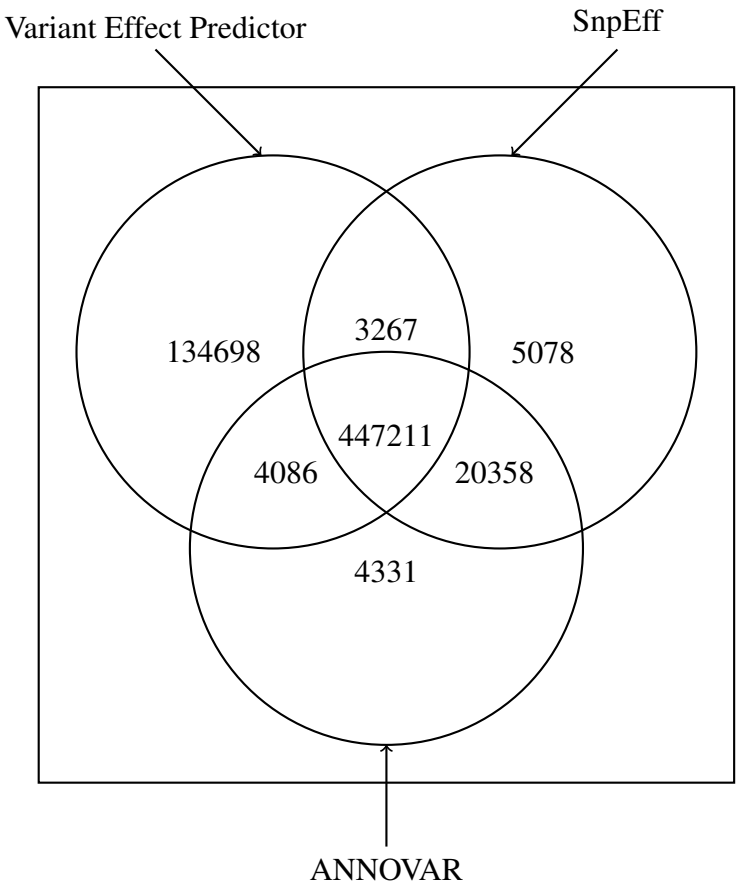


FIGURE 4.12 – Nombre de variants pLoF identifiés par les outils d'annotation de variants

4.7 Recherche de knockouts

Pour le même ensemble d'individus, la liste des knockouts a été construite à partir de la liste des variants pLoF de chaque gène. L'étude est divisée par ethnicité pour la même raison. Chaque knockout est modélisé comme un variant génétique fictif. Notons que les knockouts peuvent être dus à plusieurs variants pLoF, et pas seulement à la présence d'un variant pLoF homozygote, et que nous utilisons des génotypes phasés.

Le fichier résultat (format de fichier VCF, DANECEK et al., 2011) contient une ligne par gène, et la présence d'un knockout chez un individu est représentée par "0|1" alors que l'absence de knockout est représentée par "0|0". Ce choix d'encodage nous permet d'utiliser le même pipeline de tests d'associations afin de contrôler efficacement le risque de faux positifs dû à la structure de population et aux relations de parenté cryptiques.

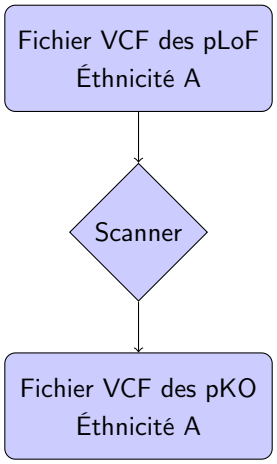


FIGURE 4.13 – Flot de données lors de la recherche de knockout

Le programme de recherche des knockouts a été écrit dans le language de programmation Haskell (MARLOW, 2010). Afin de fonctionner avec une quantité de mémoire limitée, ce programme parcours chaque gène en séquence et crée pour ce gène une table de hachage, indexée par l'identifiant de l'individu, qui contient la liste de ses variants pLoF et leurs phases. Ainsi, seules les données d'un gène sont conservées en mémoire à chaque instant.

Ce programme find-knockouts peut être téléchargé et installé à partir de son dépôt Github : https://github.com/Helkafen/find-knockouts.

4.8 Flot de travail

Le phénotype normalisé a été analysé par le modèle linéaire mixte EMMAX, implémenté par le logiciel EPACTS, avec l'aide de la matrice de parenté de l'ethnicité étudiée. EMMAX est capable de modéliser les liens de parenté cryptiques dans la population étudiée, ce qui minimise le taux de faux positifs (ASTLE & BALDING, 2010). La matrice de parenté a été calculée une fois pour chaque ethnicité.

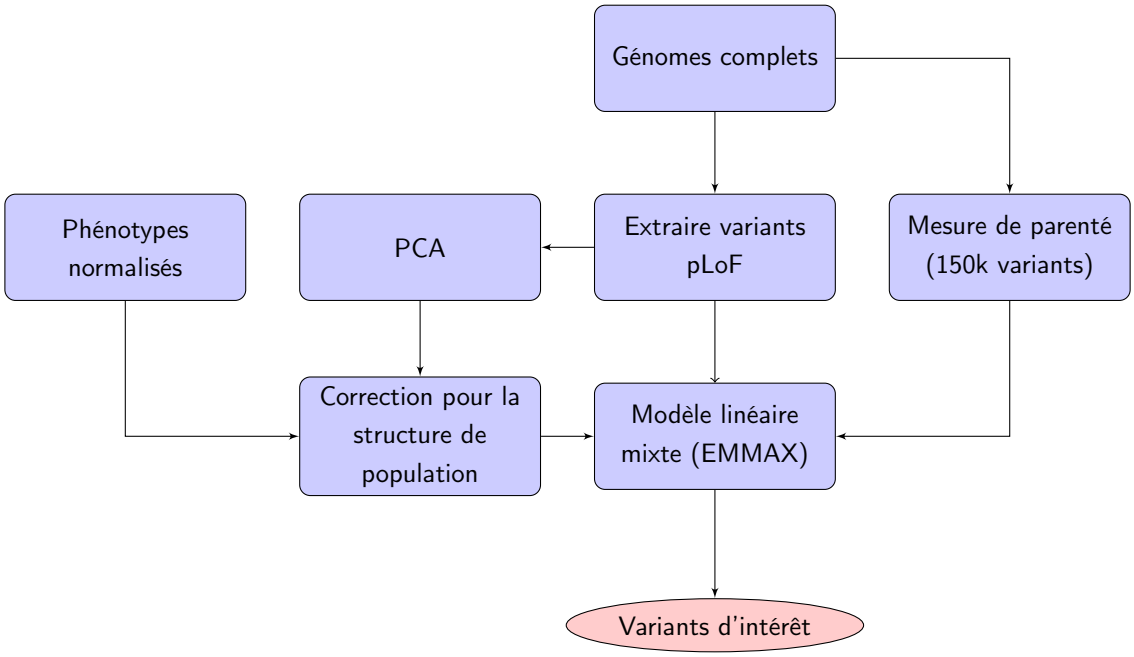


FIGURE 4.14 – Flot des tests d'associations

4.9 Méta-analyse

Nous avons réalisé une méta-analyse pour consolider les résultats d'associations des variants perte-de-fonction et des knockouts dans les quatre blocs continentaux. Les outils de méta-analyse agrègent les résultats résumés (p-valeur et taille de l'effet β) de plusieurs études d'association pour mettre en valeur des variants aux effets modérés présents dans plusieurs populations. Ceci est particulièrement utile dans le cas où des origines éthniques différentes auraient été un facteur confondant (EVANGELOU & IOANNIDIS, 2013). Cette méthode perd peu de puissance statistique par rapport à une étude qui se base sur les données individuelles (LIN & ZENG, 2010), et est plus rapide à exécuter.

Nous avons utilisé le mode STDERR du logiciel METAL (WILLER et al., 2010), dans lequel il se base sur les estimations des β (et leur erreur standard) issues des analyses d'association, sur les p-valeurs et sur la taille des échantillons. METAL nous offre une nouvelle estimation des β et de nouvelles p-valeurs. Le poids attribué aux cohortes est proportionnel à l'inverse de la variance du β.

Nous avons aussi observé de manière qualitative les associations d'un variant d'intérêt avec plusieurs phénotypes biologiquement corrélés. Par exemple, un variant qui affecte plusieurs cellules myéloïdes est probablement impliqué dans le fonctionnement des CMP (cellules progénitrices myéloides) ou de leurs progénitrices.

4.10 Recherche des sites de liaison des facteurs de transcription

En utilisant les motifs PWM décrits précédemment (section 4.3.5), nous pouvons prédire les positions des TFBS à partir des fichiers de génotype et observer les associations statistiques entre les variations du nombre de TFBS et les traits sanguins. Cette approche nous permet de formuler des hypothèses à propos de certains variants non-codants, ce qui est généralement une tâche difficile.

4.10.1 Régions de chromatine ouverte

Grâce aux données de CORCES et al., 2016, nous connaissons les régions de chromatine ouverte pour plusieurs types de cellule sanguines et pour leurs cellules progénitrices. Cette connaissance nous permet de restreindre la recherche de TFBS aux régions où nos facteurs de transcription d'intérêt ne sont pas mécaniquement dans l'impossibilité d'interagir avec la chromatine et de réguler l'expression génique.

Certaines régions se chevauchent et peuvent être analysées une seule fois pour tous les types cellulaires, ce qui accélère les calculs et le chargement des données génotypiques. Les régions qui se chevauchent sont unifiées (figure 4.15), et les régions issues de cette unification sont scannées indépendamment les unes des autres.

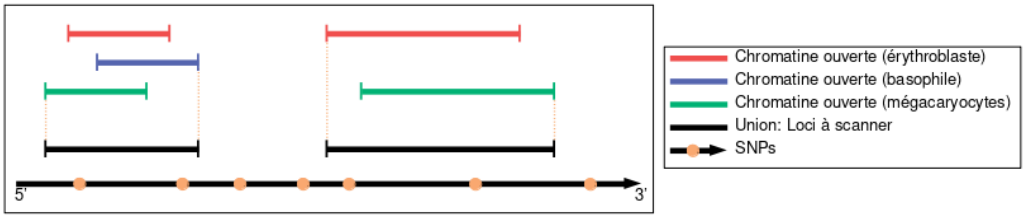


FIGURE 4.15 – Unification des régions de chromatine ouverte de tous les types cellulaires

4.10.2 Orientation des TFBS

LIS et WALTHER, 2016 suggère que l'orientation du TFBS est généralement neutre ("the binding orientation of transcription factors may generally not be relevant"). Pour cette raison, nous comptabilisons de la même manière les TFBS orientés dans les deux sens. Pour identifier les TFBS dans le sens négatif, nous avons besoin de prendre le complément de la séquence inverse, comme montré dans DAI, HE et ZHAO, 2007 : en plus d'inverser l'ordre du motif, nous prenons le complément de chaque position (A <-> T et C <-> G).

4.10.3 Encodage

Afin d'utiliser des outils standards de tests d'association, qui sont conçus pour étudier des variants ordinaires (SNV, indel, ..), nous devons représenter ces variations du nombre de TFBS sous la forme de variants génétiques fictifs. Le champs "DS" (dosage) du format VCF est parfait pour ce besoin. Pour une région de chromatine ouverte et un facteur de transcription donnés, nous comptons le nombre de TFBS dans les deux haplotypes de chaque participant et nous calculons leur somme. Ce nombre de TFBS peut varier entre une valeur minimale dans la cohorte (encodée avec la valeur $DS = 0.0$) et une valeur maximale (encodée avec la valeur $DS = 2.0$), et les valeurs intermédiaires peuvent être interpolées de manière triviale et sans perte de précision.

Pour assurer une compatibilité avec les outils de tests d'association qui ne supportent pas le champ DS, nous avons aussi traduit les variations du nombre de TFBS dans le champ GT (génotype) du format VCF. Les valeurs minimale, moyenne et maximale sont représentées par "0|0", "0|1" et "1|1". Cet encodage est évidemment moins précis que l'encodage décrit plus haut. Les deux encodages sont comparés dans la table 4.IV.

Nombre de TFBS	Champ GT	Champs DS	Perte d'information (GT/DS)
0, 1	0|0, 1|1	0.0, 2.0	Non/Non
0, 1, 2	0|0, 0|1, 1|1	0.0, 1.0, 2.0	Non/Non
0, 1, 2, 4	0|0, 0|0, 0|1, 1|1	0.0, 0.5, 1.0, 2.0	Oui/Non

TABLE 4.IV – Comparaison des deux encodages des TFBS

Les régions qui ne sont pas polymorphiques en terme du nombre de TFBS (i.e tous les participants portent le même nombre de TFBS pour un TF donné) sont ignorées.

Nous avons aussi implémenté un filtre pour ne garder que les régions suffisamment polymorphiques : par exemple si dans une cohorte de $N = 100$ participants, 95 personnes ont deux TFBS, 3 personnes ont un TFBS et 2 personnes n'en ont aucun, nous calculons une fréquence $F = (2+3)/N = 5\%$. Tous les groupes, sauf le groupe le plus nombreux (ici 95) sont ajoutés au numérateur. Ce filtre élimine les régions pour lesquelles il serait difficile d'établir une association statistique fiable.

