Université de Montréal

Évaluation de l'imputation des données génétiques Canadiennes-Françaises

Par

Justin Pelletier

Département de Biochimie et Médecine Moléculaire, Faculté de Médecine

Mémoire présenté en vue de l'obtention du grade de M. Sc.

en Bio-informatique

Avril 2022

Université de Montréal

Département de Biochimie et Médecine Moléculaire, Faculté de Médecine

---

Ce mémoire intitulé

Évaluation de l'imputation des données génétiques Canadiennes-Françaises

Présenté par

Justin Pelletier

A été évalué par un jury composé des personnes suivantes

Sarah Gagliano Taliun
Présidente-rapporteur

Julie Hussin
Directeur de recherche

Despoina Manousaki
Membre du jury

Résumé

L'imputation est désormais un outil essentiel dans l'analyse des études d'association à l'échelle du génome, permettant l'estimation de génotypes à des positions variables du génome non génotypées, via des inférences statistiques à partir d'haplotypes contenus dans un panel de référence utilisé pour l'imputation, soit une bibliothèque d'haplotype séquencés phasés en haplotypes. Les données génétiques imputés servent aux études d'associations sur les traits et maladies complexes. La population fondatrice canadienne-française est une population très utile dans les études d'association génétique en raison de sa diversité unique d'haplotypes et de l'excès de variantes rares. Ici, nous décrivons les défis qui accompagnent l'imputation de cette population fondatrice, qui n'est pas représentée dans les panels de référence disponibles, ainsi que la stratégie optimale pour imputer des ensembles de données génotypés hétérogènes, provenant de plusieurs plateformes de génotypage. Nous avons caractérisé l'imputation de 29,356 individus génotypés sur plusieurs puces de génotypage de la province du Québec constituant la cohorte CARTaGENE (CaG). Nous avons établi que le panel de référence le plus récent et le plus diversifié *Trans-Omics for Precision Medecine* (TOPMed) a surpassé le panel de référence *Haplotype Reference Consortium* (HRC) dans l'ensemble de données canadienne-française de CaG. Nous avons évalué la précision de l'imputation avec le un score de qualité (R^2) fréquemment utilisé, ainsi que l'exactitude calculée en fonction des génotypes aux variants observés par séquençage, disponibles dans CARTaGENE pour un sous-groupe d'individus. Nous avons déterminé que la stratégie optimale pour augmenter la qualité d'imputation sur des ensembles de données hétérogènes a été atteinte en fusionnant chaque sous-ensemble de données après les avoir imputés individuellement. Ce résultat ouvre la voie à l'intégration de cohortes génotypées hétérogènes dans les études d'associations. Nos résultats soulignent également les défis que représente une population fondatrice pour l'imputation, en comparant la qualité de l'imputation de CaG avec d'autres sous-cohortes canadiennes du projet CanPath, soit l'Ontario, l'Alberta, la Colombie-Britannique et les provinces atlantiques. Ces résultats mettent en évidence l'impact de l'absence de diversité

haplotypique spécifique dans les panels de référence sur l'imputation d'une population européenne fondatrice récente, démontrant l'importance de la représentativité de la population étudiée dans ces panels.

Mots-clés : Imputation, Canadien français, Population fondatrice, Génotypage, Panels de référence, Génétique des populations, Bio-informatique, Génomique

Abstract

Imputation is now an essential tool in the analysis of genome-wide association studies, allowing the estimation of genotypes at variable positions of the ungenotyped genome, via statistical inferences from haplotypes contained in a reference panel used for imputation, (a library of sequenced genotypes phased into haplotype). Imputed genetic data is used for association studies of complex traits and diseases. The French-Canadian founder population is a very useful population in genetic association studies due to its unique haplotype's diversity and excess of rare variants. Here, we describe the challenges that come with imputing this founder population, which is not represented in available reference panels, as well as the optimal strategy for imputing heterogeneous genotyped datasets, from multiple genotyping platforms. We characterized the imputation of 29,356 individuals genotyped on multiple genotyping arrays from the province of Quebec constituting the CARTaGENE (CaG) cohort. We established that the newer and more diverse Trans-Omics for Precision Medicine (TOPMed) reference panel outperformed the Haplotype Reference Consortium (HRC) reference panel in the CaG French-Canadian dataset. We evaluated the precision of the imputation with the frequently used quality score (R^2), as well as the accuracy calculated according to the genotypes observed by sequencing, available in CARTaGENE for a subgroup of individuals. We determined that the optimal strategy for increasing imputation quality on heterogeneous datasets was achieved by merging each subset of data after imputing them individually. This result opens the way to the integration of heterogeneous genotyped cohorts in association studies. Our results also highlight the challenges of a founder population for imputation, comparing the quality of CaG imputation with other Canadian sub-cohorts of the CanPath project, namely Ontario, Alberta, British-Columbia, and the Atlantic provinces. These results highlight the impact of the absence of specific haplotypic diversity in the reference panels on the imputation of a recent European founder population, demonstrating the importance of the representativeness of the population studied in these panels.

Keywords : Imputation, French-Canadian, Founder population, Genotyping, Reference panels, Population genetic, Bioinformatics, Genomic

Table des matières

Liste des tableaux

Liste des figures

Liste des sigles et abréviations

GWAS	Étude d'association pangénomique (*Genome wide association study)*
IBS	*Identical By Descent*
HRC	*Haplotype Reference Consortium*
HapMap	*International HapMap Project*
VCF	*Variant Calling Format*
Pb	Paire de Base
MAF	Fréquence de l'allèle Mineur (*Minor Allele Frequency*)
ML	Machine learning
HW	Hardy-Weinberg
FDR	Taux de fausses découvertes (ou *False Discovery Rate*)
HLA	*Human Leukocyte Antigen*
cM	Centi-Morgan
Indel	Insertion ou Délétion

CNV	Variants du Nombre de Copies
IBD	*Identical By Descent*
SNP	*Single Nucleotide Polymorphism*
SNV	*Single Nucleotide Variants*
LD	Déséquilibre de Liaison (ou *Linkage Desequilibrium*)
CF	Canadiens-Français
FC	*French-Canadian*
ADN	Acide désoxyribonucléique
NGS	Séquençage de nouvelle génération (ou *Next Generation Sequencing*)
GH	*Genotype Harmonizer*
MIS	*Michigan Imputation Server*
TIS	*TOPMed Imputation Server*
QCRef	Panel de Référence du Québec
CaG	CARTaGENE
1000G	Projet des 1000 Génomes (ou *The 1000 Genomes Project*)

PCA	*Principal Componant analysis*
QC	Contrôle de Qualité (ou *Quality control*)
HMM	Modèle de Markov caché (ou *Hidden Markov Model*)
CanPath	*Canadian Partnership for Tomorrow's Project*
OHS	*Ontario Health Study*
BCGP	*BC Generations Project*
ATL	*Atlantic Partnership for Tomorrow's Health*
ATP	*Alberta's Tomorrow Project*
MTP	*Manitoba Tomorrow's Project*
HFS	*Healthy Future Sask*
TOPMed	*Trans-Omics for Precision Medicine*
WES	*Whole Exome Sequencing*
WGS	*Whole Genome Sequencing*

Remerciements

Je voudrais tout d'abord remercier Julie Hussin de m'avoir accueilli dans son laboratoire depuis les 4 dernières années. D'avoir vu en moi la fibre d'un bio-informaticien alors que je n'étais encore qu'un néophyte à ma deuxième année de Baccalauréat. D'avoir été un modèle de chercheur engagée dans son milieu et voulant plus que tout le bien-être de ses étudiants et proches. Je la remercie de sa compréhension et de sa confiance en moi qui m'ont permis de m'épanouir autant dans ma vie personnelle que professionnelle.

Je remercie également Sarah Gagliano Taliun d'avoir prêté intérêt à mon projet en acceptant d'être sur mon jury de mémoire et de m'avoir parrainé mon projet dans un domaine dans lequel est les plus que compétente. Un remerciement spécifique à ses conseils qui m'ont orienté dans mon projet.

Un remerciement à Despoina Manousaki d'avoir accepté d'être membre du jury pour mon projet de mémoire de maîtrise.