4.10.4 Traitement d'un locus

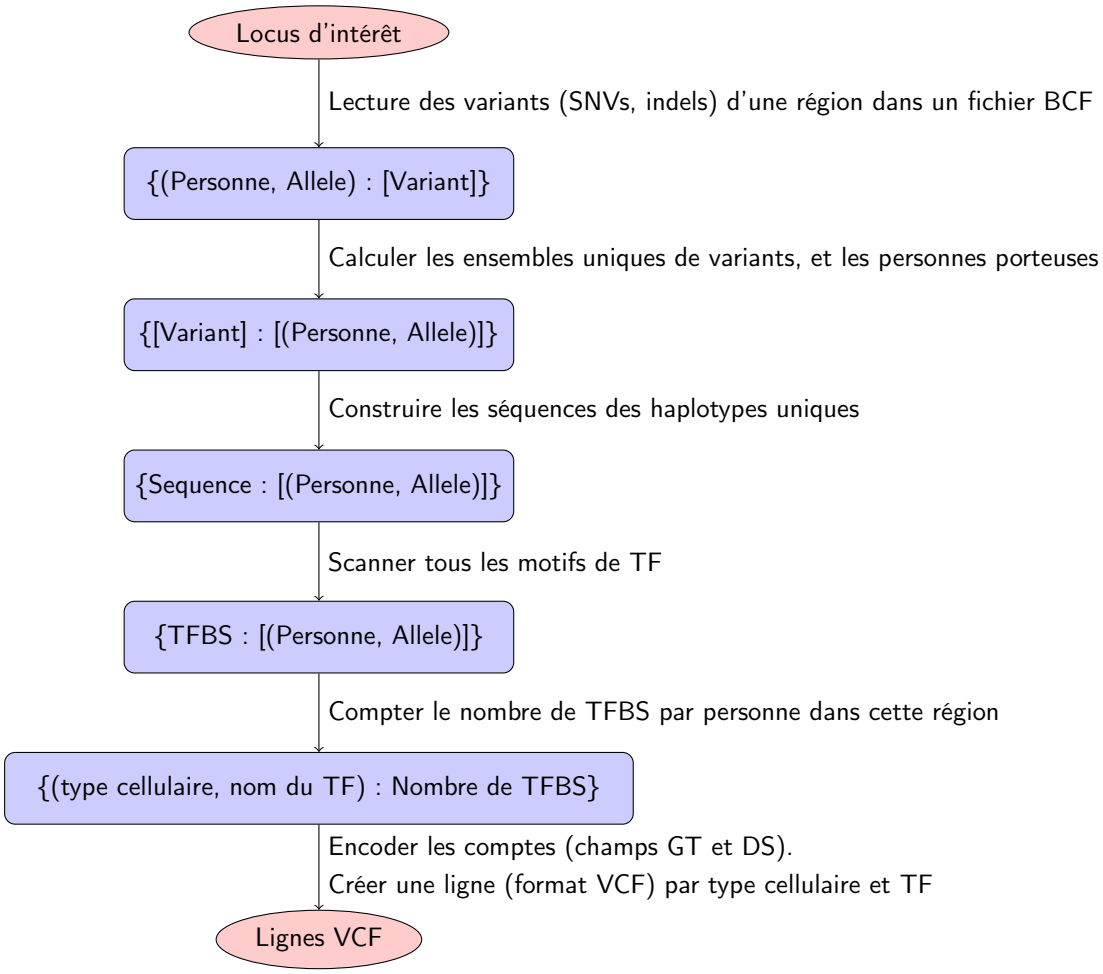


FIGURE 4.16 – Séquence de traitement d'un locus et types de données principaux. "(X,Y)" est une paire de X et Y : par exemple, (Personne, Allele) identifie l'un des allèle d'une personne. "[X]" est une liste de X : par exemple, [Variant] contient une liste de variants. "X -> Y" représente une table de hachage dont les clefs sont de type X et les valeurs sont de type Y, et "TFBS -> [(Personne, Allele)]" classifie les haplotypes par les TFBS qu'ils contiennent. Les clefs d'une table de hachage sont uniques par définition.

Pour commencer, les variants génétiques (SNVs, indels) du locus sont lus à partir d'un fichier BCF. Nous créons une table de hachage. Les clefs de cette table de hachage sont les identifiants des participants et des allèles, et les valeurs sont les listes de différence par rapport au génome de référence.

Ensuite, nous inversons les clefs et les valeurs de la table de hachage et utilisons le génome de

référence pour construire les séquences de chaque haplotype distinct. Chaque haplotype distinct est associé avec une liste de participants et d'allèles. Cette représentation est économe en mémoire vive.

Nous pouvons maintenance scanner les séquences des haplotypes pour trouver les TFBS. Chaque TFBS est associé avec une liste de participants et d'allèles.

Étant donné que la plupart des variants sont rares, le nombre d'haplotypes distincts pour un locus donné est généralement beaucoup plus petit que $2N$ (N = taille de la cohorte). Il est donc important pour la vitesse de calcul de ne scanner que les haplotypes distincts.

Les régions spécifiques aux différents types cellulaires peuvent maintenant être considérés séparemment. Nous comptons le nombre de TFBS pour chaque type cellulaire. Le résultat est sérialisé dans le format VCF, et une ligne est créée pour chaque type cellulaire et TF, à condition que la fréquence des polymorphismes soit suffisante.

4.10.5 Parallélisme

Étant donné que les loci unifiés ne se chevauchent pas (par définition), ils peuvent être analysés séparemment par un nombre arbitraire de threads. La figure 4.17 illustre un flux de traitement parallèle avec trois threads.

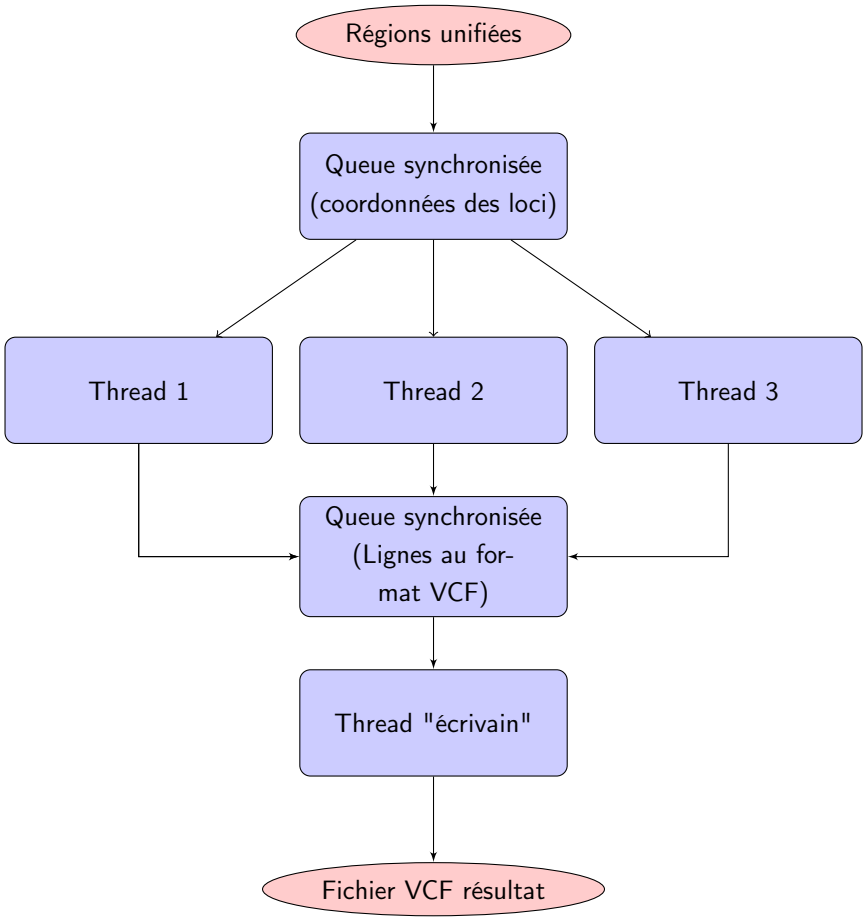
D'abord, les coordonnées de tous les loci d'intérêt sont envoyés à une queue synchronisée. Chaque thread fonctionne dans une boucle : à chaque iteration, le thread lit les coordonnées d'un locus, l'analyse, génère un résultat au format VCF et l'envoie à une deuxième queue synchronisée.

Le thread "écrivain" reçoit les morceaux de VCF et les écrit en séquence dans le fichier résultat. Après la reception du dernier morceau, l'écriture dans le fichier est complétée et le fichier est fermé.

Ce flux de travail occupe tous les threads de manière optimale, même si certains loci demandent un traitement plus long que d'autres. Le thread "écrivain" garantit que les écritures dans le fichier résultat sont réalisées l'une après l'autre, mais l'ordre du fichier n'est pas garanti.

Pour de meilleures performances, le fichier du génome de référence (format FASTA indexé) et le fichier de génotypes (format BCF indexé) ne sont ouverts qu'une seule fois à la création de chaque thread. Ceci économise 200ms de temps de traitement par locus.

FIGURE 4.17 – Parallélisme de locus et synchronisation des threads



4.10.6 Langage d'implémentation

Ce logiciel est implémenté dans le langage Rust, un langage de programmation de plus en plus utilisé pour le calcul haute performance. La gestion explicite de la mémoire de Rust aide le programmeur à minimiser le nombre d'allocations et la consommation totale de mémoire vive, ce qui améliore les performances globales.

Rust fournit des outils pour partager la mémoire de manière sûre dans les programmes multi-threads, par exemple les queues synchronisés. Le langage garantit que toute donnée qui est visible par plus d'un thread ne peut être accédée qu'en toute sécurité, ce qui protège le programmeur de certaines erreurs subtiles mais courantes.

Le langage vérifie au moment de la compilation que les erreurs potentielles sont bien gérées. Rust refuse de compiler un programme qui ne gère pas correctement plusieurs classes d'erreurs potentielles, comme les accès en mémoire non autorisés, les pointeurs nuls et les variables non initialisées.

Il n'a aucun comportement indéfini.

Rust fournit un ensemble croissant de bibliothèques bioinformatiques. Rust-bio (J. KÖSTER, 2016) gère les fichiers FASTA et BED tandis que rust-htslib (KÖSTER, 2020) gère les fichiers BCF indexés.

En résumé, Rust offre les avantages du C et du C ++, deux langages couramment utilisés pour la bioinformatique haute performance, tout en protégeant le programmeur de plusieurs classes d'erreurs de programmation.

4.11 Reproducibilité logicielle

Les projets bio-informatiques dépendent de plusieurs logiciels et librairies tiers. Leur installation peut être fastidieuse et source d'erreurs pour une équipe qui souhaite reproduire les résultats sur un nouvel ordinateur. Par exemple, l'installation d'un logiciel a demandé plusieurs semaines de travail à l'un de mes collègues bio-informaticien.

Pour faciliter cet effort de collaboration, il est utile de fournir une méthode d'installation aussi simple que possible.

4.11.1 Installation de l'environnement

HOSTE, 2018 compare plusieurs outils modernes de gestion de dépendances informatiques. Parmi ceux-ci, le gestionnaire de paquets Nix (DOLSTRA, 2006) garantit la reproductibilité totale d'un environnement informatique.

L'installation des logiciels a été réalisée avec Nix. Tous les programmes et librairies tiers et leurs versions exactes ont été énumérés dans le fichier descripteur *shell.nix*, et la commande "nix-shell" de Nix a téléchargé et installé cet environnement complet de manière reproductible. Les programmes installés sont uniquement visibles par le projet courant, et ne peuvent pas affecter d'autres projets.

Nous avons aussi défini des paquets Nix pour les trois logiciels qui étaient indisponibles dans le dépôt officiel de Nix : *FlashPCA*, *tabix* et *SnakeMake*. La définition du paquet *SnakeMake* a été partagée en open source avec le projet Nix, ce qui permettra à d'autres équipes de recherche d'utiliser *SnakeMake* dans un environnement reproductible. La soumission a été réalisée en deux étapes : ajout de la dépendance Python *ratelimiter* version 1.2.0.post0 [6] puis ajout du logiciel *SnakeMake* version 5.2.2 [7]. *SnakeMake* est maintenant utilisé par la communauté et mis à jour régulièrement [8] de

6. https://github.com/NixOS/nixpkgs/commit/c49e507bbc3286827d4610b08f1ccc18463e25b0
7. https://github.com/NixOS/nixpkgs/commit/93ce77af405b0be6a6f5f5108b8e59cbac97249d
8. https://github.com/NixOS/nixpkgs/commits/cff5adc2fbd8838cc8451d7e58c6770f7b032ae1/pkgs/applications/science/misc/snakemake

l'historique).

Nix offre plusieurs avantages par rapport à d'autres gestionnaires de paquets tels que Conda :

— Herméticité (i.e indépendance vis-à-vis de la version du système d'exploitation sous-jacent et des versions des librairies partagées) : les paquets Nix spécifient de manière transitive toutes leurs dépendances exactes, ce qui garantit la fiabilité de l'installation sur tous les ordinateurs Linux (BOESPFLUG & HUFSCHMITT, 2018). Pour la même raison, une mise à jour du système d'exploitation n'affecte pas le projet courant

— Un changement de version ou la désinstallation d'un logiciel ou d'une librairie ne laisse aucune trace résiduelle dans l'environnement, comme si l'environnement entier était réinstallé à neuf

— Support des dépendances en diamant : si deux logiciels requièrent deux versions différentes d'un outil, Nix expose la bonne version à chaque logiciel. Ceci permet d'installer des logiciels qui sont d'ordinaire incompatibles

— Composabilité : les environnements de deux projets peuvent toujours être joints automatiquement

— Atomicité : Un changement dans la définition de l'environnement est appliqué soit entièrement, soit pas du tout, même si l'ordinateur s'éteint à cause d'une coupure de courant

Les gestionnaires de conteneurs tels que Docker (MERKEL, 2014) et Singularity (SYLABS, 2020) répondent principalement au besoin d'isolation entre chaque projet, mais leur garantie de reproductibilité est perdue dès que le conteneur n'est plus distribué en ligne.

La première installation de l'environnement peut être longue car Nix compile et optimise les logiciels pour l'architecture de l'ordinateur courant. Les fichiers générés sont conservés par Nix pour les utilisations futures, y compris après la désinstallation d'un logiciel.

Nix est compatible avec un environnement de calcul distribué, et est utilisé par les grappes de Calcul Canada.

4.11.2 Orchestration

La transformation des fichiers d'entrée (e.g fichier de génotypes et de phénotypes) en fichier de sortie (associations statistiques, méta-analyses, figures) est divisée en plusieurs étapes.

Ces étapes sont définies dans un fichier Snakemake (J. KÖSTER & RAHMANN, 2012). Chaque étape est composée d'une liste de fichiers d'entrée, d'une liste de fichiers de sortie et d'une commande pour créer les seconds à partir des premiers.

SnakeMake exécute la commande quand les fichiers de sortie sont absents et quand l'un des fichiers d'entrée est plus récent que les fichiers de sortie. Grâce à ce mécanisme, les résultats finaux sont toujours en accord avec les fichiers de départ et le bio-informaticien n'a pas besoin de se rappeler quelles commandes doivent être lancées et dans quel ordre.

Snakemake est capable de lancer ces commandes dans un environnement distribué tel que les grappes de Calcul Canada. Afin de respecter les conditions d'utilisation, nous définissons un temps de calcul et la consommation de mémoire maximale pour chaque règle. Pour les analyses qui portent sur un chromosome, il est pratique de demander des ressources proportionnelles à la taille de ce chromosome.

Ce mécanisme, en tandem avec le versionnage des scripts dans Git (BLISCHAK, DAVENPORT & WILSON, 2016), rend toutes les analyses reproductibles.

Cet exemple basique de règle SnakeMake compte le nombre de variants perte-de-fonction pour chaque ethnicité (*pop*). Le script *count_lof.py* lit un fichier VCF et crée un fichier *lof_count.tab*. Snakemake demande deux heures de calcul et 1Go de mémoire vive à Compute Canada :

```
rule lof_count:
        input: "data/{pop}/genotype_lof/all.vcf.gz"
        output: "data/{pop}/genotype_lof/lof_count.tab"
        threads: 1
        resources: mem=1000, runtime=120
        shell: "python bin/count_lof.py {output} {input}"
```

4.11.3 Flot de travail

Un projet de recherche basé sur Nix et SnakeMake peut être téléchargé, installé et exécuté sur n'importe quel ordinateur Linux ou grappe de calcul. La génération de l'article final, avec les figures mises à jour, peut faire partie de ce système. Voici les instructions (hors téléchargement des données confidentielles) :

```
~$ git clone <project>           # Copie l'ensemble du projet
~$ cd project
~$ nix-shell                     # Installe tous les logiciels requis
[nix-shell:~/project]$ snakemake # Lance les analyses
```

4.12 Reproduction des résultats

Grâce aux données du projet UK Biobank, une base de données de participants Britanniques, nous avons pu confirmer ou infirmer certaines associations issues de TOPMed. Le UK Biobank fournit les exomes complets de 50k individus (90% d'origine Européenne, 27233 femmes et 22675 hommes, 8959608 variants), leur sexe, age et statut de fumeur, les composantes principales de leur génome et leurs phénotypes normalisés. Ceci nous permet de répliquer les associations des régions codantes. Nous avons utilisé un modèle linéaire simple avec une correction pour les covariables principales et la structure de population.

D'autres variants ont été trouvés dans le GWAS catalog (J. MACARTHUR et al., 2017, D1), une base de données qui consolide les résultats de 5687 études pan-génomiques.

4.13 Usage du logiciel

Un exemple d'usage du logiciel find-tfbs ci-dessous :

```
~$ count_pwm --usage
count_pwm
  --input chr1.bcf
  --bed CMP.bed,Erythro.bed
  --chromosome chr1
  --output out.vcf.fz
  --pwm_file HOCOMOCOv11_full_pwms_HUMAN_mono.txt
  --pwm_names GATA1_HUMAN.H11MO.1.A,GATA2_HUMAN.H11MO.1.A
  --pwm_threshold 0.0001
  --pwm_threshold_directory thresholds
  --reference hg38.fa
  --threads 1
  --min_maf 5
  --samples samples
```

Le fichier *samples* contient les identifiants de tous les participants d'intérêt (un identifiant par ligne), et les fichiers *bed* contiennent les régions d'intérêt (un fichier par type cellulaire).

CHAPITRE 5

RÉSULTATS ET DISCUSSION

5.1 Associations des variants perte-de-fonction et des knockouts

La population Africaine fournit un nombre important de pKO malgré un nombre de participants réduit. Les pertes-de-fonction et les 2714 gènes pour lesquels nous avons observé au moins un pKO sont distribués de la manière suivante entre les quatre ethnicités (figure 5.1) :

Ethnicité	Africaine	Asiatique	Européenne	Hispanique
pLoF	55750	4377	114401	53105
pKO	1617	395	1634	1557

TABLE 5.I – Nombre de variants pLoF et nombre de KO par bloc continental

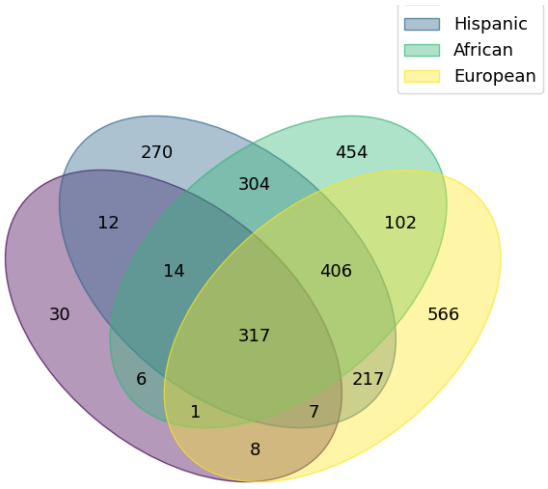


FIGURE 5.1 – Diagramme de Venn des gènes pKO par bloc continental

Pour les associations de variants pLoF, le seuil de p-valeur déterminé par Bonferroni ($0.05/N_{variants}$) est de 1.0×10^{-5} pour la population Asiatique, 9.0×10^{-7} pour la population Africaine et Hispanique, et 4.0×10^{-7} pour la population Européenne.

Chez la population Africaine, les associations avec des variants qui se trouvent à proximité du locus Duffy ont été retirées de la liste de résultats. Le mécanisme responsable du phénotype Duffy-négatif est bien caractérisé dans REICH et al., 2009 et DUCHENE et al., 2017. Ce locus subit une pression de sélection positive car il protège contre la malaria.

Trait	Gène/Variant	Population	Statistiques résumées	Réplication	Nouveauté
WBC	LEO1 chr15 :51941557	Africaine	$p = 3.54 \times 10^{-6}$ $\beta = -0.51$, $se = 0.11$		Non
LYMPH	ALOX12B chr17 :8077106	Africaine	$p = 8.55 \times 10^{-6}$ $\beta = 1.87$, $se = 0.42$	Pas dans UKBB	
NEUTRO	MYRFL chr12 :69936623	Africaine	$p = 2.00 \times 10^{-6}$ $\beta = 1.33$, $se = 0.28$	Pas dans UKBB	
MCV	SNX25 chr4 :185339389	Hispanique	$p = 7.08 \times 10^{-6}$ $\beta = 2.44$, $se = 0.54$	Pas dans UKBB	
BASO	NBPF14 chr1 :148579167	Européenne	$p = 5.96 \times 10^{-6}$ $\beta = -1.74$, $se = 0.38$		
LYMPH	S1PR1 chr1 :101236991	Méta-analyse	$p = 3.99 \times 10^{-6}$ $\beta = -0.07$, $se = 0.015$	Pas dans UKBB	
MCV	SERPINB11 chr18 :63712688	Méta-analyse	$p = 5.60 \times 10^{-6}$ $\beta = -0.68$, $se = 0.15$	Pas dans UKBB	
MCV	WDSUB1 chr2 :159236041	Africaine	$p = 6.09 \times 10^{-6}$ $\beta = -2.68$, $se = 0.59$	Pas répliqué	
EOSIN	CD52 chr1 :26318046	Méta-analyse	$p = 6.68 \times 10^{-6}$ $\beta = 0.13$, $se = 0.03$	Pas répliqué	
HGB	CD6 chr11 :61017803	Européenne	$p = 7.39 \times 10^{-6}$ $\beta = -2.57$, $se = 0.58$	Pas dans UKBB	
HCT	CD6 chr11 :61017803	Européenne	$p = 7.73 \times 10^{-6}$ $\beta = -2.57$, $se = 0.57$	Pas dans UKBB	

FIGURE 5.2 – Associations de variants pLoF

Pour les associations de variants pKO, le seuil de p-valeur déterminé par Bonferroni ($0.05/N_{genes}$) est de 1.3×10^{-4} pour la population Asiatique et 3.1×10^{-5} pour les autres populations.

Trait	Gène/Variant	Population	Statistiques résumées	Réplication	Nouveauté
BASO	MAGEB16 chrX :35802369	Hispanique	$p = 1.84 \times 10^{-5}$ $\beta = -0.10$, $se = 0.02$	Pas dans UKBB	
MCV	ZNF3 chr7 :100064789	Africaine	$p = 3.60 \times 10^{-5}$ $\beta = 0.27$, $se = 0.06$	Pas dans UKBB	
MCH	ZNF3 chr7 :100064789	Africaine	$p = 8.53 \times 10^{-5}$ $\beta = 0.28$, $se = 0.07$	Pas dans UKBB	
BASO	CSAG1 chrX :152728120	Hispanique	$p = 7.03 \times 10^{-5}$ $\beta = -0.22$, $se = 0.05$	Pas répliqué	
PLT	OR1AD1 chr12 :48202411	Africaine	$p = 7.12 \times 10^{-5}$ $\beta = -0.15$, $se = 0.04$		

FIGURE 5.3 – Associations de variants pKO

5.2 Associations des sites de liaison des facteurs de transcription

Nous avons collecté 1127 associations entre des variations du nombre de TFBS et des phénotypes sanguins dont la p-valeur est inférieure à 10^{-9}. La plupart de ces associations concernent trois loci qui sont déjà bien caractérisés : Duffy/DARC, complexe HLA et α-globine. Il est difficile d'établir une interprétation pour ces trois loci, car les déséquilibres de liaison avec de nombreux variants causaux peuvent créer des associations fictives.

Parmi les 90 autres associations significatives ($p < 10^{-9}$), nous trouvons plusieurs candidats plausibles. Par exemple, une variation du nombre de TFBS de GATA3 et ETS2 dans une région de chromatine ouverte des cellules MEP (progénitrices des plaquettes) à proximité du gène TAOK1 est liée au volume moyen des plaquettes de la population Africaine. Nous trouvons aussi une association entre les TFBS (TWST1, TFE2 et ITF2) et le nombre de plaquettes, dans une région proche du gène JMJD1C dont la chromatine est ouverte chez les cellules MPP (qui est aussi un type de cellule progénitrice des plaquettes). Les résultats de TAOK1 et JMJD1C sont cohérents avec des études d'association antérieures, qui avaient découvert d'autres associations avec des traits de plaquettes.

Nous avons extrait des données supplémentaires pour les TFBS d'intérêt qui se situent dans un promoteur, car ceux-ci se prêtent à une interprétation plus aisée. Nous avons téléchargé la liste des promoteurs définis dans le "Regulatory Build" de Ensembl (ZERBINO, WILDER, JOHNSON, JUETTEMANN & FLICEK, 2015), qui est défini à partir de la présence de marques d'histones dans plusieurs types cellulaires. Pour chaque promoteur, nous avons sélectionné tous les variants liés à un trait hématopoïetiques d'une étude pan-génomique antérieure (CHEN et al., 2020 ; VUCKOVIC et al., 2020), et

nous avons réalisé une analyse conditionnelle où ces variants sont considérés comme des covariables, de la même manière que l'âge et le sexe. Les résultats de cette analyse conditionnelle nous indiquent si les variants TFBS sont statistiquement indépendants des variants connus.

Aucune variation des TFBS dans les promoteurs (figure 5.II), à l'exception du promoteur de RENBP, n'est statistiquement indépendante, ce qui pourrait indiquer que les variants TFBS expliquent partiellement ou entièrement les effets des variants de l'analyse pan-génomique.

Population	Trait	Type cellulaire	Région de chromatine ouverte	TFBS	P-valeur	P-valeur (cond.)	Gène du promoteur
Européenne	MPV	HSC	chr12 :121789119-121790240	ZEB1	1.33^{-29}	0.5036	LINC01089,RHOF
Européenne	WBC	CD8	chr17 :39863834-39864579	STAT6	3.06^{-10}	0.4858	IKZF3,ZPBP2
Européenne	WBC	Mono	chr17 :40062711-40063806	CTCF	2.43^{-15}	0.8519	THRA
Européenne	MCV	MPP	chr7 :100626407-100627165	CTCFL	8.63^{-14}	0.09284	TFR2
Européenne	MCH	MPP	chr7 :100626407-100627165	CTCFL	1.14^{-13}	0.04977	TFR2
Européenne	MBC	MPP	chr7 :100626407-100627165	CTCFL	2.23^{-12}	0.4069	TFR2
Européenne	MCH	Erythro	chr7 :100642442-100642911	FLI1	3.79^{-15}	0.01461	TFR2
Européenne	MCV	Erythro	chr7 :100642442-100642911	FLI1	4.47^{-15}	0.9462	TFR2
Européenne	RBC	Erythro	chr7 :100642442-100642911	FLI1	4.09^{-13}	0.2606	TFR2
Européenne	WBC	CMP	chr7 :28684549-28685114	STAT1	1.39^{-13}	0.2477	CREB5
Africaine	RBC	MEP	chrX :153945909-153946614	CTCFL	1.47^{-18}	8.91^{-05}	RENBP
Africaine	RBC	MEP	chrX :153945909-153946614	CTCFL	2.12^{-14}	1.45^{-07}	RENBP
Africaine	RBC	MEP	chrX :153945909-153946614	CTCFL	2.77^{-12}	1.71^{-04}	RENBP
Hispanique	RBC	MEP	chrX :153945909-153946614	CTCFL	5.56^{-10}	3.27^{-05}	RENBP
Hispanique	RBC	MEP	chrX :153945909-153946614	CTCFL	9.65^{-14}	1.542^{-04}	RENBP
Hispanique	RBC	MEP	chrX :153945909-153946614	CTCFL	8.08^{-11}	2.51^{-06}	RENBP

TABLE 5.II – Associations entre les variations des TFBS situés dans les promoteurs et les traits sanguins ($p < 10^{-9}$). L'analyse conditionnelle a utilisé tous les variants d'une étude pan-génomique dans une fenêtre de une mégabase. La dernière colonne indique le gène lié au promoteur.

Dans le cas de RENBP, une étude plus approfondie nous montre une région complexe (figure 5.4) : les 46 variants du pic de chromatine ouverte de 705 nucléotides forment 50 haplotypes distincts dans la population Africaine. Parmi ces 46 variants, trois variants changent les prédictions du modèle PWM : deux variants détruisent un TFBS du facteur de transcription CTCTL (X:153946252_G_A, X:153946429_C_T) et un variant crée un TFBS de CTCFL. Ce dernier variant, rs7889328 est le seul variant dont l'association avec les traits d'érythrocytes (concentration, distribution de leur largeur, et volume moyen) reste significative lors d'une analyse conditionnelle.

Le possible rôle de RENBP dans l'hématopoïèse n'est pas évident. RENBP est connu pour être un inhibiteur du système rénine–angiotensine–aldostérone qui régule la pression artérielle et n'a pas

de lien direct connu avec ces caractéristiques des érythrocytes. RENBP semble aussi jouer un rôle inconnu dans le système immunitaire (van BILSEN et al., 2020) et est un catalyseur dans le metabolisme de l'acide sialique (LUCHANSKY, YAREMA, TAKAHASHI & BERTOZZI, 2003).

Le rôle du facteur de transcription CTCFL est aussi difficile à interpréter. En conditions normales, CTCFL est exprimé dans la spermatogénèse. Cependant, des études ont observé une surexpression de CTCFL dans plusieurs types de cancers (BERGMAIER et al., 2018). Étant donnée la similarité des motifs de CTCFL et de CTCF, nous avons vérifié que les résultats d'association étaient exclusifs à CTCFL.

Le tableau et la figure suivants illustrent la structure et les associations du locus de RENBP (5.III, 5.4).