Une pensée pour tous les membres du laboratoire que j'ai côtoyé tout au long de ces années et qui ont su m'aider dans mes recherches et me divertir le temps d'un midi. Un remerciement spécifique pour Jean-Christophe Grenier qui depuis ses débuts dans le laboratoire est un atout majeur pour les étudiants grâce à sa vaste connaissance du domaine. Participant activement dans mon projet et me conseillant semaine après semaine. Je le remercie spécialement pour son assiduité et ses réponse aux messages Slack peu importe l'heure du jour et peu-importe qu'il travaille ou qu'il soit en congé.

Je remercie Élaine Meunier pour son efficacité et son encadrement lors de mon BAC et ma Maîtrise. De sa patience et de ses rappels de dates limites sans lesquels j'aurais probablement manqué plusieurs événements.

Je tenais également à remercier Marie Pageau de m'avoir fait confiance et de m'avoir supporté en tant qu'auxiliaire d'enseignement pendant les 4 dernières années.

Un remerciement pour la Faculté des Études Supérieure et BioTalent Canada qui ont subventionné ma maîtrise.

Je termine en remerciant tous mes amis et proches de m'avoir encouragé durant toute ces années et d'avoir écouté mes longs monologues sur des sujets avec lesquels ils n'étaient pas toujours très familier. Votre curiosité et vos questions m'ont poussé à me surpasser au fils des ans. Remerciement spécial à mes parents qui m'ont toujours encouragé à pousser mes ambitions et qui m'ont donné un moyen de les atteindre. Sans vous tous, je ne serais peut-être pas le bio-informaticien et l'homme que je suis aujourd'hui.

Introduction

La nouvelle ère du séquençage à haut débit a permis plusieurs avancées significatives dans la compréhension des causes génétiques des maladies humaines. Les découvertes rendues sur les traits complexes et maladies rares trouvée par études d'association sur le génome entier, dit GWAS (pour *Genome-Wide Association Studies*) ouvrent la voie à l'implantation de la médecine personnalisée, la médecine préventive et la pharmacogénomique. Cependant, malgré les avancées technologiques marquantes des dernières années, le séquençage du génome entier demeure dispendieux. L'alternative moins coûteuse au séquençage se trouve ainsi dans le génotypage, qui consiste à caractériser certains variants déjà répertoriés dans le génome grâce aux puces de génotypages.

Cette technique permet ainsi, d'obtenir les données génétiques « partielles » d'un plus grand échantillon d'individus pour les mêmes couts moins élevés que le séquençage du génome. Puisque la taille d'échantillon est directement corrélée au pouvoir statistique, l'utilisation de données de génotypage d'un plus grand nombre d'individus permet, conséquemment, d'augmenter le nombre de découvertes dans les études d'associations GWAS. Le génotypage ne caractérisant qu'une fraction du génome, l'inférence statistique des données génétiques manquantes (ou imputation) est nécessaire. Cette stratégie utilise un panel de référence qui contient de nombreux haplotypes (série de variations d'ADN, ou polymorphismes, qui ont de fortes chances d'être hérités ensemble) issus de données de séquençage. Plusieurs panels de références sont disponibles publiquement et contiennent des échantillons issus de diverse populations humaines. Avant 2021, la majorité des individus inclus dans les panels de référence étaient d'origine Européenne et en minorité Africaine (1). Cela fait en sorte que la diversité haplotypique qu'ils renferment ne reflète pas l'entièreté de la diversité haplotypique humaine existante, rendant plus difficile l'imputation d'individus d'origines sous-représentées dans le panel utilisé. La publication, en 2021, du plus grand panel jamais constitué, le *Trans-Omics for Precision Medicine* (TOPMed) (2), visant à diversifier l'origine des haplotypes, démontre une amélioration sans équivoque de la qualité

d'imputation des données génotypées diverses (2). Ceci est rendu possible en incluant des individus d'origine Océanienne, Asiatique et d'Amérique centrale et du sud, d'Amérique du Nord (origine européenne métissée), en plus des Européens et Africains qui sont traditionnellement présents. Malgré l'augmentation de la diversité dans les panels d'imputation, des problèmes demeurent. Certaines études ont d'ailleurs démontré que l'imputation de populations non comprises dans le panel de référence utilisé pouvait entrainer une faible qualité d'imputation (1, 3-6). Le défi principal de l'imputation demeure dans l'inférence des variants à faibles fréquence dans la population (7). Leur intrinsèque rareté se manifeste autant dans les haplotypes utilisés pour faire l'imputation que dans les données utilisées en GWAS. Ces variants sont très importants puisqu'ils contribuent grandement dans les causes des maladies complexes (8-11).

Les populations fondatrices telles que la population Canadienne Française (CF) constituent un attrait en génétique des populations et médicale. Elles sont très utiles à la découverte de nouveaux variants associés aux maladies génétiques complexes lors des études d'association pangénomiques puisque plusieurs variants rares ont vu leur fréquence allélique augmenter dans ces populations. La population fondatrice CF, issue majoritairement d'un groupe de fondateurs venus de France il y a entre 9 et 17 générations (12), est prisée pour réaliser des analyses de type GWAS. De plus, une cohorte Québécoise presqu'entièrement génotypées, majoritairement CF, nommée CARTaGENE (CaG) (13) est disponible. Toutefois, la qualité potentielle de l'imputation de cette population demeure incertaine, dû à la diversité haplotypique différente de sa population Européenne ancestrale (14).

Ce mémoire se divise en quatre parties. Le Chapitre 1 présente une revue de la littérature portant sur les notions génétiques, populationnelles, statistiques et techniques permettant de situer nos connaissances actuelles sur l'imputation de données génétiques ainsi que de faciliter la compréhension du projet sur la caractérisation de l'imputation des données génétiques Canadiennes-Française. J'y présente les jeux de données utilisés dans le projet. Le Chapitre 2 présente une section discutant des hypothèses et des objectifs de mon projet. Au Chapitre 3, je présente un article scientifique reportant les résultats originaux obtenus sur la caractérisation de l'imputation dans la cohorte CaG avec les techniques bio-informatiques de pointe en génomique humaine. Finalement, au Chapitre 4, je présente une analyse et discussion de ces résultats et une conclusion énonçant les perspectives conclura le mémoire.

Chapitre 1 - Revue de la littérature

1.1 Génération de données génétiques

1.1.1 Séquençage du génome entier

Le séquençage d'ADN est une méthode servant à déterminer l'ordre de la suite d'acide nucléique qui compose la séquence d'ADN (15). Grâce aux techniques de séquençage, il est possible d'identifier les variants, ou mutations, présents dans l'ADN qu'importe leur type. On parle entre autres de *Single Nucleotide Polymorphism* (SNP) présents dans au moins 1% de la population, de *Single Nucleotide Variant* (SNV) pouvant avoir une fréquence de moins de 1% , de courtes insertion/délétions (indels) ou encore de variation du nombre de copies d'un fragment d'ADN (CNV). Le séquençage peut servir à déterminer la suite de nucléotides de plusieurs types de séquences de différentes longueurs. Il peut s'agir d'une courte séquence, de certaines parties du génome tels que les exons (exome) ou du génome complet.

La première technologie de séquençage développée est le séquençage Sanger qui est basé sur l'incorporation sélective de didésoxynucléotides de terminaison par l'ADN polymérase lors de la réplication de l'ADN in vitro (16). Cette méthode permet d'obtenir des lectures de plus de 500 paires de base et a typiquement un taux d'exactitude de 99,99% (17). Ces raisons expliquent qu'elle est encore utilisée aujourd'hui pour les projets de plus petite taille et pour confirmer les résultats des nouvelles technologies de séquençage (16). Les désavantages de cette techniques sont le temps et les coûts engendrés pour des expériences non-ciblées (17).

Une nouvelle génération de technologie de séquençage (NGS) est venue remédier partiellement aux limites du séquençage Sanger, il s'agit des technologies de séquençage à haut débit. Ces technologies comprennent le séquençage par synthèse (18), par ligature (19), par semi-

conducteur ionique (20) et plus encore. La plus communément utilisée est la technologie de séquençage par synthèse employée par Illumina (21). Le séquençage Illumina fonctionne en identifiant simultanément les bases d'ADN se liguant à la séquence d'ADN à séquencer par leur émission d'un signal fluorescent unique. Chaque brin d'ADN initialement fragmenté, hybridé à la puce micro fluidique, puis amplifié est ensuite séquencé grâce à l'incorporation de nucléotides fluorescent émettant une longueur d'onde et une intensité caractéristique lors de l'excitation à chaque cycle d'extension de la chaîne par synthèse (21). Bien que le séquençage Illumina se base sur la technique du séquençage Sanger, les deux techniques se distinguent par l'ordre des étapes des manipulations ainsi que par l'étape de restauration. En effet, le séquençage Sanger permet d'obtenir généralement une séquence d'ADN alors que les NGS peut produire jusqu'à 250 millions de lectures (*reads*) (21). De plus, le haut débit de séquençage permet de séquencer beaucoup plus rapidement des longues séquences et pour un coût moindre.

Ces nouvelles technologies (NGS) ont révolutionné l'étude de la génétique et la biologie moléculaire (22). Le séquençage de l'ADN est maintenant indispensable dans les domaines biologiques tels que la médecine personnalisée, la biotechnologie, la virologie ou de biologie médico-légale.

1.1.2 Séquençage de l'exome

L'exome est l'ensemble des exons des gènes, soit la partie codante de l'ADN. Puisque cette partie du génome code directement pour des protéines, une mutation peut en altérer la séquence d'acides aminés. Le séquençage de l'exome entier (*Whole Exome Sequencing*, WES) permet donc de cibler les variants codants qui risquent d'avoir un impact direct sur un trait particulier ou une maladie (23).

Le séquençage de l'exome entier se fait en deux étapes. Tout d'abord, un enrichissement de l'ADN cible est effectué. Deux principales techniques sont utilisées soit la capture par puce contenant les séquences oligonucléotides simples-brins des exons humain (24) ou la capture en solution qui utilise des sondes se liant aux séquences exomiques (25). Ensuite les fragments d'ADN ciblés peuvent être séquencés à l'aide des multiples méthodes de séquençages à haut débit.

Ces techniques permettent d'obtenir efficacement les séquence codantes des gènes (26). Le séquençage ciblé des exons est grandement utilisé dans la recherche sur les maladies mendéliennes puisque celles-ci sont souvent causées par un ou quelques rares variants ayant un impact fonctionnel important. Malgré son coût inférieur au séquençage du génome complet (*Whole Genome Sequencing*, WGS), le WES ne permet pas de découvrir l'ensemble des variations ayant des impacts sur les traits et les maladies complexes (27).

1.1.3 Génotypage

1.1.3.1 Stratégie du génotypage

Le génotypage est une alternative plus économique, avec des prix variant de 75 à 500$ (CAD) par échantillon, et plus rapide que les méthodes de séquençage. Cette méthode consiste à identifier les allèles variants que possède un individu pour des positions données. Il est utilisé pour regarder plusieurs variants à la fois et est surtout utilisé pour des variants plus communs. Les puces de génotypage utilisent les SNPs qui sont des variants mutés d'un nucléotide présent dans au moins 1% de la population et sont conçues selon des critères populationnels et/ou pour des maladies communes ou rares en visant à caractériser les variants d'intérêt.

En comparaison avec le séquençage qui permet aussi de caractériser les SNPs chez les individus, le génotypage est beaucoup moins dispendieux et permet de facilement et rapidement génotyper directement les allèles aux sites d'intérêt. Toutefois, le génotypage ne donne pas directement d'information sur les sites autres que ceux génotypés. Le génotypage est utilisé pour de nombreuses applications, y compris le diagnostic clinique de maladies, GWAS, la cartographie fine des loci connus, les études de liaison (28) ou encore les CNV.

1.1.3.2 Puces de génotypages

La technique du génotypage utilise des puces à ADN, dont le fait appel à l'hybridation de l'ADN monocaténaire fragmenté à des puces contenant des centaines de milliers de séquences de sondes nucléotidiques uniques (29). Les puces de génotypage sont largement utilisées et éprouvées afin de caractériser les variants d'un individu à des sites préalablement caractérisés grâce au séquençage (29). La principale utilisation de ces puces est la caractérisation de SNPs, toutefois elles peuvent aussi servir pour les CNV et indels.

La conception des puces de génotypage vise des utilisations spécifiques, caractérisant uniquement certains loci sur quelques chromosomes plutôt que sur le génome entier. Effectivement, plusieurs compagnies, telles qu'Illumina et Affymetrix, offrent des architectures de matrice avec un nombre fixe de variants qui ont été sélectionnés dans l'objectif de caractériser les SNPs qui impactent des phénotypes observés (28) ou permettent de faire de la caractérisation populationnelle. Certaines puces sont aussi conçues pour des populations spécifiques en tirant profit des schéma de déséquilibre de liaison, soit l'association non aléatoire de loci sur un chromosome dépendant de la distance génétique (voir section 1.1.3.3), pour étendre le plus possible la couverture du génome (30), toutefois, ce dernier type de puce est souvent plus dispendieux.

Les puces de génotypage sont utilisées pour toutes les raisons mentionnées plus haut dans les GWAS (28) qui nécessitent un grand nombre d'échantillons afin d'avoir un plus grand pouvoir statistique (29). L'objectif du génotypage étant de limiter le nombre de variants à caractériser directement, il omet cependant de génotyper certains variants d'intérêt. De manière générale, les puces sont conçues en se basant sur les variants des populations Européennes (en raison du contexte socio-économique ayant mené à plus de découvertes de variants dans ces cohortes) qui ne sont pas représentatifs de l'ensemble de la variabilité existante mondialement (3). La conception de ces puces doit prendre en compte plusieurs facteurs génétiques afin d'optimiser les découvertes qui découlent des données produites.

1.1.3.3 Recombinaison Génétique

Un des phénomènes qui affecte la diversité génétique est la recombinaison génétique. Ces événements impliquent l'échange de matériel génétique entre plusieurs chromosomes ou entre différentes régions du même chromosome. Le processus s'effectue généralement dans des régions homologues des chromosomes. Chez l'humain (organisme diploide[1]), la recombinaison se fait entre les chromosomes homologues lors de la méiose, où un enjambement entre deux chromatides entraine un échange de matériel génétique (31).

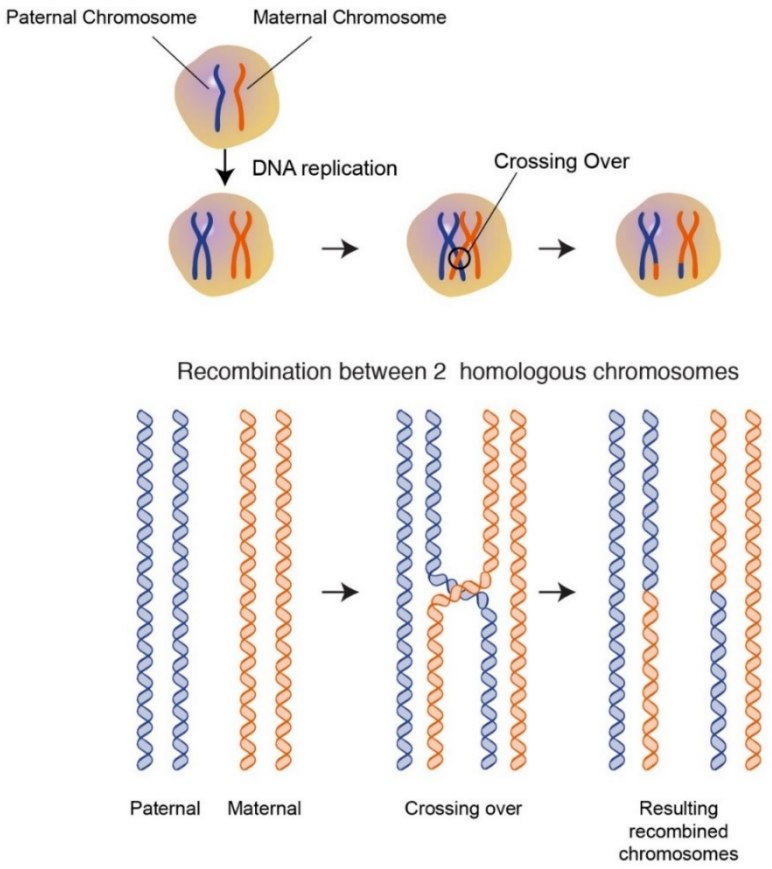


Figure 1.1. Illustration de la recombinaison par enjambement méiotique (32) (https://www.genome.gov/genetics-glossary/homologous-recombination)

[1] Cellule qui contient deux copies de chaque chromosomes (2n)

Les chromosomes résultants sont des croisements entre les chromosomes parentaux, contribuant ainsi à diversité génétique par la création de nouvelles combinaisons de variants. Cette suite de variations dans l'ADN (combinaison de SNPs ou d'allèle[2]) qui est héritée ensemble, parce qu'elle se trouve sur le même chromosome, est nommée haplotype. Les nouveaux haplotypes issus de la recombinaison verront leurs fréquences varier dans la population en fonction des différentes forces de sélections (33).