Population	Trait	TFBS	TFBS (cond.)	rs7889328 (cond.)
Africaine	RBC	p=1.5e-18, β=-0.32	p=8.9e-5, β=-0.18	p=1.3e-4, β=-0.12
Africaine	RDW	p=2.1e-14, β=-0.38	p=1.5e-7, β=-0.34	p=1.5e-7, β=-0.22
Africaine	MCV	p=2.8e-12, β=0.23	p=1.7e-4, β=0.16	p=2.1e-4, β=0.10
Hispanique	RBC	p=5.5e-10, β=-0.40	p=3.3e-5, β=-0.23	p=3.3e-5, β=-0.23
Hispanique	RDW	p=9.6e-14, β=-0.51	p=1.5e-4, β=-0.22	p=1.5e-4, β=-0.22
Hispanique	MCV	p=8.1e-11, β=0.41	p=2.5e-6, β=0.26	p=2.5e-6, β=0.26

TABLE 5.III – Associations entre les TFBS situés dans le promoteur de RENBP et les traits sanguins. β représente le changement du phénotype normalisé pour un incrément du nombre de TFBS, exprimé en nombre de déviations standards. L'analyse conditionnelle a utilisé tous les variants d'une étude pan-génomique dans une fenêtre de un mégabase. Les deux dernières colonnes présentent les résultats d'association conditionnelle pour les TFBS et pour le variant rs7889328.

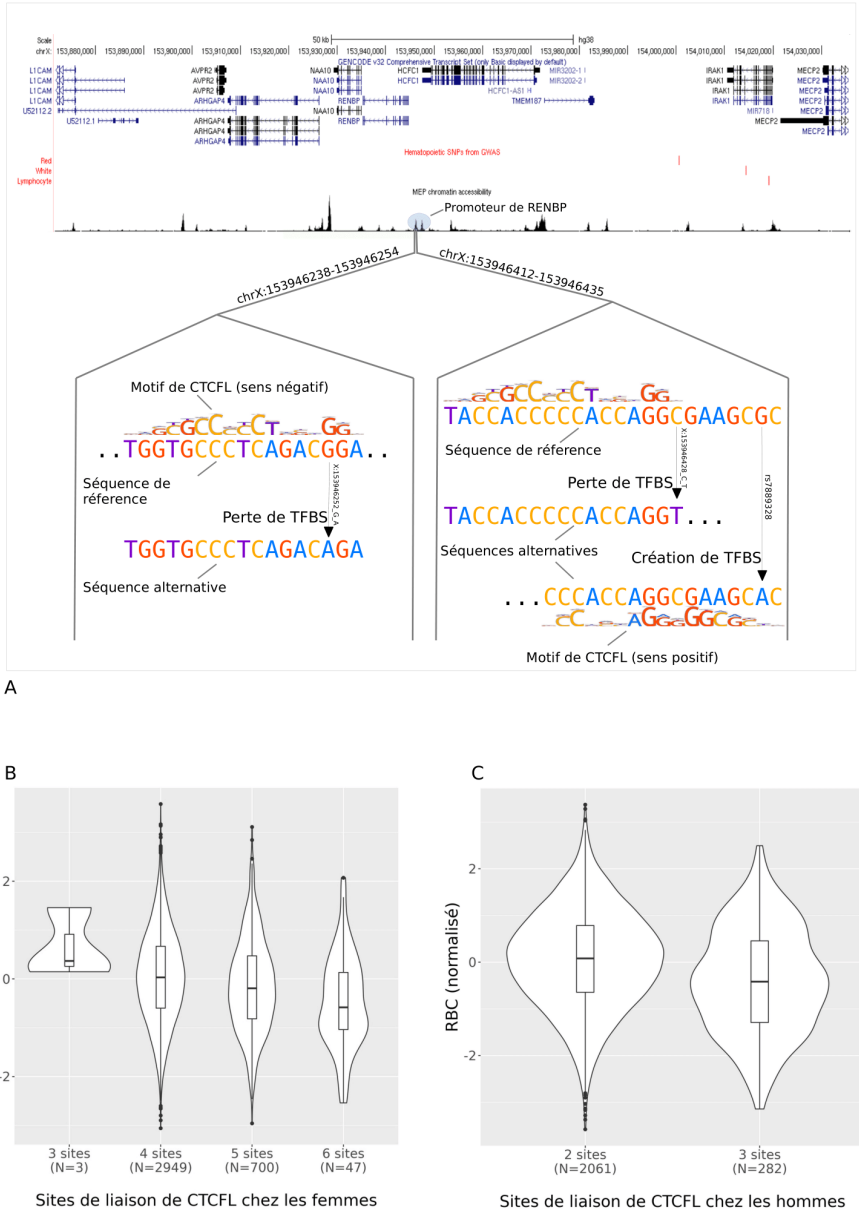


FIGURE 5.4 – Trois variants situés dans un pic de chromatine ouverte des cellules progénitrices mégacaryocytes-érythroïde (MEP), situé au sein du promoteur de RENBP, changent les prédictions du modèle PWM pour le facteur de transcription CTCFL. (A) La première figure montre le contexte du gène RENBP : les régions de chromatine ouverte et les variants associés à des traits sanguins lors d'une étude pan-génomique antérieure. Les deux variants X:153946252_G_A et X:153946429_C_T détruisent les deux TFBS qui sont présents dans le génome de référence, alors que le variant rs7889328 crée un site de liaison pour CTCFL. Notons que ces variants ne changent pas toujours les nucléotides les plus conservés et peuvent paraître surprenants ; par exemple X:153946429_C_T transforme un C (score=0.04) en T (score=-0.96). (B) et (C) Nombre d'érythrocytes (axe des ordonnées) en fonction du nombre de TFBS de CTCFL dans le promoteur de RENBP (axe des abscisses). Le nombre de TFBS est réduit chez les hommes car ce locus est situé sur le chromosome X.

Dans les cohortes TOPMed, le variant rs7889328 est relativement fréquent chez la population Africaine (MAF=0.11), et relativement rare chez les populations Hispanique (MAF=0.02) et Européenne (MAF=0.0002).

CHAPITRE 6

CONCLUSION

Nos deux approches se sont focalisées sur des classes de variants génétiques dont il est possible de prédire certaines conséquences : perte-de-fonction, knockout, et altération de l'expression d'un gène via l'altération d'un TFBS.

La priorisation des variants perte-de-fonction et la création des pseudo-variants knockouts pour la recherche d'associations nous a permis d'étudier des variants codants rares et d'augmenter notre puissance statistique. Cette approche a confirmé l'implication de plusieurs gènes dont le rôle dans l'hématopoïese était déjà caractérisé, ce qui valide notre approche et notre implémentation, et nous avons découvert plusieurs nouvelles associations. La réplication de ces découvertes est difficile, car ces variants sont rares et ne se retrouvent que rarement dans d'autres populations telles que le UKBB. Un effort de réplication basé sur des lignées cellulaires et sur l'édition du génome est techniquement possible, mais son coût élevé doit être justifiable. Il est possible que de futurs jeux de données (WGS ou éxome) permettront de confirmer ou d'infirmer nos résultats, et nous espérons que les outils d'annotation des variants continueront de progresser afin de mieux prédire quels variants sont susceptibles de causer une perte-de-fonction.

La priorisation des variants qui affectent les TFBS facilite l'interprétation des résultats d'association, ce qui est généralement difficile pour les variants des régions non codantes. En particulier, nous savons que les variants TFBS qui se trouvent dans un promoteur affectent probablement le gène voisin, et nous savons grâce aux cartes de chromatine ouverte dans quel type cellulaire ce mécanisme a des chances de jouer un rôle. Ces informations ne sont pas produites par une étude pan-génomique classique, qui de plus souffre d'un nombre d'hypothèses plus élevé. En faisant le choix de nous restreindre aux variants TFBS, nous rendons possible la détection d'effets plus faibles sur le phénotype et la détection de variants plus rares. Dans le futur, la cartographie des régions de chromatine ouverte sera disponible pour d'autres types cellulaires, et les bases de données de motifs seront enrichies avec des motifs plus précis.

Nous espérons que des cohortes plus nombreuses seront étudiées grâce à un coût de séquençage réduit et permettront de poursuivre ces approches. Nous espérons aussi que les réductions des coûts d'autres techniques permettront de générer des informations complémentaires, telles que des informations individuelles sur la méthylation de la chromatine et sur les marques d'histone. Ces informations nous aideraient à mieux prioriser les variants TFBS et à réduire le nombre d'hypothèses.

BIBLIOGRAPHIE

ABRAHAM, G. & INOUYE, M. (2016, avril 5). FlashPCA : fast sparse canonical correlation analysis of genomic data. doi :10.1101/047217

ASTLE, W. & BALDING, D. J. (2010, octobre 22). Population structure and cryptic relatedness in genetic association studies. doi :10.1214/09-STS307

AVSEC, Ž., WEILERT, M., SHRIKUMAR, A., KRUEGER, S., ALEXANDARI, A., DALAL, K., ... ZEITLINGER, J. (2020). Base-resolution models of transcription factor binding reveal soft motif syntax. *bioRxiv*. Publisher : Cold Spring Harbor Laboratory _eprint : https ://www.biorxiv.org/content/early/202 doi :10.1101/737981

BAO, K., CARR, T., WU, J., BARCLAY, W., JIN, J., CIOFANI, M. & REINHARDT, R. L. (2016, décembre 1). BATF modulates the Th2 locus control region and regulates CD4+ T cell fate during anti-helminth immunity. *Journal of immunology (Baltimore, Md. : 1950)*, *197*(11), 4371-4381. doi :10.4049/jimmunol.1601371

BATISTA, C. R., LI, S. K. H., XU, L. S., SOLOMON, L. A. & DEKOTER, R. P. (2017, janvier 6). PU.1 regulates ig light chain transcription and rearrangement in pre-b cells during b cell development. *The Journal of Immunology*. Publisher : American Association of Immunologists Section : IMMUNE SYSTEM DEVELOPMENT. doi :10.4049/jimmunol.1601709

BAUER, D. E., KAMRAN, S. C., LESSARD, S., XU, J., FUJIWARA, Y., LIN, C., ... ORKIN, S. H. (2013, octobre 11). An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. *Science (New York, N.Y.) 342*(6155), 253-257. doi :10.1126/science.1242088

BERGMAIER, P., WETH, O., DIENSTBACH, S., BOETTGER, T., GALJART, N., MERNBERGER, M., ... RENKAWITZ, R. (2018, août 21). Choice of binding sites for CTCFL compared to CTCF is driven by chromatin and by sequence preference. *Nucleic Acids Research*, *46*(14), 7097-7107. Publisher : Oxford Academic. doi :10.1093/nar/gky483

BLISCHAK, J. D., DAVENPORT, E. R. & WILSON, G. (2016, janvier 19). A Quick Introduction to Version Control with Git and GitHub. *PLoS Computational Biology*, *12*(1). doi :10.1371/journal.pcbi.1004668

BOESPFLUG, M. & HUFSCHMITT, T. (2018, mars 15). Tweag I/O - Nix + Bazel = fully reproducible, incremental builds. Récupérée 21 novembre 2018, à partir de https://www.tweag.io/posts/2018-03-15-bazel-nix.html

BOETTCHER, M. & MCMANUS, M. T. (2015, mai 21). Choosing the Right Tool for the Job : RNAi, TALEN or CRISPR. *Molecular cell*, *58*(4), 575-585. doi :10.1016/j.molcel.2015.04.028

BUSH, W. S. (2012). Genome-Wide Association Studies. Récupérée 20 novembre 2018, à partir de https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002822

CAMPBELL, M. C. & TISHKOFF, S. A. (2008). AFRICAN GENETIC DIVERSITY : Implications for Human Demographic History, Modern Human Origins, and Complex Disease Mapping. *Annual review of genomics and human genetics*, *9*, 403-433. doi :10.1146/annurev.genom.9.081307.164258

CARLSON, C. S., MATISE, T. C., NORTH, K. E., HAIMAN, C. A., FESINMEYER, M. D., BUYSKE, S., ... PAGE CONSORTIUM. (2013, septembre). Generalization and dilution of association results from European GWAS in populations of non-European ancestry : the PAGE study. *PLoS biology*, *11*(9), e1001661. doi :10.1371/journal.pbio.1001661

CHEN, M.-H., RAFFIELD, L. M., MOUSAS, A., SAKAUE, S., HUFFMAN, J. E., MOSCATI, A., ... LETTRE, G. (2020, septembre 3). Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations. *Cell*, *182*(5), 1198-1213.e14. Publisher : Elsevier. doi :10.1016/j.cell.2020.06.045

CINGOLANI, P., PLATTS, A., WANG, L. L., COON, M., NGUYEN, T., WANG, L., ... RUDEN, D. M. (2012, avril 1). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly*, *6*(2), 80-92. doi :10.4161/fly.19695

CLAUSSNITZER, M., DANKEL, S. N., KIM, K.-H., QUON, G., MEULEMAN, W., HAUGEN, C., ... KELLIS, M. (2015, septembre 3). FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *The New England Journal of Medicine*, *373*(10), 895-907. doi :10.1056/NEJMoa1502214

CORCES, M. R., BUENROSTRO, J. D., WU, B., GREENSIDE, P. G., CHAN, S. M., KOENIG, J. L., ... CHANG, H. Y. (2016). Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nature Genetics*, *48*(10), 1193-1203. doi :10.1038/ng.3646

DAI, X., HE, J. & ZHAO, X. (2007, juillet). A new systematic computational approach to predicting target genes of transcription factors. *Nucleic Acids Research*, *35*(13), 4433-4440. doi :10.1093/nar/gkm454

DANECEK, P., AUTON, A., ABECASIS, G., ALBERS, C. A., BANKS, E., DEPRISTO, M. A., ... DURBIN, R. (2011, août 1). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156-2158. doi :10.1093/bioinformatics/btr330

DEBOEVER, C., TANIGAWA, Y., MCINNES, G., LAVERTU, A., CHANG, C., BUSTAMANTE, C. D., ... RIVAS, M. A. (2017, septembre 2). Medical relevance of protein-truncating variants across 337,208 individuals in the UK biobank study. *bioRxiv*, 179762. doi :10.1101/179762

DOLSTRA, E. (2006). *The purely functional software deployment model* (thèse de doct., Utrecht, S.l.). OCLC : 71702886. Récupérée à partir de https://nixos.org/~eelco/pubs/phd-thesis.pdf

DUCHENE, J., NOVITZKY-BASSO, I., THIRIOT, A., CASANOVA-ACEBES, M., BIANCHINI, M., ETHERIDGE, S. L., ... ROT, A. (2017, juillet). Atypical chemokine receptor 1 on nucleated erythroid cells regulates hematopoiesis. *Nature immunology*, *18*(7), 753-761. doi :10.1038/ni.3763

EVANGELOU, E. & IOANNIDIS, J. P. A. (2013, juin). Meta-analysis methods for genome-wide association studies and beyond. *Nature Reviews Genetics*, *14*(6), 379-389. doi :10.1038/nrg3472

FERNÁNDEZ, J. R., ETZEL, C., BEASLEY, T. M., SHETE, S., AMOS, C. I. & ALLISON, D. B. (2002). Improving the power of sib pair quantitative trait loci detection by phenotype winsorization. *Human Heredity*, *53*(2), 59-67. doi :10.1159/000057984