La recombinaison est quantifiée grâce à un taux de recombinaison *r* mesuré en Centimorgan par Mégabase (cM/Mb), soit la distance génétique entre deux loci (le pourcentage de chance que les loci sur le même chromosome soient séparés par un événement de recombinaison pour une génération) par million de pair de bases (pb). Cette distance génétique entre deux loci varie le long du génome. Certaines régions chromosomiques sont moins susceptibles à recombiner et sont nommées *coldspots* de recombinaison alors que d'autres, de 2-5Kb, recombinent hautement et sont nommées *hotspots* (34). Ainsi, les allèles se trouvant avant et après un *hotspot* de recombinaison sont moins souvent observés conjointement sur un même haplotype. Pour des SNPs qui se retrouvent au sein d'un même *coldspot*, les associations d'allèles sont alors rarement cassées et ils sont transmis plus souvent ensemble.

1.1.3.4 Déséquilibre de liaison

Le déséquilibre de liaison (ou LD pour *Linkage Desiquilibrium*) se définit comme étant l'association non-aléatoire d'allèles à différents loci (35), soit une deux loci qui sont hérités ensemble à une fréquence différentes de ce qui serait attendu pour deux loci indépendants. Le LD est une conséquence de la présence et de la fréquence d'événements de recombinaison méiotique (36). Ce phénomène permet alors de séparer des variants, créant de nouveaux haplotypes, ou alors de garder certains variants ensemble sur le même haplotype. On dit alors qu'il y a déséquilibre de liaison lorsque la fréquence des allèles de deux loci est différente de ce que donnerait une association aléatoire de ces allèles (35).

[2] Versions d'une séquence d'ADN (un ou une suite de nucléotide) à un emplacement génomique donné.

Le déséquilibre de liaison est influencé par plusieurs facteurs tels que la sélection naturelle, le taux de recombinaison, le taux de mutations, la dérive génétique, l'appariement non-aléatoire ou encore la structure de population (37). La combinaison de l'effet de ces facteurs crée différents patrons de déséquilibres de liaison selon les populations. Il existe des cartes génétiques qui rapportent la distance génétique entre les variants génomiques identifiés grâce au séquençage. Ces distances sont calculées en fonction de la fréquence des croisements chromosomiques se produisant pendant la méiose, mesurée en Centimorgan (cM), et non en fonction de leur emplacement physique sur le chromosome en paire de base (38). Ces cartes génétiques furent développées en premier grâce au projet génome humain (*Human genome project*) et au projet international HapMap (39) et sont perfectionnées grâce à l'information provenant des nouvelles technologies de séquençage (NGS) (15, 38).

Tel que mentionné dans la section 1.1.3, la connaissance des patrons de LD est mise à profit lors d'études GWAS autant dans la conception des puces de génotypage que dans l'imputation des variants non-caractérisés. En effet, puisque les puces de génotypages qui sont utilisées dans les études GWAS doivent caractériser le plus de variations au travers du génome entier en génotypant le moins de variants possible, le LD permet d'identifier des variants qui serviront de proxy pour les variants qui les entourent sans avoir à les génotyper directement.

1.1.3.5 HapMap

Le projet international HapMap est le premier projet de grande envergure visant à caractériser la variation génétique humaine. L'objectif de ce consortium était de déterminer les modèles communs de variation de séquence d'ADN dans le génome humain en caractérisant les variants, leurs fréquences et leurs corrélations entre eux (39, 40). Le projet utilisait initialement 180 échantillons d'ADN provenant de trios familiaux de populations Européennes et Africaines ainsi que 90 échantillons non apparentés Asiatiques(39). Une première étape de séquençage ciblé d'un sous-ensemble de variants suivi d'un génotypage à l'aide d'un total de cinq technologies de génotypage à haut débit furent utilisées afin de comparer leur précision, leur efficacité et leurs coûts. La motivation pour l'utilisation du génotypage fût de réduire les coûts et le temps d'obtention de résultats sur les variants au travers du génome. Pour ce faire, l'approche utilise les

informations d'un ensemble relativement petit de variants qui capturent la plupart de la variation dans le génome (proxy), de sorte que toute région ou gène puisse être testé pour une association avec une maladie particulière (39). La dernière version de HapMap (phase 3) répertorie plus de 1.6 millions de SNPs tirées du génome de 1184 individus de 11 populations différentes (41). Ce premier projet d'envergure a inspiré d'autres cohortes de génotypage diversifiées telles que le *Human Genome Diversity Project* (HGDP) (42), le *Population Reference Sample* (POPRES) (43) et le *1000 Genomes Project* (1000G) (44) visant à caractériser et à comprendre la diversité génétique humaine.

1.2 Études d'association pangénomique (GWAS)

Les GWAS sont un outil majeur pour identifier des variants génétiques qui confèrent une susceptibilité aux traits complexes. Elles testent une corrélation entre l'état de la maladie ou d'un phénotype et la variation génétique pour identifier les régions du génome qui contribuent à ce trait. Une association génétique significative peut être interprétée de deux manières. Soit, comme une association directe, dans laquelle le SNP identifié est le véritable variant causal conférant une susceptibilité à la maladie, soit comme une association indirecte, dans laquelle un SNP en déséquilibre de liaison avec le vrai variant causal est identifié. Il est aussi possible qu'un résultat soit un faux positif : cette fausse association peut être due au hasard ou à une confusion systématique. Toutefois, certaines stratégies existent afin de palier à ce type d'erreur. Les tests multiples tels que la correction de Bonferroni (45), l'utilisation de valeurs P ajustées pour contrôler le taux de fausses découvertes (*False Discovery Rate*, FDR) ou l'utilisation des probabilités a posteriori pour contrôler par l'approche FDR bayésienne (46).

Une approche utilisée initialement, lorsque les jeux de données étaient de petite taille, est l'approche par gène candidat qui requiert une hypothèse a priori sur le mécanisme biologique du trait ou de la maladie observée. Un gène candidat est présumé associé à une maladie particulière ou à un trait phénotypique. Les fonctions biologiques de ce gène sont dérivées d'autres études (comparatives, clonage positionnel, etc.) (47). Bien que cette méthode soit maintenant critiquée en raison du haut taux de non-reproductibilité et de son caractère trop ciblé, elle fut à l'origine de

plusieurs découvertes (47, 48). Cette stratégie est surtout efficace pour identifier la cause des maladies mendéliennes classiques qui sont causées par un ou quelques variants très rares dans un seul gène. Toutefois, lorsque l'on parle de maladies ou de traits phénotypiques complexes (qui sont influencées par plusieurs variants qui modulent le risque), cette approche est peu efficace puisqu'elle ne considère qu'une partie du génome, souvent choisie de façon biaisée. Une approche exploratoire, n'étant basée sur une hypothèse de départ, est nécessaire pour découvrir de nouveaux mécanismes biologiques. Une alternative aux études d'association gène candidats est l'étude de liaison plus traditionnelle basée sur des liens familiaux pour cartographier les variants génétiques qui sous-tendent les maladies humaines courantes (49-51). Cette stratégie requiert toutefois d'avoir les données génétiques et le pedigree familial.

Les technologies de génotypage et séquençage NGS mentionnées précédemment (Section 1.1.1) ont grandement contribué à l'émergence des études d'association pangénomique ou GWAS. Ce type d'étude d'association pangénomique tente d'établir des associations entre de nombreux variants génétiques et certains traits phénotypiques. D'un point de vue biologique, les maladies complexes sont, contrairement aux maladies mendéliennes, caractérisées par un impact faible de plusieurs variants rares et communs (8, 9) classifiés par des fréquences respectives de moins de 1% et de plus de 1% dans la population. Les traits complexes étant déterminés par une multitude de loci ayant chacun un impact plus ou moins important sur le phénotype étudié, il est nécessaire de travailler avec une grande taille d'échantillon pour atteindre une puissance statistique significative. Étant donné la contrainte de grande taille d'échantillon pour une étude GWAS, elles ont été principalement réalisées en utilisant des cohortes européennes importantes et bien caractérisées, laissant beaucoup de variations génétiques et de diversité non-explorées afin de comprendre l'architecture génétique de traits complexes (52).

L'avancement des bases de données de SNPs de génomes entiers, des catalogues substantiels de variations d'haplotypes et le progrès dans les technologies de séquençage et de génotypage (49) ont étés nécessaires à l'apparition des études GWAS. Aujourd'hui ces études peuvent utiliser un ensemble de marqueurs, quelques centaines de millier de SNPs, à travers le génome des individus atteints et non-atteints à l'aide de puces commerciales de génotypage. Les SNPs typés sont ensuite testés un par un avec le phénotype d'intérêt afin de donner une valeur d'association (51, 53). Puisqu'il est peu probable que l'ensemble de SNPs sur la puce inclut le

véritable variant causal, cette approche considère l'utilisation des marqueurs sur la puce de génotypage comme prédicteurs de variants non typés mais génétiquement associés aux variants identifiés (à cause du LD) (49, 54). Un défi majeur dans ce domaine est de discriminer les variants causaux des variants en LD avec les variants associés. Un moyen de gagner de la valeur ajoutée pour l'analyse consiste à combiner les informations entre les marqueurs génotypés et les catalogues de variations existants tels que HapMap (39, 51). En effet, ce projet a contribué grandement à l'avancement des découvertes par GWAS en répertoriant les haplotypes existants dans la population humaine. Il a permis à la fois de trouver le variant possiblement causal en LD avec les variants significatifs trouvés dans les GWAS (55) et d'améliorer les données génétiques utilisées pour faire les GWAS.

Cependant, la structure de corrélation qui existe entre les variants d'ADN dans le génome humain actuel, de par les patrons de LD, cependant très importante dans l'analyse des résultats des GWAS. En raison des forces évolutives, en particulier la taille de la population, les mutations, le taux de recombinaison et la sélection naturelle (53), cette structure génomique varie entre les populations humaines. Ces facteurs font en sorte que peu de résultats GWAS sont transposables d'une population humaine à l'autre et que la réplication est difficile lorsque l'on change d'origine ethnique sur laquelle le GWAS est fait (56). Les populations métissées, bien que sous-représentées dans les GWAS (56-58), présentent une opportunité intéressante pour capturer l'architecture génétique de plusieurs populations et donc d'augmenter le pouvoir statistique (59). En contrepartie, la structure des populations métissées, si elle n'est pas correctement considérée et intégrée, peut entraîner des signaux de corrélation erronés et donc des taux de faux positifs plus élevés (60).

1.3 Imputation

L'imputation permet d'obtenir l'information de plusieurs SNPs non-génotypés à partir d'un sous-ensemble de SNPs génotypés par des puces de séquençage (61). Cette méthode se base sur les haplotypes existants dans une cohorte de référence pour l'imputation. Grâce à ces haplotypes et au concept de déséquilibre de liaison, elle peut inférer les variants se situant dans un bloc de LD proche des variants typés. Les facteurs importants à considérer pour augmenter la performance de

l'imputation et sa qualité sont nombreux. Tout d'abord, il est important que les SNPs obtenus par la puce de génotypage soient répartis uniformément dans le génome et que leur qualité soit bonne. Ceci permet d'avoir de l'information sur l'ensemble du génome et donc de ne manquer aucune région lors de l'imputation. Ensuite, il faut disposer ou constituer une cohorte de référence pour l'imputation qui soit d'origines diversifiées (différents haplotypes et patrons de LD), la plus grande possible et de meilleure qualité possible. Finalement, il faut que le pré-phasage (l'inférence d'haplotypes à partir des données génotypées) des données génétiques ainsi que l'imputation se fasse avec un logiciel et une méthode statistique qui conviennent le mieux à une grande cohorte de référence. Chacun de ces facteurs influence grandement la qualité d'imputation et la capacité à détecter les variants ayant un effet sur les maladies génétiques dans les GWAS.

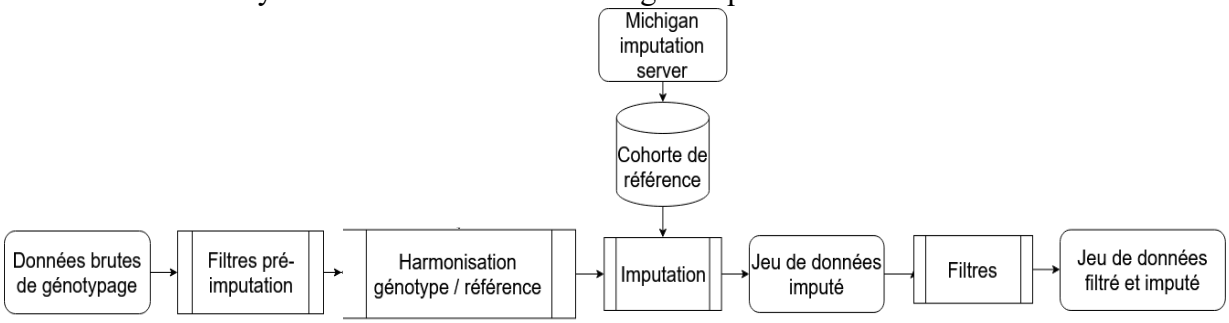


Figure 1.2. Pipeline d'imputation développé pour les données génotypées. Ce pipeline inclus celui du Michigan Imputation Server (62).

L'imputation en génétique est une technique permettant de complémenter les variants génotypés en utilisant l'information documentée dans des catalogues de variants génomiques et les cartes génétiques (49). Le principe de l'imputation est d'utiliser le LD qui existe entre les différents variants du génome humain pour inférer (imputer) les allèles aux sites manquants grâce aux sites observés sur les puces de génotypage utilisées (49) (Fig. 1.3). À chaque fois qu'un tronçon particulier de chromosome est examiné en détail chez au moins un individu, nous apprenons les génotypes de nombreux autres individus qui héritent de ce même tronçon, identique par descendance (ou IBD pour «*identical by descent*») (63). Les tronçons d'haplotypes partagés sont plus longs lorsque l'on impute des individus apparentés que des individus non-apparentés (63). Ce phénomène se justifie par le fait que les ancêtres communs sont plus éloignés dans le temps et en termes de générations, ce qui augmente la diversité des haplotypes (recombinaisons et mutations) (64). Les régions haplotypiques partagées sont plus courtes et peuvent donc être plus difficiles à identifier avec confiance (63). L'avantage de l'imputation est l'utilisation d'un modèle de

génétique des populations approximatif qui donne plus de poids aux génotypes cohérents avec les modèles locaux de LD (3, 65).