FRAZER, K. A., MURRAY, S. S., SCHORK, N. J. & TOPOL, E. J. (2009, avril). Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics*, *10*(4), 241-251. doi :10.1038/nrg2554

GAUTAM, S., FIORAVANTI, J., ZHU, W., GALL, J. B. L., BROHAWN, P., LACEY, N. E., ... GATTINONI, L. (2019, mars). The transcription factor c-myb regulates CD8 + t cell stemness and antitumor immunity. *Nature Immunology*, *20*(3), 337-349. Number : 3 Publisher : Nature Publishing Group. doi :10.1038/s41590-018-0311-z

GIEGER, C., RADHAKRISHNAN, A., CVEJIC, A., TANG, W., PORCU, E., PISTIS, G., ... SORANZO, N. (2011, novembre 30). New gene functions in megakaryopoiesis and platelet formation. *Nature*, *480*(7376), 201-208. doi :10.1038/nature10659

GRUSDAT, M., MCILWAIN, D. R., XU, H. C., POZDEEV, V. I., KNIEVEL, J., CROME, S. Q., ... LANG, P. A. (2014, juillet). IRF4 and BATF are critical for CD8+ T-cell function following infection with LCMV. *Cell Death and Differentiation*, *21*(7), 1050-1060. doi :10.1038/cdd.2014.19

HANNA, R. N., CARLIN, L. M., HUBBELING, H. G., NACKIEWICZ, D., GREEN, A. M., PUNT, J. A., ... HEDRICK, C. C. (2011, août). The transcription factor NR4a1 (nur77) controls bone marrow differentiation and the survival of ly6c- monocytes. *Nature Immunology*, *12*(8), 778-785. Number : 8 Publisher : Nature Publishing Group. doi :10.1038/ni.2063

HENG LI, BOB HANDSAKER, PETR DANECEK, SHANE MCCARTHY & JOHN MARSHALL. (2019, novembre 19). BCFtools (Version 1.7). Récupérée à partir de https://samtools.github.io/bcftools/bcftools-man.html

HILL, W. G., GODDARD, M. E. & VISSCHER, P. M. (2008, février 29). Data and theory point to mainly additive genetic variance for complex traits. *PLoS genetics*, *4*(2), e1000008. doi :10.1371/journal.pgen.1000008

HIRSCHHORN, J., LOHMUELLER, K., BYRNE, E. & HIRSCHHORN, K. (2002). A comprehensive review of genetic association studies | Genetics in Medicine. Récupérée 6 novembre 2018, à partir de https://www.nature.com/articles/gim200210

HONG, E. P. & PARK, J. W. (2012, juin). Sample Size and Statistical Power Calculation in Genetic Association Studies. *Genomics & Informatics*, *10*(2), 117-122. doi :10.5808/GI.2012.10.2.117

HOSTE, K. (2018). Installing software for scientists on a multi-user HPC system, 35.

ITOH-NAKADAI, A., MATSUMOTO, M., KATO, H., SASAKI, J., UEHARA, Y., SATO, Y., ... IGARASHI, K. (2017, mars). A bach2-cebp gene regulatory network for the commitment of multipotent hematopoietic progenitors. *Cell Reports*, *18*(10), 2401-2414. doi :10.1016/j.celrep.2017.02.029

JAGANNATHAN-BOGDAN, M. & ZON, L. I. (2013, juin 15). Hematopoiesis. *Development (Cambridge, England)*, *140*(12), 2463-2467. doi :10.1242/dev.083147

JASON D. BUENROSTRO. (2018). Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation : Cell. Récupérée 24 avril 2020, à partir de https://doi.org/10.1016/j.cell.2018.03.074

KANG, H. M., SUL, J. H., SERVICE, S. K., ZAITLEN, N. A., KONG, S.-Y., FREIMER, N. B., ... ESKIN, E. (2010, avril). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, *42*(4), 348-354. doi :10.1038/ng.548

KATSUMURA, K. R., DEVILBISS, A. W., POPE, N. J., JOHNSON, K. D. & BRESNICK, E. H. (2013, septembre). Transcriptional Mechanisms Underlying Hemoglobin Synthesis. *Cold Spring Harbor Perspectives in Medicine*, *3*(9). doi :10.1101/cshperspect.a015412

KOHU, K., OHMORI, H., WONG, W. F., ONDA, D., WAKOH, T., KON, S., ... SATAKE, M. (2009, décembre 15). The runx3 transcription factor augments th1 and down-modulates th2 phenotypes by interacting with and attenuating GATA3. *The Journal of Immunology*, *183*(12), 7817-7824. Publisher : American Association of Immunologists Section : CELLULAR IMMUNOLOGY AND IMMUNE REGULATION. doi :10.4049/jimmunol.0802527

KÖSTER. (2020, janvier 17). rust-htslib. Rust-Bio. Récupérée 20 janvier 2020, à partir de https://github.com/rust-bio/rust-htslib

KÖSTER, J. (2016, février 1). Rust-bio : a fast and safe bioinformatics library. *Bioinformatics*, *32*(3), 444-446. doi :10.1093/bioinformatics/btv573

KÖSTER, J. & RAHMANN, S. (2012, octobre 1). Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, *28*(19), 2520-2522. doi :10.1093/bioinformatics/bts480

KULAKOVSKIY, I. V., VORONTSOV, I. E., YEVSHIN, I. S., SHARIPOV, R. N., FEDOROVA, A. D., RUMYNSKIY, E. I., ... MAKEEV, V. J. (2018, janvier 4). HOCOMOCO : towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-seq analysis. *Nucleic Acids Research*, *46*, D252-D259. Publisher : Oxford Academic. doi :10.1093/nar/gkx1106

LE SAOUT, C., LUCKEY, M. A., VILLARINO, A. V., SMITH, M., HASLEY, R. B., MYERS, T. G., ... CATALFAMO, M. (2017). IL-7–dependent STAT1 activation limits homeostatic CD4+ T cell expansion. *JCI Insight*, *2*(22). doi :10.1172/jci.insight.96228

LESSARD, S., GATOF, E. S., BEAUDOIN, M., SCHUPP, P. G., SHER, F., ALI, A., ... LETTRE, G. (2017, août 1). An erythroid-specific ATP2B4 enhancer mediates red blood cell hydration and malaria susceptibility. *The Journal of Clinical Investigation*, *127*(8), 3065-3074. doi :10.1172/JCI94378

LETTRE, G., SANKARAN, V. G., BEZERRA, M. A. C., ARAÚJO, A. S., UDA, M., SANNA, S., ... ORKIN, S. H. (2008, août 19). DNA polymorphisms at the BCL11A, HBS1L-MYB, and β-globin loci associate with fetal hemoglobin levels and pain crises in sickle cell disease. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(33), 11869-11874. doi :10.1073/pnas.0804799105

LI, Z., LI, X., LIU, Y., SHEN, J., CHEN, H., ZHOU, H., ... LIN, X. (2019). Dynamic Scan Procedure for Detecting Rare-Variant Association Regions in Whole-Genome Sequencing Studies. *American Journal of Human Genetics*, *104*(5), 802-814. doi :10.1016/j.ajhg.2019.03.002

LIN, D. Y. & ZENG, D. (2010, janvier). Meta-Analysis of Genome-Wide Association Studies : No Efficiency Gain in Using Individual Participant Data. *Genetic epidemiology*, *34*(1). doi :10.1002/gepi.20435

LIS, M. & WALTHER, D. (2016, mars 3). The orientation of transcription factor binding site motifs in gene promoter regions : does it matter ? *BMC Genomics*, *17*. doi :10.1186/s12864-016-2549-x

LUCHANSKY, S. J., YAREMA, K. J., TAKAHASHI, S. & BERTOZZI, C. R. (2003, juillet 3). GlcNAc 2-epimerase can serve a catabolic role in sialic acid metabolism. *Journal of Biological Chemistry*, *278*(10), 8035-8042. Publisher : American Society for Biochemistry and Molecular Biology. doi :10.1074/jbc.M212127200

MACARTHUR, D. G. [D. G.], MANOLIO, T. A., DIMMOCK, D. P., REHM, H. L., SHENDURE, J., ABECASIS, G. R., ... GUNTER, C. (2014, avril 24). Guidelines for investigating causality of sequence variants in human disease. *Nature*, *508*(7497), 469-476. doi :10.1038/nature13127

MACARTHUR, D. G. [Daniel G.], BALASUBRAMANIAN, S., FRANKISH, A., HUANG, N., MORRIS, J., WALTER, K., ... TYLER-SMITH, C. (2012, février 17). A systematic survey of loss-of-function variants in human protein-coding genes. *Science (New York, N.Y.) 335*(6070), 823-828. doi :10.1126/science.1215040

MACARTHUR, J., BOWLER, E., CEREZO, M., GIL, L., HALL, P., HASTINGS, E., ... PARKINSON, H. (2017, janvier 4). The new NHGRI-EBI catalog of published genome-wide association studies (GWAS catalog). *Nucleic Acids Research*, *45*, D896-D901. doi :10.1093/nar/gkw1133

MÄKI-TANILA, A. & HILL, W. G. (2014, septembre 1). Influence of gene interaction on complex trait variation with multilocus models. *Genetics*, *198*(1), 355-367. doi :10.1534/genetics.114.165282

MARLOW, S. (2010). *Haskell 2010 Language Report*.

MAURANO, M. T., HUMBERT, R., RYNES, E., THURMAN, R. E., HAUGEN, E., WANG, H., ... STAMATOYANNOPOULOS, J. A. (2012, septembre 7). Systematic localization of common disease-associated variation in regulatory DNA. *Science (New York, N.Y.) 337*(6099), 1190-1195. doi :10.1126/science.1222794

MCCARTHY, D. J., HUMBURG, P., KANAPIN, A., RIVAS, M. A., GAULTON, K., ASDS, ... DONNELLY, P. (2014). Choice of transcripts and software has a large effect on variant annotation. *Genome Medicine*, *6*(3), 26. doi :10.1186/gm543

MCGREGOR, T. L., HUNT, K. A., YEE, E., MASON, D., NIOI, P., TICAU, S., ... van HEEL, D. A. (2020). Characterising a healthy adult with a rare HAO1 knockout to support a therapeutic strategy for primary hyperoxaluria. *eLife*, *9*. doi :10.7554/eLife.54363

MCLAREN, W., GIL, L., HUNT, S. E., RIAT, H. S., RITCHIE, G. R. S., THORMANN, A., ... CUNNINGHAM, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, *17*(1), 122. doi :10.1186/s13059-016-0974-4

MEDVEDOVIC, J., EBERT, A., TAGOH, H. & BUSSLINGER, M. (2011). Pax5 : a master regulator of B cell development and leukemogenesis. *Advances in Immunology*, *111*, 179-206. doi :10.1016/B978-0-12-385991-4.00005-2

MERKEL, D. (2014, mars 1). Docker : lightweight Linux containers for consistent development and deployment. Belltown Media.

MONIR, M. M. & ZHU, J. (2017, décembre). Comparing GWAS results of complex traits using full genetic model and additive models for revealing genetic architecture. *Scientific Reports*, *7*(1). doi :10.1038/srep38600

MORRIS, S. A. (2019, juin 15). The evolving concept of cell identity in the single cell era. *Development*, *146*(12). Publisher : Oxford University Press for The Company of Biologists Limited Section : SPOTLIGHT. doi :10.1242/dev.169748

MORRISON, A. C., HUANG, Z., YU, B., METCALF, G., LIU, X., BALLANTYNE, C., ... BOERWINKLE, E. (2017). Practical Approaches for Whole-Genome Sequence Analysis of Heart- and Blood-Related Traits. *American Journal of Human Genetics*, *100*(2), 205-215. doi :10.1016/j.ajhg.2016.12.009

MORRISON, A. C., VOORMAN, A., JOHNSON, A. D., LIU, X., YU, J., LI, A., ... COHORTS FOR HEART AND AGING RESEARCH IN GENETIC EPIDEMIOLOGY (CHARGE) CONSORTIUM. (2013, août). Whole-genome sequence-based analysis of high-density lipoprotein cholesterol. *Nature Genetics*, *45*(8), 899-901. doi :10.1038/ng.2671

MOUSAS, A., NTRITSOS, G., CHEN, M.-H., SONG, C., HUFFMAN, J. E., TZOULAKI, I., ... REINER, A. P. (2017, août 7). Rare coding variants pinpoint genes that control human hematological traits. *PLoS Genetics*, *13*(8). doi :10.1371/journal.pgen.1006925

MUSUNURU, K., STRONG, A., FRANK-KAMENETSKY, M., LEE, N. E., AHFELDT, T., SACHS, K. V., ... RADER, D. J. (2010, août 5). From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature*, *466*(7307), 714-719. doi :10.1038/nature09266

NARASIMHAN, V. M., HUNT, K. A., MASON, D., BAKER, C. L., KARCZEWSKI, K. J., BARNES, M. R., ... van HEEL, D. A. (2016, avril 22). Health and population effects of rare gene knockouts in adult humans with related parents. *Science (New York, N.Y.) 352*(6284), 474-477. doi :10.1126/science.aac8624

NARASIMHAN, V. M., XUE, Y. & TYLER-SMITH, C. (2016, avril 1). Human Knockout Carriers : Dead, Diseased, Healthy, or Improved ? *Trends in Molecular Medicine*, *22*(4), 341-351. doi :10.1016/j.molmed.2016.02.006

NATARAJAN, P., PELOSO, G. M., ZEKAVAT, S. M., MONTASSER, M., GANNA, A., CHAFFIN, M., ... NHLBI TOPMED LIPIDS WORKING GROUP. (2018). Deep-coverage whole genome sequences and blood lipids among 16,324 individuals. *Nature Communications*, *9*(1), 3391. doi :10.1038/s41467-018-05747-8

NHGRI. (2018, octobre 29). GWAS Catalog. Récupérée 3 novembre 2018, à partir de https://www.ebi.ac.uk/gwas/

PARNAS, O., JOVANOVIC, M., EISENHAURE, T. M., HERBST, R. H., DIXIT, A., YE, C. J., ... REGEV, A. (2015, juillet 30). A genome-wide CRISPR screen in primary immune cells to dissect regulatory networks. *Cell*, *162*(3), 675-686. doi :10.1016/j.cell.2015.06.059

POLFUS, L. M., KHAJURIA, R. K., SCHICK, U. M., PANKRATZ, N., PAZOKI, R., BRODY, J. A., ... SANKARAN, V. G. (2016, août 4). Whole-Exome Sequencing Identifies Loci Associated with Blood Cell Traits and Reveals a Role for Alternative GFI1B Splice Variants in Human Hematopoiesis. *American Journal of Human Genetics*, *99*(2), 481-488. doi :10.1016/j.ajhg.2016.06.016

PRICE, A. L., PATTERSON, N. J., PLENGE, R. M., WEINBLATT, M. E., SHADICK, N. A. & REICH, D. (2006, août). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 904-909. doi :10.1038/ng1847

PRUITT, K. D., TATUSOVA, T. & MAGLOTT, D. R. (2005, janvier 1). NCBI reference sequence (RefSeq) : a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, *33*, D501-D504. Publisher : Oxford Academic. doi :10.1093/nar/gki025

PURCELL, S., NEALE, B., TODD-BROWN, K., THOMAS, L., FERREIRA, M. A. R., BENDER, D., ... SHAM, P. C. (p. d.). *PLINK : a toolset for whole genome association and*.