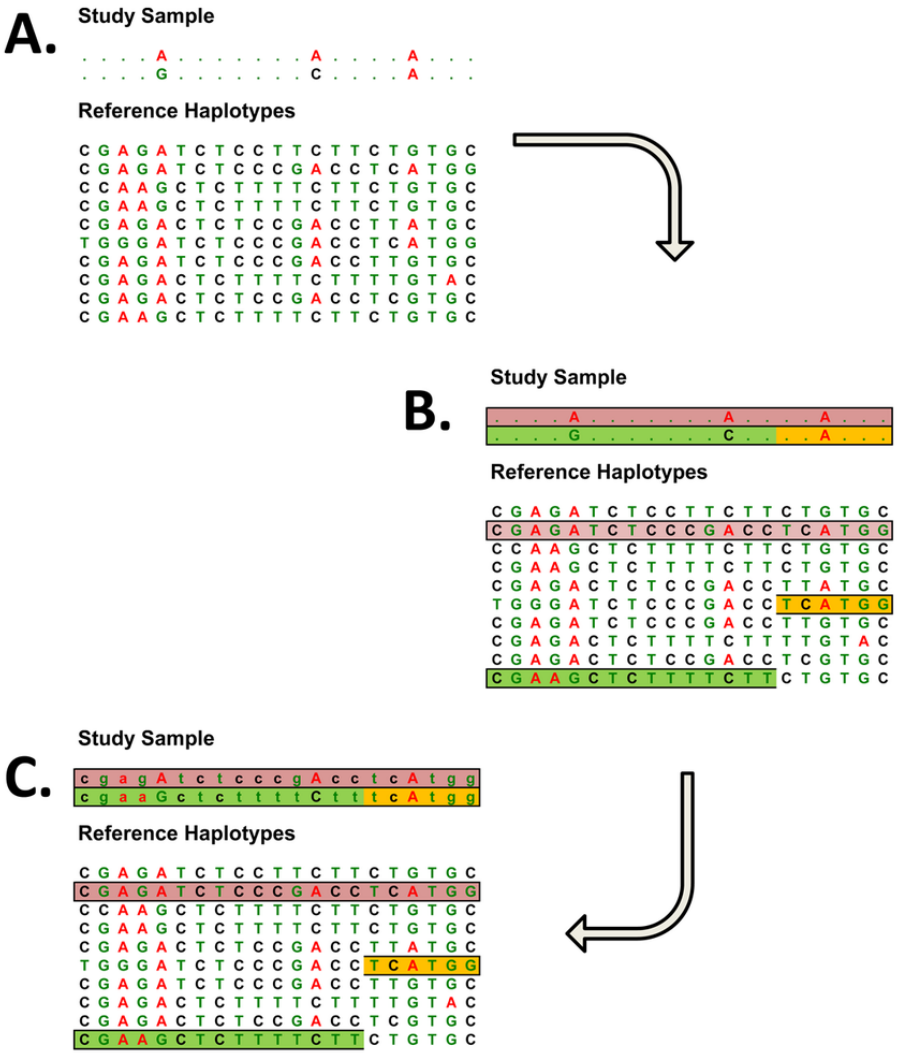


Figure 1.3. Schéma d'imputation du génotype d'individus non-apparentés (63).

Les haplotypes phasés sont représentés par des suites horizontales de nucléotides (chaque ligne est un haplotype). Les haplotypes contenant des données manquantes sont issus du génotypage (*Study Sample*) sont comparés aux haplotypes contenus dans les cohortes de référence d'imputation (*Reference Haplotypes*). En identifiant les blocs haplotypiques auquel les variants génotypés ont le plus de chance d'appartenir, il est possible d'inférer, avec les haplotypes de référence, les nucléotides aux sites manquants (66).

1.3.1 Harmonisation

Deux problèmes persistent encore en ce qui concerne les données génotypées. D'une part, le grand nombre de formats de fichiers utilisés par la communauté en génétique. D'autre part, l'ambiguïté du brin d'ADN des polymorphismes nucléotidiques. L'ADN est composé de deux brins antiparallèles et complémentaires pour les bases A-T et C-G. Ces deux brins sont nommés *foward* et *reverse*, représentant respectivement les brins 5'→3' et 3'→5'. Les différentes techniques de génotypage caractérisent uniquement l'un des deux brins, choisi de manière intentionnelle ou accidentelle (67). Pour ajouter au problème, la plupart des formats de fichiers ne définissent pas le brin utilisé. Pour certains types de SNPs, il est simple de détecter et de corriger les différences de brins. Par exemple, un SNP A/C à un complément, sur l'autre brin T/G et est donc non-ambigu. Cependant, les variants G/C et T/A sont ambigus car leurs allèles complémentaires sont respectivement C/G et A/T et le brin est plus difficilement discernable.

L'imputation de génotype étant généralement effectuée sur de multiples jeux de données, il est essentiel de tenter d'harmoniser l'ensemble des variants genotypés. Une des méthodes d'harmonisation consiste à comparer l'allèle mineur entre deux jeux de données. Cette technique n'est toutefois pas idéale car l'allèle mineur peut différer entre deux jeux de données et deux populations, spécialement pour les variants communs. Le plus récent outil permettant de remédier au problème de brins se nomme *Genotype Harmonizer* (GH) (67), qui permet de lire presque tous les formats de fichiers de données génétiques (PLINK, binary PLINK, VCF, SHAPEIT2 & Oxford GEN). Cet outil aligne l'échantillon à l'étude à une référence spécifié sans connaissance à priori du brin utilisé lors du génotypage. La méthode automatisée de GH attribue le brin des SNPs ambigus en sélectionnant des SNPs non-ambigus proches en LD à la fois dans les données à l'études et dans la référence. Plus spécifiquement, le logiciel corrèle les fréquences estimées d'haplotypes entre les données de l'étude et les données de référence dans un ratio positif ou négatif. Si le ratio présente plus de corrélations négatives que positives dans les fréquences d'haplotype, le SNP ambigu est permuté sur l'autre brin. Dans le cas où un SNP ambigu ne dispose pas d'assez de SNPs non-ambigus en LD, il sera exclu de l'ensemble. Une alternative au GH, consiste à utiliser la définition donnée par les compagnies desquelles sont issus les puces de génotypage qui précisent le brin sur lequel est situé les variants. Une dernière méthode utilise une suite d'outil du groupe

McCarthy qui est conseillée par plusieurs serveurs d'imputation et harmonise les données de génotypage en comparant les allèles avec une cohorte de référence selon les fréquences calculées dans le 1000G (https://www.well.ox.ac.uk/~wrayner/tools/). L'harmonisation est désormais la première étape dans l'imputation des données pour un pipeline de méta-analyses GWAS ou d'imputation (68).

1.3.2 Phasage

Les données issues du séquençage ou du génotypage sont dites non-phasées. La grande quantité d'informations contenues dans ces données est, toutefois, mieux exploitée grâce à des haplotypes en phase, qui identifient les allèles qui sont colocalisés sur le même brin chromosomique. Les allèles que l'on obtient grâce aux puces de génotypage ne sont pas assignées à l'un ni l'autre des chromosomes parentaux et ne sont donc pas associés à un haplotype (69). Cependant, l'information sur la phase des haplotypes est nécessaire à l'imputation (70). Les méthodes statistiques utilisées pour faire le phasage des données génétiques suivent le modèle multinomial simple, dans lequel chaque haplotype possible cohérent avec l'échantillon a reçu un paramètre de fréquence inconnu et ces paramètres ont été estimés avec un algorithme de maximisation des attentes (ML). Les chaînes de Markov cachées (HMM) (71) et les modèles haplotypiques applicables aux ensembles de données de grande taille (72-75) sont utilisés. Tous les logiciels de phasage utilisent ces méthodes et permettent de localiser l'ensemble des variants génotypés d'un individu sur un des haplotypes, permettant ainsi l'imputation des autres variants non-génotypés à partir de ceux-ci (61, 69). Les principaux logiciels utilisés aujourd'hui pour cette tache sont PHASE (71), ShapeIT (76) et Eagle2 (77).

1.3.3 Logiciels d'imputation

Les logiciels d'imputation peuvent être utilisés afin d'inférer les variants non génotypés. MACH (78), BEAGLE (79), Minimac (80, 81), IMPUTE v2 (49) sont des logiciels très efficaces et couramment utilisés pour faire l'imputation de données génomiques.

1.3.3.1 IMPUTE v2

IMPUTE v2 et Minimac qui sont basés sur une extension des modèles de Markov cachés (HMM) initialement développés dans le cadre de schémas d'échantillonnage d'importance pour simuler des arbres coalescents, pour modéliser le déséquilibre de liaison et estimer les taux de recombinaison (82). La méthode IMPUTE v1 est basée sur un HMM du vecteur de génotypes de chaque individu, G*i*, conditionnel à H, et un ensemble de paramètres. Ce modèle est décrit dans la formule :

$$P(G_i|H,\theta,\rho) = \sum_z P(G_i|Z,\theta), P(Z|H,\rho)$$

dans laquelle Z = {Z1, ..., ZL} avec Zj = {Z*j*1, Z*j*2} et Z*j*k = {1, ..., N}. Z*j* peut être considéré comme la paire d'haplotypes de la cohorte de référence au SNP *j* qui sont copiés pour former le vecteur génotype. Le terme P (Z | H, ρ) modélise comment la paire d'haplotypes copiés change le long de la séquence et est défini par une chaîne de Markov dans laquelle la commutation entre les états dépend d'une estimation de la carte de recombinaison (ρ) à travers le génome. Le terme P (G*i* | Z, θ) permet à chaque vecteur de génotypes observé de changer via les mutations des génotypes déterminés par la paire d'haplotypes copiés et est modulé avec le paramètre de mutation θ (49). IMPUTE v2 quant à lui, sépare d'abord les SNPs en deux ensembles : un ensemble T qui est génotypé dans l'échantillon et dans la cohorte de référence et un ensemble U qui est génotypé uniquement dans la cohorte de référence. L'algorithme estime les haplotypes au niveau des SNPs de T par IMPUTE v1, puis impute des allèles au niveau des SNPs de U en fonction des haplotypes actuels estimés.