R CORE TEAM. (2013). *R : A Language and Environment for Statistical Computing*. Vienna, Austria : R Foundation for Statistical Computing. Récupérée à partir de http://www.R-project.org/

REICH, D., NALLS, M. A., KAO, W. H. L., AKYLBEKOVA, E. L., TANDON, A., PATTERSON, N., ... WILSON, J. G. (2009, janvier). Reduced neutrophil count in people of African descent is due to a regulatory variant in the Duffy antigen receptor for chemokines gene. *PLoS genetics*, *5*(1), e1000360. doi :10.1371/journal.pgen.1000360

RENTZSCH, P., WITTEN, D., COOPER, G. M., SHENDURE, J. & KIRCHER, M. (2018, octobre 29). CADD : predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*. doi :10.1093/nar/gky1016

SALEHEEN, D., NATARAJAN, P., ARMEAN, I. M., ZHAO, W., RASHEED, A., KHETARPAL, S., ... KATHIRESAN, S. (2017, avril 12). Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. *Nature*, *544*(7649), 235-239. doi :10.1038/nature22034

SERRE, D., MONTPETIT, A., PARÉ, G., ENGERT, J. C., YUSUF, S., KEAVNEY, B., ... ANAND, S. (2008, janvier 2). Correction of Population Stratification in Large Multi-Ethnic Association Studies. *PLoS ONE*, *3*(1). doi :10.1371/journal.pone.0001382

SHANG, J., ZHANG, J., SUN, Y., LIU, D., YE, D. & YIN, Y. (2011, décembre 15). Performance analysis of novel methods for detecting epistasis. *BMC Bioinformatics*, *12*(1), 475. doi :10.1186/1471-2105-12-475

SHETE, S., BEASLEY, T. M., ETZEL, C. J., FERNÁNDEZ, J. R., CHEN, J., ALLISON, D. B. & AMOS, C. I. (2004, mars). Effect of winsorization on power and type 1 error of variance components and related methods of QTL detection. *Behavior Genetics*, *34*(2), 153-159. doi :10.1023/B:BEGE.0000013729.26354.da

SIEBERT, M. & SÖDING, J. (2016, juillet 27). Bayesian markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. *Nucleic Acids Research*, *44*(13), 6055-6069. Publisher : Oxford Academic. doi :10.1093/nar/gkw521

SIMONE MESMAN & MARTEN P. SMIDT. (2017). Frontiers | Tcf12 Is Involved in Early Cell-Fate Determination and Subset Specification of Midbrain Dopamine Neurons | Frontiers in Molecular Neuroscience. Récupérée 24 avril 2020, à partir de https://doi.org/10.3389/fnmol.2017.00353

SIU, G., WURSTER, A. L., LIPSICK, J. S. & HEDRICK, S. M. (1992, avril). Expression of the CD4 gene requires a Myb transcription factor. *Molecular and Cellular Biology*, *12*(4), 1592-1604. Récupérée 24 avril 2020, à partir de https://www.ncbi.nlm.nih.gov/pmc/articles/PMC369602/

SUDLOW, C., GALLACHER, J., ALLEN, N., BERAL, V., BURTON, P., DANESH, J., ... COLLINS, R. (2015, mars 31). UK biobank : an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine*, *12*(3), e1001779. Publisher : Public Library of Science. doi :10.1371/journal.pmed.1001779

SYLABS. (2020). Singularity [Sylabs.io]. Récupérée 16 avril 2020, à partir de https://sylabs.io/singularity/

TATTINI, L., D'AURIZIO, R. & MAGI, A. (2015). Detection of genomic structural variants from next-generation sequencing data. *Frontiers in Bioengineering and Biotechnology*, *3*. Publisher : Frontiers Media SA. doi :10.3389/fbioe.2015.00092

THE LEVERHULME CENTRE FOR DEMOGRAPHIC SCIENCE. (2020). GWAS Diversity Monitor. Récupérée 16 avril 2020, à partir de https://www.gwasdiversitymonitor.com/

TURNER, S., ARMSTRONG, L. L., BRADFORD, Y., CARLSON, C. S., CRAWFORD, D. C., CRENSHAW, A. T., ... RITCHIE, M. D. (2011, janvier). Quality Control Procedures for Genome Wide Association Studies. *Current protocols in human genetics / editorial board, Jonathan L. Haines ... [et al.] CHAPTER*, Unit1.19. doi :10.1002/0471142905.hg0119s68

VAGAPOVA, E. R., SPIRIN, P. V., LEBEDEV, T. D. & PRASSOLOV, V. S. (2018). The Role of TAL1 in Hematopoiesis and Leukemogenesis. *Acta Naturae*, *10*(1), 15-23. Récupérée 24 avril 2020, à partir de https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5916730/

van der HARST, P., ZHANG, W., LEACH, I. M., RENDON, A., VERWEIJ, N., SEHMI, J., ... CHAMBERS, J. C. (2012, décembre 20). Seventy-five genetic loci influencing the human red blood cell. *Nature*, *492*(7429), 369-375. doi :10.1038/nature11677

VAN ROSSUM, G. & DRAKE JR, F. L. (1995). *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam.

van BILSEN, J. H. M., DULOS, R., van STEE, M. F., MEIMA, M. Y., ROUHANI RANKOUHI, T., NEERGAARD JACOBSEN, L., ... KRISHNAN, S. (2020). Seeking Windows of Opportunity to Shape Lifelong Immune Health : A Network-Based Strategy to Predict and Prioritize Markers of Early Life Immune Modulation. *Frontiers in Immunology*, *11*, 644. doi :10.3389/fimmu.2020.00644

VASQUEZ, L. J., MANN, A. L., CHEN, L. & SORANZO, N. (2016, janvier). From GWAS to function : lessons from blood cells. *Isbt Science Series*, *11*, 211-219. doi :10.1111/voxs.12217

VISSCHER, P. M., WRAY, N. R., ZHANG, Q., SKLAR, P., MCCARTHY, M. I., BROWN, M. A. & YANG, J. (2017, juillet 6). 10 Years of GWAS Discovery : Biology, Function, and Translation. *American Journal of Human Genetics*, *101*(1), 5-22. doi :10.1016/j.ajhg.2017.06.005

VUCKOVIC, D., BAO, E. L., AKBARI, P., LAREAU, C. A., MOUSAS, A., JIANG, T., ... SORANZO, N. (2020, septembre 3). The Polygenic and Monogenic Basis of Blood Traits and Diseases. *Cell*, *182*(5), 1214-1231.e11. Publisher : Elsevier. doi :10.1016/j.cell.2020.08.008

WANG, K., LI, M. & HAKONARSON, H. (2010, septembre). ANNOVAR : functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), e164. doi :10.1093/nar/gkq603

WEEKS, N. T., LUECKE, G. R., GROTH, B. M., KRAEVA, M., MA, L., KRAMER, L. M., ... REECY, J. M. (2018, mai 1). High-performance epistasis detection in quantitative trait GWAS. *The International Journal of High Performance Computing Applications*, *32*(3), 321-336. doi :10.1177/1094342016658110

WENDLING, F. (1999, janvier 1). Thrombopoietin : its role from early hematopoiesis to platelet production. *Haematologica*, *84*(2), 158-166. Récupérée 20 novembre 2018, à partir de http://www.haematologica.org/content/84/2/158

WICKHAM, H. (2016). *ggplot2 : Elegant Graphics for Data Analysis*. Springer-Verlag New York. Récupérée à partir de https://ggplot2.tidyverse.org

WICKHAM, H., AVERICK, M., BRYAN, J., CHANG, W., MCGOWAN, L. D., FRANÇOIS, R., ... YUTANI, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. doi :10.21105/joss.01686

WILLER, C. J., LI, Y. & ABECASIS, G. R. (2010, septembre 1). METAL : fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, *26*(17), 2190-2191. doi :10.1093/bioinformatics/btq340

ZERBINO, D. R., WILDER, S. P., JOHNSON, N., JUETTEMANN, T. & FLICEK, P. R. (2015, mars 24). The Ensembl Regulatory Build. *Genome Biology*, *16*(1), 56. doi :10.1186/s13059-015-0621-5

ZHU, Y. P., THOMAS, G. D. & HEDRICK, C. C. (2016, septembre). Transcriptional Control of Monocyte Development. *Arteriosclerosis, thrombosis, and vascular biology*, *36*(9), 1722-1733. doi :10.1161/ATVBAHA.116.304054

ZHUANG, Y., CHENG, P. & WEINTRAUB, H. (1996, juin). B-lymphocyte development is regulated by the combined dosage of three basic helix-loop-helix genes, E2A, E2-2, and HEB. *Molecular and Cellular Biology*, *16*(6), 2898-2905. doi :10.1128/mcb.16.6.2898

Chapitre 7

find-tfbs : a tool to identify functional non-coding variants associated with complex human traits using open chromatin maps and phased whole-genome sequences

find-tfbs: a tool to identify functional non-coding variants associated with complex human traits using open chromatin maps and phased whole-genome sequences

Sébastian Méric de Bellefon[1,2,*], Other authors, TOPMed banner, Other authors, Guillaume Lettre[1,2,*]

[1]Montreal Heart Institute, Montréal, Québec, H1T 1C8, Canada. [2]Faculté de Médecine, Université de Montréal, Montréal, Québec, H3T 1J4, Canada

*To whom correspondence should be addressed.

Motivation: Whole-genome DNA sequencing (WGS) enables the discovery of non-coding variants, but tools are lacking to prioritize the subset that functionally impacts human phenotypes. DNA sequence variants that disrupt or create transcription factor binding sites (TFBS) can modulate gene expression. find-tfbs efficiently scans phased WGS in large cohorts to identify and count TFBSs in regulatory sequences. This information can then be used in association testing to find putatively functional non-coding variants associated with complex human diseases or traits.

Results: We applied find-tfbs to discover functional non-coding variants associated with hematological traits in the NHLBI Trans-Omics for Precision Medicine (TOPMed) WGS dataset (N_{max}=44,709). We identified >2000 associations at $P<1x10^{-9}$, implicating specific blood cell-types, transcription factors and causal genes. The vast majority of these associations are captured by variants identified in large genome-wide association studies (GWAS) for blood-cell traits. find-tfbs is computationally efficient and robust, allowing for the rapid identification of non-coding variants associated with multiple human phenotypes in very large sample size.

Availability: https://github.com/Helkafen/find-tfbs and https://github.com/Helkafen/find-tfbs-demo

Contacts: sebastian.meric.de.bellefon@umontreal.ca and guillaume.lettre@umontreal.ca

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Genome-wide association studies (GWAS) have identified thousands of common genetic variants (hereby defined as variants with minor allele frequency (MAF) ≥5%) associated with complex human diseases and other quantitative traits. Functional annotation of these variants with maps of open chromatin regions and histone tail modifications has revealed their enrichment within non-coding regulatory sequences (Maurano et al., Vierstra et al.). This suggests that a large fraction of human phenotypic variation is modulated by common variants in sequences that control gene expression. Recently, this hypothesis has been confirmed experimentally in a few robust examples (Musunuru et al., Bauer et al., Lessard et al., Claussnitzer et al.).

In contrast to common variants, most rare genetic variants implicated in human phenotypes have been found in protein-coding exons, mostly because sequencing whole-genomes remained prohibitively expensive until recently. Thus, we still do not know to what extent rare genetic variants in the non-coding human genome can influence inter-individual phenotypic variation. Because rare variants are often missed by the GWAS framework, their identification could yield new loci and genes, or help focus on strong candidate genes at GWAS loci.

Limited statistical power is an important issue for rare variants association testing because of the small number of carriers (by definition) and their very large number in the human genome (Zuk et al., 2014). To lower the multiple hypothesis burden and therefore increase the chance of finding significant associations, analyses of whole-exome sequencing (WES) datasets often collate rare coding variants by genes. For rare non-coding variants found by WGS, collapsing methods based on sliding windows or scanning algorithms have been proposed (e.g. SCANG) (Li et al., 2019; Morrison et al. 2013 and 2017; Natarajan et al., 2018). However, these methods do not consider our current understanding of gene expression regulation, and in particular the important fact that gene expression is controlled by transcription factors (TF) that bind regulatory DNA sequences.

To address this need, we developed a new tool, find-tfbs, that can efficiently scan a large number of phased WGS and identify genetic variants that create or disrupt transcription factor-binding sites (TFBS) within pre-specified regulatory elements. find-tfbs uses position weight matrices (PWM) to identify and count the number of TFBS occurrences found in each regulatory element of each individual. This information can then be used in standard association testing pipelines. To demonstrate its utility and robustness, we used find-tfbs to analyze associations between TFBS and 15 blood-cell traits in WGS data from 44,709 participants sequenced by the NHLBI TOPMed Project.

2 Methods

2.1 Scanning phased WGS data to find TFBS

find-tfbs (**Figure 1**) takes three files as inputs: (1) phased WGS data in the BCF format, (2) genomic coordinates of regions of interest, and (3) PWM of prioritized transcription factors. The genomic coordinates file can contain regions identified by open chromatin experiments (e.g. ATAC-seq, DNase1 hypersensitivity), histone tail marks profiling/segmentation or genomic annotation (e.g. gene promoters). It can come from a single cell type or complex tissue. It is also possible to submit multiple coordinates files at once. Since some of these regions overlap, processing time is reduced by merging the overlapping regions, which allows find-tfbs to extract each genetic variant and scan the merged regions only once. At the end of the analyses, find-tfbs dispatches the TFBS it has found and counted in each region to the corresponding cell type/tissue and TF.

Each merged locus is scanned independently. First, find-tfbs creates a hash table of differences from the reference genome (single nucleotide variants (SNVs) and small insertions-deletions (indels)) in the locus. find-tfbs indexes these differences by haplotype identifiers, where a haplotype identifier represents one strand of one participant. The genetic variants are read from the indexed BCF file. Then, find-tfbs reverses the keys and values of this hash table and uses the reference genome to build the sequence of each unique haplotype: each unique haplotype sequence is associated with a list of haplotype identifiers. This representation is memory efficient. Using the TF PWMs, find-tfbs scans the unique haplotype sequences for TF motifs on both the forward and reverse strand (Lis et al., 2016; Dai et al., 2007), and the TFBS are associated with a list of haplotype identifiers. Since most variants are rare, the number of distinct haplotypes for a given locus is usually much smaller than 2N (N=cohort size). Scanning unique haplotypes reduces the amount of redundant computing.