1.3.3.2 Minimac 4

Minimac 4 est une implémentation à mémoire réduite et plus efficace en termes de temps de calcul que les algorithmes Minimac développés auparavant (7). L'algorithme est basé sur une réduction de l'espace d'état des HMM décrivant le partage d'haplotypes (62). Ce modèle divise le génome en blocs consécutifs et itère uniquement sur les haplotypes uniques dans chaque bloc génomique. Il utilise ensuite une fonction de *reversible mapping* qui peut reconstruire exactement

l'espace d'état utilisé par les méthodes plus anciennes Minimac (81) et IMPUTE (49). Ce logiciel d'imputation récent démontre une amélioration de la vitesse d'exécution, d'espace mémoire utilisé et d'exactitude d'imputation par rapport aux autres méthodes (7).

1.3.3.3 Serveur d'imputation de l'université du Michigan (*Michigan imputation server*, MIS)

Le serveur d'imputation de l'université du Michigan (MIS) est un service web d'imputation qui facilite l'accès à plusieurs panels de référence et facilite ce processus complexe, qui comprend plusieurs étapes de gestion du jeu de données, pour l'utilisateur. Le serveur d'imputation permet l'utilisation des panels de référence suivants : *Haplotype Reference Consortium* (HRC) (83), *Multi-ethnic human leukocyte antigen* (HLA) (84), Genome Asia Pilot & v2 (85), 1000 Genomes Phase 3 (1000G), CAAPA - African American Panel (86), HapMap2 (87) et le *Trans Omic for Precision Medecine* (TOPMed) (2). Il est aussi possible d'héberger les panels de référence propre à un usager sur ce serveur.

Le pipeline du MIS est rigoureux et utilise les logiciels Eagle2 pour le phasage des données et Minimac4 pour l'imputation afin d'offrir la meilleure qualité d'imputation disponible sur de gros jeux de données. Les données d'entrée de génotypage harmonisées téléversées par l'utilisateur sont, lors de la première étape, séparées en *segments* de 20 Mégabases. Les étapes suivantes seront effectuées individuellement sur chacun des *segments* d'ADN. La deuxième étape consiste en un contrôle de Qualité (QC) effectués à trois niveaux : les *segments*, les variants et les échantillons. Les étapes sont détaillées dans le Tableau 1. Si les données de génotypage entrées par l'utilisateur ne sont pas sur le même génome de référence que le panel utilisé, les données sont *converties* vers ce dernier. Le génomes de référence pour l'humain répertorient les nucléotides selon leur positions sur les chromosomes grâce au séquençage profond. Celui-ci sert de référence pour toute les données génétiques en uniformisant les positions et les identifiants des variants. La troisième étape du pipeline s'occupe du phasage des données génétiques avec comme référence la population Européenne avec un chevauchement de 5Mb fait par Eagle2 (77). La quatrième étape consiste ensuite à l'imputation des *segments* avec le logiciel Minimac4 sur des fenêtres de 500 000 paires de bases. La cinquième et dernière étape combine les différents *segments* imputés sur chaque

chromosome et encrypte les données afin de les rendre disponibles de manière sécuritaire à l'utilisateur.

Table 1.1. Étapes du contrôle de qualité du MIS.
(https://imputationserver.readthedocs.io/en/latest/pipeline/) (62)

Étapes	Filter
Chunk	Déterminer le nombre de variants valides: Un variant est valide s'il est inclus dans le panel de référence. Au moins trois variants doivent être inclus.
	Déterminer la quantité de variants présent dans le panel de référence : Au moins 50% des variants doivent être inclus.
	Déterminer le *call rate* : au moins 50 % des variants doivent être appelées pour chaque échantillon.
	Exclusion du *Chunk* si (#variants < 3 || chevauchement < 50% || *sampleCallRate* < 50%)
Variant	Vérifier les allèles : seuls A, C, G, T sont autorisés.
	Calculer la fréquence allélique alternative (AF) : Marquez tout avec un AF > 0,5.
	Calculer le *call rate*.
	Calculer le Chi-carré pour chaque variant (panel de référence vs. échantillons à l'étude).
	Déterminer les permutations d'allèles : Comparez référence (ref) et alternatif (alt) du panel de référence avec les données de l'étude (les variantes A/T et C/G sont ignorées).
	Déterminer les inversement de brin : après avoir éliminé les éventuelles permutations d'allèles, inverser et comparer ref/alt du panel de référence avec les données à l'étude.
	Déterminez les permutations d'allèles en combinaison avec les inversions de brin : combinez les deux règles ci-dessus.
Échantillon	Pour chr1-22, un *Chunk* est exclu si un échantillon a un call rate < 50 %. Seuls les *Chunks* complets sont exclus, pas les échantillons (voir "Au niveau du *chunk*" ci-dessus).

1.3.3.4 Mesure de qualité d'imputation

L'imputation des variants est par définition une inférence statistique du génotype pour un individu. L'exactitude d'imputation est calculée par le coefficient de corrélation r^2 entre les dosages d'allèles imputés et les génotypes masqués par Minimac (80, 81) ou par le score INFO de IMPUTE (49). Le dosage imputé pour un haplotype est estimé en trouvant la probabilité postérieure (suite à l'imputation) de l'allèle alternatif à un site. Le dosage génotypés est ensuite évalué en faisant la somme de chaque dosages haplotypiques (homozygote vs hétérozygote). Par exemple, si la probabilité postérieure de l'allèle alternatif au site X est de 0.98 et 0.96 pour chaque haplotype, le dosage génotypique de cet individu sera de 0.94 + 0.98 = 1.93 (c'est-à-dire très proche de 2 allèles alternatifs, avec une petite incertitude). Le r^2 est calculé selon la formule suivante :

$$\hat{r}^2 = \frac{\frac{1}{2n} \times \sum_{i=1}^{2n}(D_i - \hat{p})}{\hat{p} \times (1 - \hat{p})}$$

où $\hat{p}$ est la fréquence de l'allèle alternatif dans la référence, D_i la probabilité d'imputation de l'allèle alternatif au $i^{ème}$ haplotype et n le nombre d'échantillon GWAS. Le r^2 varie de 0 à 1, 0 étant considéré comme la moins bonne et 1 la meilleure qualité (81). Cette métrique dépend donc grandement de la différence de fréquence entre la fréquence imputée et celle dans le panel de référence, abaissant le score pour des variants ayant des fréquences divergentes entre les populations. De plus, le r^2 est une estimation de la qualité et non une mesure de l'exactitude d'imputation. Il est possible de mesurer la réelle exactitude d'imputation en comparant les variants imputés aux variants génotypés pour les mêmes individus (1), toutefois cette technique requiert les données de séquençage pour les mêmes individus, ce qui contredit l'objectif du génotypage et de l'imputation.

1.3.4 Panels de référence

L'imputation se base sur une cohorte de génomes de référence afin d'inférer les sites manquants suite au génotypage d'un individu. Sachant que l'imputation utilise aussi le LD pour inférer les SNPs manquants, il est donc important d'avoir de larges cohortes séquencées et variées (65, 88).

Les premières cohortes de références disponibles furent le *Human Genome Diversity Project* (HGDP) (89) et HapMap (40) composées respectivement de 1064 lignées cellulaires issues de 52 populations et de 270 individus issus de 3 populations (39). Ces deux cohortes contiennent principalement des individus d'origines européennes, africaines ou asiatiques. Depuis, plusieurs nouvelles cohortes de référence ont été créés afin d'être utilisées pour faire de l'imputation, elles se composent toutes principalement d'individus d'origine Européenne. Entre autres, le *Haplotype Reference Consortium* (HRC) (90) composé de 64,976 haplotypes provenant principalement d'individus d'origine Européenne et le UK10K avec près de 10,000 individus Britanniques (91). Afin de mieux représenter la diversité des haplotypes chez l'humain, les données de certains projets multi-ethniques sont aussi utilisées comme panel de référence pour l'imputation. C'est le cas du projet 1000G (92) composé de 2 504 individus d'origine diverses (Européennes, Africaines, Asiatiques, Américaines ou encore Afro-américaine) et du plus récent et plus complet, TOPMed qui rassemble plus de 80 différentes études pour un total d'environ 155,000 individus (93). Ce dernier est composé de 41% d'individus d'origine européenne, de 31% d'individus d'origine africaine, 15% d'individus d'origine latino/hispanique, 9% d'individus d'origine asiatique et de 4% d'individus d'origines autres (93).