Finally, find-tfbs counts the number of TFBS in each cell-/tissue-specific locus for all TFs. The result is serialized using the VCF format, and one line is created for each cell type/tissue and transcription factor. find-tfbs discards the lines where the TFBS count frequency (MTCF, defined below) is lower than the threshold (default threshold=0).

2.2. Recoding TFBS counts for association testing

First, find-tfbs counts in each participant the number of TFBS found on both alleles for a given region and adds them together. If every individual in the cohort has the same number of TFBS, find-tfbs ignores this region since it is not polymorphic. find-tfbs encodes polymorphic region using the dosage (DS) and genotype (GT) fields of the VCF file format (Danecek et al., 2011). While some association testing pipelines (e.g EPACTS) accept both fields, others only accept the GT field. The DS field accepts any number between 0.0 and 2.0. For our purposes, 0.0 represents the lowest number of TFBS found in a region in the cohort and 2.0 represents the

highest number. We interpolate the intermediate TFBS counts and lose no accuracy. The possible values of the GT field are '0|0', '0|1' and '1|1'. The lowest and highest numbers of TFBS are encoded as '0|0' and '1|1'. The average of the lowest and highest values is encoded as '0|1'. Every other value is encoded as the closest encoded neighbour, which makes the GT field less accurate than the DS field for some regions.

2.3 Minor TFBS count frequency (MTCF)

For each scanned region and TF, find-tfbs calculates a frequency of TFBS count variations. This allows for simple filtering of regions with too few alternate TFBS counts for association testing. For example, in a cohort of N=100 persons, if 95 individuals have two TFBS in a given region, three participants have one TFBS and two participants have none, then MTCF=(2+3)/N=5%. find-tfbs sums all the groups but the most frequent one. If several groups share the highest frequency, only one of them is taken out of the sum.

2.4 Parallelism and performance

Since the merged loci do not overlap (by definition), they can be analyzed separately by any number of worker threads. First, the coordinates of all the loci are sent to a synchronized channel. Each worker thread works in a loop: at the beginning of an iteration, the worker reads the coordinates of one locus, then analyzes it, generates a VCF-formatted result and sends the result to another synchronized channel. The writer thread receives the VCF-formatted strings and writes them sequentially to the result VCF file. Upon completion, the output file handle is flushed and closed by the writer thread. This workflow keeps all the worker threads busy, even if some loci require more processing time than others, and the writer thread guarantees that file writes are sequential. However, the output order is undefined. We save about 200ms of processing time per locus by opening the reference genome file (an indexed FASTA) and the input genotype file (an indexed BCF) at the creation of each worker thread and by keeping the file handles open.

In find-tfbs, the largest data structures are flat arrays. To minimize the number of CPU cache misses and improve performance, they follow the order of the individuals from the input BCF file. These data structures include the ordered list of participants in the input BCF file, the list of participants who share a TFBS or a haplotype sequence, and the number of matches per participant in a locus. None of these data structures are written on disk, in order to increase performance.

2.5 Implementation language

find-tfbs is implemented in Rust (Matsakis and Klock, 2014), a programming language that is increasingly used for high performance computing. The explicit memory management of Rust helps the programmer minimize memory allocations and total memory usage, thereby increasing overall performance. Rust provides tools to share data safely in multi-threaded programs, for instance synchronized queues and channels. The language guarantees that any

piece of data that is seen by more than one thread can only be accessed safely, which protects the programmer from subtle but common mistakes.

The language also enforces sound error management during compilation. The Rust compiler refuses to compile programs that fail to address several classes of potential runtime errors (e.g memory safety errors, null pointers and uninitialized variables). It has no undefined behavior, unlike C and C++. Rust provides a growing set of bioinformatics libraries. Rust-bio (Köster, 2016) manages FASTA and BED files while rust-htslib (Köster, 2020) manages indexed BCF files.

2.6 Application to hematologic traits

A large number of TFs play a role during the proliferation and differentiation of blood cells. However, in many cases, the downstream target genes of these TFs remain unknown. As an example to test find-tfbs, we explored how variation in TFBS counts for 97 TFs modulate 15 blood-cell phenotypes. Phased WGS data and complete blood count (CBC) came from 44,709 participants sequenced by the NHLBI Trans-Omics for Precision Medicine (TOPMed) whole-genome sequencing project, freeze 8 (Taliun et al.). 9870 and 9757 participants have African and Hispanic ancestry, and 25,569 have European ancestry (**Supplementary Table 1**). The TOPMed WGS dataset (freeze 8) is 781 gigabytes after compression.

To prioritize regions more likely to control gene expression, we analyzed open chromatin regions identified in 16 hematopoietic cell types by ATAC-seq (Corces et al., 2016). The list of TFs was based on a literature review. When the literature was imprecise, we tested all the relevant progenitor and blood-cell types. For instance, when a source indicated that knockout of a TF was associated with erythrocyte count, we tested this TF in open chromatin regions of erythroblasts and all their available progenitors. When a more precise mechanism was known, we only tested the specific cell type. We restricted the list to the 97 TFs that have a known DNA binding motif in the HOCOMOCO database, version 11 (Kulakovskiy et al., 2018).

We corrected blood-cell traits for age, sex and smoker status by ethnicity and cohort, and then normalized the residuals using inverse normal transformation. We then corrected the normalized phenotypes for population structure within each ethnicity, using the first 10 principal components calculated using 149,454 variants in linkage equilibrium. We used EPACTS for association testing, separately for each ethnicity, using the *q.emmax* algorithm which accounts for cryptic relatedness. Since the variant frequency was already controlled by find-tfbs, we removed the frequency filter in EPACTS, and kept the default values for all other parameters.

Some of the supplementary materials for this experiment can be found in the find-tfbs-demo repository. In particular, the list of relevant transcription factors per cell type and the list of phenotypes per cell type are provided. Genomic coordinates for the open chromatin regions from the different blood cell-types (Corces et al., 2016) are included in the repository for convenience.

3.2 Performance

Our experiment was run on a Compute Canada cluster equipped with Intel Xeon Gold 6148 processors. The genotype files occupied a total of 1.006 terabyte for all chromosomes. find-tfbs analyzed 0.41 merged peaks per second on average, using two cores. We used the Linux profiling tool *perf* and observed that loading and decompressing the genotype files was the most resource-intensive task. Building and scanning the unique haplotypes used a relatively small amount of resources.

3 Results

3.1 Blood-cell trait association results

In this study, we used blood-cell traits to test the implementation of find-tfbs. We arbitrarily defined statistical significance as nominal P-value $<1x10^{-9}$. We acknowledge that this threshold does not rigorously take into account the large number of hypotheses tested and emphasize that association results presented here need to be further replicated. All results that meet this statistical significance threshold are available in **Supplementary Table 2**. The vast majority of the significant associations map to the Duffy/*DARC*, HLA and α-globin loci. Because these loci are already known and genetically complex due to their respective linkage disequilibrium patterns, we did not consider them further in our downstream analyses.

Outside of these three regions, we found 90 combinations of "blood-cell traits/open chromatin regions/TFBS" associated at $P<1x10^{-9}$ (**Supplementary Table 2**). This list includes a few highly plausible associations, such as an ATAC-seq peak found in the gene *TAOK1* in megakaryocyte-erythroid progenitor (MEP) cells, which is polymorphic for GATA3 and ETS2 TFBS and associated with mean platelet volume (MPV) in African-ancestry participants. Another interesting association signal, found in European-ancestry individuals, highlights an open chromatin region found in the gene *JMJD1C* in multipotential progenitor (MPP) cells that is associated with platelet counts and include a variable number of binding sites for the TF TWST1, TFE2 and ITF2.

Focusing on associations that map to promoters as annotated in the Ensembl Regulatory Build (Zerbino et al., 2015), we identified signals in the promoters of several genes (**Table 1**). By conditional analyses, we tested if these promoter-based signals were statistically independent from the variants at the same loci that were identified by previous large-scale GWAS for blood-cell traits (Vuckovic et al., Chen et al.). For all but one gene, conditional results were not significant (**Table 1**), suggesting that genetic variants that create or disrupt TFBS in these promoters might explain, at least in part, the GWAS signals. For an ATAC-seq peak in the promoter of *RENBP* located on chromosome X, the association signal remained significant (**Table 1** and **Figure 2**). *RENBP* is an inhibitor in the renin–angiotensin–aldosterone system that regulates arterial blood pressure and it plays an undefined role in early life immune systems (van Bilsen et al.). *RENBP* also serves a catabolic role in sialic acid metabolism (Luchansky et al.). This association signal is present in individuals of African and Hispanic ethnicity and is associated with several red blood cell (RBC) indices (RBC count, RBC distribution width, mean corpuscular volume). This 705-bp open chromatin peak was identified in MEP and encompasses a genetically complex locus: we found 50 distinct haplotypes in the African-ancestry population due to 46 variants (SNPs, indels) and the reference haplotype contains two binding sites for CTCFL, a transcriptional repressor with a similar binding motif to CTCF and that is expressed during spermatogenesis and in certain cancer types (Bergmaier et al.). Out of these 46 variants, two of them disrupt a CTCFL TFBS (X:153946252_G_A, X:153946429_C_T) and one of them creates a CTCFL TFBS (rs7889328)(**Figure 2**). While

other variants overlap with the putative binding sites, their individual effects on the PWM scores do not change our model predictions. Further conditional analyses indicated that alleles at rs7889328 accounted for the remaining association signal after controlling for the known GWAS variants at the locus (**Supplementary Table 3**).

4 Conclusion

We developed find-tfbs, a robust algorithm to scan phased WGS data to identify and count TFBS. We tested our tool on the large TOPMed dataset, with an initial focus on hematological traits. We identified many open chromatin regions that harbor genetic variants that create or disrupt TFBS, and that are associated with blood-cell phenotypes. By conditional analyses, we showed that the majority of these associations capture previously identified GWAS loci. Because of our experimental design, such results are interesting because they highlight a possible molecular mechanism. Indeed, find-tfbs combined with association testing tools (e.g. EPACTS) outputs the location of the open chromatin region, the cell-type in which the region was found, and the TF involved, allowing for guided functional characterization of promising GWAS loci.

find-tfbs was purposely designed to be flexible. It will consider all types of genetic variants, including rare variants, and simple plugins can be added to customize the find-tfbs output for other association testing tools. Because of its optimization, find-tfbs can test many phenotypes and regulatory sequence data types in parallel in very large WGS datasets. In the future, considering alternatives to standard PWMs could further improve find-tfbs. For instance, dinucleotide PWMs (Kulakovskiy et al., 2016) and Bayesian Markov models (Siebert and Söding, 2016) outperform mononucleotide PWMs over a variety of datasets by encoding nucleotide correlations. More recently, a promising convolutional neural network called BPNet was able to discover spacing information between motifs, in agreement with known TF-TF interactions (Avsec et al., 2020). As the number and ethnic diversity of WGS data available increase, we expect that find-tfbs will become an extremely useful bioinformatic tool to explore the non-coding regulatory genome implicated in human phenotypic variation.

Author contributions

S.M.d.B. and G.L. designed the study. All authors contributed data. S.M.d.B. and G.L. wrote the manuscript with contributions from all other authors.

Acknowledgements

We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed. The contributions of the investigators of the NHLBI TOPMed Consortium (https://www.nhlbiwgs.org/topmed-banner-authorship) are gratefully acknowledged.

The views expressed in this manuscript are those of the authors and do not necessarily represent the views of the National Institutes of Health or the U.S. Department of Health and Human Services.

Funding

This work has been supported by the Canadian Institutes of Health Research (PJT #168902), the Canada Research Chair Program and the Montreal Heart Institute Foundation.

Whole genome sequencing (WGS) for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). WGS for “NHLBI TOPMed: The Jackson Heart Study” (phs000964) was performed at the University of Washington Northwest Genomics Center (HHSN268201100037C). WGS for “NHLBI TOPMed: ARIC” (phs001211) was performed at the Baylor College of Medicine Human Genome Sequencing Center (3U54HG003273-12S2 and HHSN268201500015C). WGS for “NHLBI TOPMed: Amish” (phs000956) was performed at the Broad Institute Genomics Platform (3R01HL121007-01S1). WGS for “NHLBI TOPMed: CARDIA” (phs001612), “NHLBI TOPMed: CHS” (phs001368) and “NHLBI TOPMed: HCHS_SOL” (phs001395) was performed at the Baylor College of Medicine Human Genome Sequencing Center (HHSN268201600033I). WGS for “NHLBI TOPMed: FHS” (phs000974) was performed at the Broad Institute Genomics Platform (3U54HG003067-12S2). WGS for “NHLBI TOPMed: GeneSTAR” (phs001218) was performed at Psomagen (3R01HL112064-04S1). WGS for “NHLBI TOPMed: JHS” (phs000964) was performed at the University of Washington Northwest Genomics Center (HHSN268201100037C). WGS for “NHLBI TOPMed: MESA” (phs001416) was performed at the Broad Institute and Beth Israel Proteomics Platform (HHSN268201600034I). WGS for “NHLBI TOPMed: SAFS” (phs001215) was performed at Illumina (3R01HL113323-03S1). WGS for “NHLBI TOPMed: WHI” (phs001237) was performed at the Broad Institute Genomics Platform (HHSN268201500014C).

Core support including centralized genomic read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Core support including phenotype harmonization, data management, sample-identity QC, and general program coordination were provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed.

Conflict of Interest

The authors declare no conflicts of interest.

References

Avsec,Ž. et al. (2020) Base-resolution models of transcription factor binding reveal soft motif syntax, BioRxiv, doi:10.1101/737981.

Bauer,D. E. et al. (2013) An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. Science (New York, N.Y.), 342(6155), 253–257.

Bergmaier,P., et al. (2018) Choice of binding sites for CTCFL compared to CTCF is driven by chromatin and by sequence preference, Nucleic Acids Res, **46**(14), 7097–7107.

Buenrostro,J. (2018) Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation, Cell. **173**, 1535-1548.

Chen,M. H. et al. (2020) Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations. Cell, 182(5), 1198–1213.e14.

Claussnitzer,M. et al. (2015) FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. The New England journal of medicine, 373(10), 895–907.

Corces, M. et al. (2016) Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. Nat Genet **48,** 1193–1203.

Dai,X. et al. (2007) A new systematic computational approach to predicting target genes of transcription factors. Nucleic Acids Res, **35**, 4433–4440.

Danecek,P. et al. (2011) The variant call format and vcftools. Bioinformatics, **27**,2156-2158.

Gate,R et al. (2018) Genetic determinants of co-accessible chromatin regions in activated T cells across humans, Nat. Genet, **50**, 1140-1150.

Köster,J. (2016) Rust-Bio: a fast and safe bioinformatics library. Bioinformatics., **32**, 444-446.

Kulakovskiy,I. et al. (2018) HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Res, **46**, D252--D259.

Kulakovskiy,I. et al. (2016) HOCOMOCO: expansion and enhancement of the collection of transcription factor binding sites models. Nucleic Acids Res, **44**, D116--D125.

Lessard,S. et al. (2017) An erythroid-specific ATP2B4 enhancer mediates red blood cell hydration and malaria susceptibility. The Journal of clinical investigation, 127(8), 3065–3074.

Li,Z et al (2019) Dynamic Scan Procedure for Detecting Rare-Variant Association Regions in Whole-Genome Sequencing Studies, Am. J. Hum. Genet., **104**(5), 802-814.

Lis,M and Walther,D. (2016) The orientation of transcription factor binding site motifs in gene promoter regions: does it matter? BMC Genom, **17**, 185.

Luchansky,S. J. et al. (2003) GlcNAc 2-epimerase can serve a catabolic role in sialic acid metabolism. The Journal of biological chemistry, 278(10), 8035–8042.

Matsakis,N. and Klock,F. (2014) The Rust Language. Ada Lett, **34**(3), 103-104.

Maurano MT et al. (2012) Systematic localization of common disease-associated variation in regulatory DNA, Science, **337**(6099), 1190-1195.

Morrison,AC. et al. (2013) Whole-genome sequence-based analysis of high-density lipoprotein cholesterol. Nat Genet., **45**(8), 899-901.

Morrison,AC. et al. (2017) Practical Approaches for Whole-Genome Sequence Analysis of Heart- and Blood-Related Traits. Am J Hum Genet., **100**(2), 205-215.

Musunuru,K. et al. (2010) From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature, 466(7307), 714–719.

Natarajan,P. et al. (2018) Deep-coverage whole genome sequences and blood lipids among 16,324 individuals. Nat Commun, **9**(1), 3391.

Siebert,M. and Söding,J. (2016), Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences, Nucleic Acids Res, **44**, 6055-6069.

Sung,M. et al. (2016) Selected heterozygosity at cis-regulatory sequences increases the expression homogeneity of a cell population in humans. Genome Biol, **17,** 164.

Taliun,D. et al. (2019) Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program, BioRxiv, doi:10.1101/563866.

van Bilsen,J. (2020) Seeking Windows of Opportunity to Shape Lifelong Immune Health: A Network-Based Strategy to Predict and Prioritize Markers of Early Life Immune Modulation. Frontiers in immunology, 11, 644.

Vierstra, J. et al. (2020). Global reference mapping of human transcription factor footprints. Nature, **583**(7818), 729–736.

Vuckovic,D. et al. (2020) The Polygenic and Monogenic Basis of Blood Traits and Diseases. Cell, 182(5), 1214–1231.e11.

Zerbino,D.R. et al. (2015) The Ensembl Regulatory Build, Genome Biol, **16**, 56.

Zuk,O et al. (2014) Searching for missing heritability: Designing rare variant association studies, PNAS, **111(4)**, E455-E464.

Table 1. Polymorphic transcription factor binding sites (TFBSs) in gene promoters associate with blood-cell traits. ATAC-seq peaks from different blood-cell types that overlap with ENSEMBL-annotated gene promoters and that include a polymorphic number of TFBS associated with hematological traits. We calculated P-values as described in the **Methods** section; for conditional analyses, we controlled for all genetic variants identified by large blood-cell traits genome-wide association studies located in a 1-Mb window. MPV, mean platelet volume; WBC, white blood cell count; MCV, mean corpuscular volume; MCH, mean corpuscular hemoglobin; RBC, red blood cell count; RDW, RBC distribution width; HSC, hematopoietic stem cell; CD8, CD8+ T lymphocyte; Mono, monocyte; MPP, multipotent progenitor; Erythro, erythroid; CMP, common myeloid progenitor; MEP, megakaryocyte-erythroid progenitor.

Population	Trait	CellType	ATAC coordinates (hg38)	TFBS	P-value	Conditional P-value	Gene promoter
European	MPV	HSC	chr12:121789119-121790240	ZEB1	1.33E-29	0.5036	*LINC01089, RHOF*
European	WBC	CD8	chr17:39863834-39864579	STAT6	3.06E-10	0.4858	*IKZF3, ZPBP2*
European	WBC	Mono	chr17:40062711-40063806	CTCF	2.43E-15	0.8519	*THRA*
European	MCV	MPP	chr7:100626407-100627165	CTCFL	8.63E-14	0.09284	*TFR2*
	MCH				1.14E-13	0.04977	
	RBC				2.23E-12	0.4069	
European	MCH	Erythro	chr7:100642442-100642911	FLI1	3.79E-15	0.01461	*TFR2*
	MCV				4.47E-15	0.9462	
	RBC				4.09E-13	0.2606	
European	WBC	CMP	chr7:28684549-28685114	STAT1	1.39E-13	0.2477	*CREB5*
African	RBC	MEP	chrX:153945909-153946614	CTCFL	1.47E-18	8.91E-05	*RENBP*
	RDW				2.12E-14	1.45E-07	
	MCV				2.77E-12	0.0001713	
Hispanic	RBC				5.56E-10	3.27E-05	
	RDW				9.65E-14	0.0001542	
	MCV				8.08E-11	2.51E-06	

Figure legends

Figure 1. Main processing sequence and data types of find-tfbs. “(X,Y)” is a pair of X and Y: for instance, (Person, Allele) identifies one of the haplotypes of one person. “[X]” is a list of X: for instance, [Variant] contains a list of variants. “X -> Y” represents a hash table with keys of type X and values of type Y, and “TFBS -> [(Person, Allele)]” classifies haplotypes by the binding sites they contain. The keys of a hash table are unique by definition. The sequence within the grey area can be processed independently and in parallel for each locus.

Figure 2. Three genetic variants located in the promoter of *RENBP* and included in an open chromatin region found in megakaryocyte-erythroid progenitor (MEP) cells change the number of CTCFL binding sites and associate with red blood cell (RBC) traits. (**A**) The top panel shows gene annotations at the locus as well as ATAC-seq peaks in MEP. In the bottom panel, we zoom-in two sub-regions in the *RENBP* promoter. The region on the left includes a variant (X:153946252_G_A) where the alternative allele disrupts a CTCFL binding site in the reference sequence (PWM scores: G=-0.188, A=-2.06). The region on the right includes two variants: X:153946428_C_T disrupts a CTCFL motif (PWM scores: C=0.04, T=-0.96) whereas rs7889328 creates a TFBS (PWM scores G=-1.13, A=-0.65). (**B**) Normalized RBC count (y-axis) per number of CTCFL motifs (x-axis) found in the promoter of *RENBP* in women. We summed the number of CTCFL binding sites found in both haplotypes. (**C**) As in (B) but in men. The number of CTCFL motifs is lower in men as RENBP is located on the X-chromosome.

Figure 1

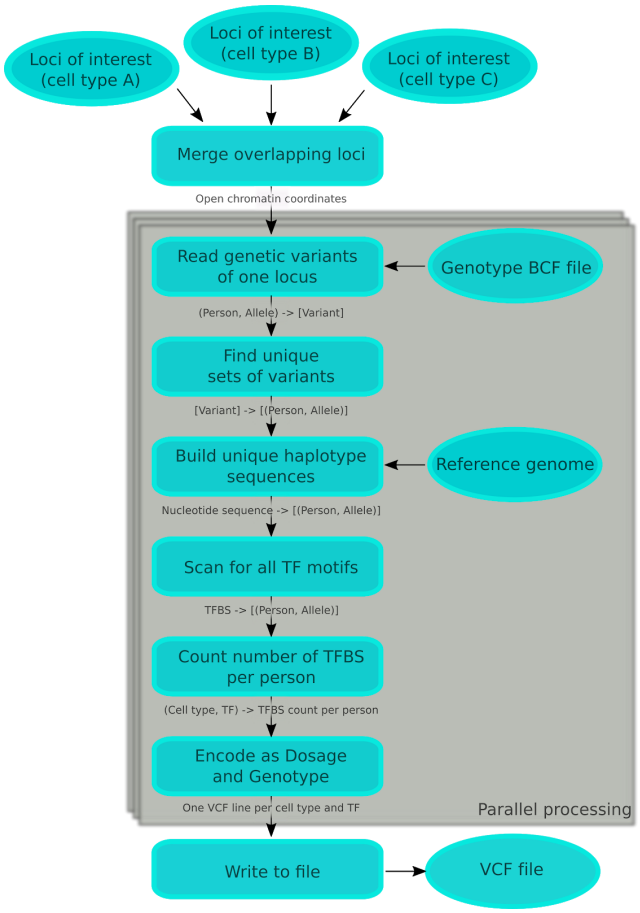
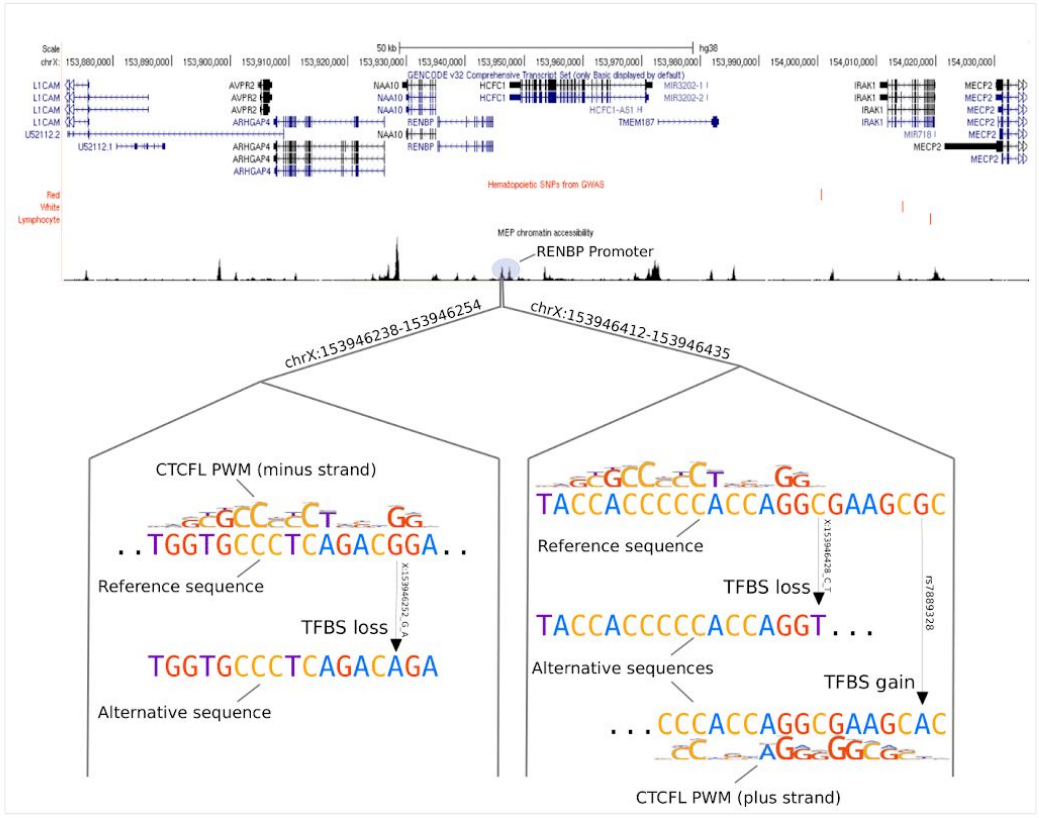
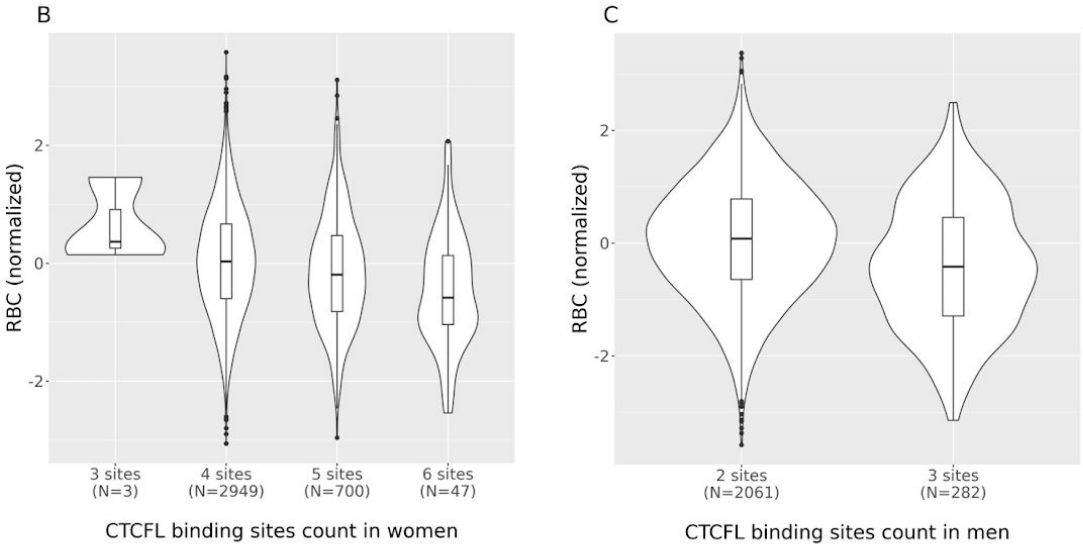

Loci of interest (cell type A)
Loci of interest (cell type B)
Loci of interest (cell type C)
Merge overlapping loci
Open chromatin coordinates
Read genetic variants of one locus
Genotype BCF file
(Person, Allele) -> [Variant]
Find unique sets of variants
[Variant] -> [(Person, Allele)]
Build unique haplotype sequences
Reference genome
Nucleotide sequence -> [(Person, Allele)]
Scan for all TF motifs
TFBS -> [(Person, Allele)]
Count number of TFBS per person
(Cell type, TF) -> TFBS count per person
Encode as Dosage and Genotype
One VCF line per cell type and TF
Parallel processing
Write to file
VCF file

Figure 2


RENBP Promoter
chrX:153946238-153946254
chrX:153946412-153946435
CTCFL PWM (minus strand)
..TGGTGCCCTCAGACGGA..
Reference sequence
TFBS loss
TGGTGCCCTCAGACAGA
Alternative sequence
TACCACCCCCACCAGGCGAAGCGC
Reference sequence
TFBS loss
TACCACCCCCACCAGGT...
Alternative sequences
TFBS gain
...CCCACCAGGCGAAGCAC
CTCFL PWM (plus strand)

A


B
RBC (normalized)
3 sites (N=3)
4 sites (N=2949)
5 sites (N=700)
6 sites (N=47)
CTCFL binding sites count in women
C
RBC (normalized)
2 sites (N=2061)
3 sites (N=282)
CTCFL binding sites count in men