En bref, les cohortes de référence utilisées pour l'imputation sont souvent basées sur des individus d'origine européenne et donc n'ont pas la capacité d'inférer les variants spécifiques aux populations d'origines différentes (3, 63). Toutefois les avancées technologiques permettent aujourd'hui d'inclure de plus en plus d'individus d'origines plus variées afin d'augmenter le pouvoir statistique d'inférence et la qualité générale d'imputation.

1.3.5 Difficultés et limites

1.3.5.1 Densité des variants

Bien que l'imputation soit une méthode utile, spécialement pour les cohortes d'individus non apparentés (63), certains facteurs influencent sa performance. La densité de variants dans les données de génotypage affecte grandement l'imputation. En effet, puisque l'imputation se base sur les positions génotypés et sur les tronçons d'haplotypes partagées, une couverture restreinte/concentrée de génotypage du génome ne permet pas d'inférer par imputation l'entièreté de la variation haplotypique existante (49, 63). Concernant les données utilisées pour l'imputation, il est essentiel de filtrer les données pour enlever les erreurs de génotypage probables afin de s'assurer d'une meilleure qualité d'imputation. Les filtres couramment utilisés sont effectués sur les variants ayant des taux de données manquantes faibles, un écart à l'équilibre de Hardy-Weinberg (HW) qui stipule que les fréquences des allèles et des génotypes demeurent constantes au fil des générations en l'absence de forces évolutives (94). Le pré-phasage des données de génotypage est aussi très important puisqu'il permet d'obtenir, pour tous les individus, les haplotypes estimés (81).

1.3.5.2 Diversité populationnelle

Un autre facteur qui affecte de manière significative la qualité de l'imputation est la diversité populationnelle présente dans les panels de référence utilisées pour l'imputation, principalement dans la mesure où la qualité de l'imputation du génotype peut être affectée par des panels de référence qui ne correspondent pas exactement à l'ascendance des populations cibles (95). L'imputation de SNPs rares est plus difficile car ceux-ci sont souvent spécifiques à une population et leurs associations reflètent une structure de LD affectée par des événements démographiques récents (96). De manière générale, on observe une qualité d'imputation plus faible pour les populations métissées qui ont des blocs de LD plus courts (97, 98) et une plus grande hétérogénéité dans leur structure d'haplotype. Celles-ci peuvent bénéficier de l'utilisation

de populations de référence plus diversifiées et/ou plus grandes (3). L'augmentation de la qualité par l'utilisation de panels de référence plus grands et diversifiés peut surpasser la qualité d'imputation obtenue en imputant avec un cohorte de même ascendance que la population métissée imputée (95). Les options pour imputer plus précisément les variants rares dans une population spécifique comprennent l'augmentation de la taille de la cohorte de référence d'imputation pour capturer plus de diversité d'haplotypes de référence et l'augmentation de la profondeur de séquençage dans le panel de référence afin de minimiser les taux d'erreur des génotypes (99). Un cas plus spécifique de population présentant des difficultés d'imputation est la Finlande (100). Étant une population fondatrice, elle présente des patrons de LD, des fréquences alléliques et des haplotypes différents des autres populations européennes. Ces différences influencent la qualité d'imputation des génotypes des Finlandais s'ils ne sont pas ou peu représentés dans le panel de référence utilisé pour l'imputation.

1.3.5.3 Uniformité des données

Les données génétiques utilisées lors de l'imputation sont souvent hétéroclites, que ce soit au niveau de l'ethnicité des individus comme mentionné dans la section précédente ou encore au niveau des méthodes utilisées pour obtenir les données de génotypage. Effectivement, il est possible que les données de génotypage proviennent de plusieurs puces de conception et de marques différentes (101, 102). De ce fait, les sites caractérisés diffèrent d'un jeu de donné à l'autre, le but de l'imputation étant justement d'homogénéiser l'ensemble de sites étudiés. Il est parfois possible d'observer des effets de lots (ou *batch effect*) dans les données, c'est-à-dire une distinction marquante entre les échantillons issus des différentes techniques/lots. Ces différences, si elles ne sont pas contrôlées et que les données ne sont pas uniformisées, peuvent se transmettre dans les données imputées et mener à de fausses conclusions lors d'études d'associations. Les solutions possibles à ce problème ne sont pas infaillibles mais consistent à procéder à un QC rigoureux et identique sur les différents lots de données. Les étapes de filtrage pour le taux de données manquantes, la fréquence de l'allèle mineur (ou MAF pour *minor allele frequency*) et l'équilibre de HW sont essentielles, ainsi que l'harmonisation des allèles afin d'obtenir des données uniformes et comparables (voir section 3.3.2.). Un moyen de vérifier que l'effet de lot n'est pas

introduit dans les données est en faisant des analyses en composantes principales (ou PCA pour *Principal Component Analyses*) (103) pré et post uniformisation. En analysant les composantes principales, si un effet de lot est présent, une distinction entre les différents lots sera détectable. Il est toutefois important de ne pas confondre structure de population et effet de lot lors des analyses en composantes principales (voir section 1.4.3).

1.4 Génétique des populations

La génétique des populations est un champ au carrefour de la biologie expérimentale, computationnelle et théorique qui étudie la composition génétique des populations et ses changements résultant de l'effet de divers facteurs tels que la sélection naturelle, la migration, la dérive ainsi que les changements démographiques (104). Elle consiste à utiliser divers modèles statistiques utilisant la dynamique des fréquences alléliques afin de mettre en lumière les processus évolutifs qui ont façonné le génome des populations humaines (104).

1.4.1 Dérive génétique

La dérive génétique est un processus qui affecte la fréquence allélique des variants dans une population d'une génération à l'autre. Cette variation dans la fréquence des allèles s'explique par une sélection aléatoire d'allèles de la population lors de la transmission à la prochaine génération. En effet, l'ensemble des gamètes qui contribueront à la génération suivante se fait en fonction des individus qui vont se reproduire et transmettre leur code génétique, on peut alors parler d' « erreur d'échantillonnage » (105) dans les populations finies. Par ce processus, certains allèles seront transmis alors que d'autres ne le seront pas et verront leur fréquence diminuer ou même disparaitre de la population. Le cas de figure inverse survient aussi, certains allèles verront leur fréquence augmenter, parfois même jusqu'à la fixation (fréquence de 100% dans la population). Cette fluctuation des fréquences est non-directionnelle et est grandement dépendante de la taille effective de la population (106). La taille effective d'une population représente le nombre d'individus qui participent à la formation de la génération suivante. Cette taille est

généralement plus petite que la taille totale réelle de la population. En conséquence, plus cette taille est petite, plus la dérive est forte puisqu'un événement aléatoire de changement de fréquence a un impact proportionnel plus grand sur le pool génétique (107). Sur plusieurs générations, la dérive diminue généralement la variabilité existante dans une population bien que certains variants rares voient leur fréquence culminer (108). La dérive contribue ainsi à la différentiation des populations.

1.4.2 Démographie

Les changements de taille d'une population ont une grande influence sur les fréquences alléliques des variations génétiques inter-individuelles. Deux principaux cas de figure qui impactent la démographie d'une population de manière opposée sont l'expansion et le goulot d'étranglement (109).

L'expansion populationnelle est un phénomène de croissance démographique considérable. Lorsqu'une population subit une croissance rapide, un excès d'allèles rares survient et une augmentation du nombre de sites polymorphes peut être observée (110, 111). Malgré l'augmentation du nombre de variants, leur effet délétère moyen est plus faible (110).

À l'inverse, un goulot d'étranglement se définit par une diminution drastique de la taille d'une population (112). Ce phénomène survient entre autres par une contrainte environnementale ou sociale telle qu'une famine, une catastrophe naturelle ou encore un évènement de colonisation. Suite à une baisse de la taille effective de la population, la diversité est diminuée de par l'échantillonnage restant de la population initiale et la dérive agit de manière plus importante. Certains allèles sont perdus alors que d'autres voient leur fréquence augmenter (106).

La migration influence aussi la démographie d'une population. Elle survient lorsqu'un groupe d'individus se déplace d'une population vers une autre. Le matériel génétique distinct des populations vient donc à se mélanger (113). Lorsque les fréquences alléliques sont différentes entre la population immigrante et la population hôte, les fréquences de cette dernière seront affectées et changeront (114). La migration permet de contrecarrer les effets de la dérive génétique et de limiter par le fait même la différenciation populationnelle en diminuant le taux de consanguinité (104).

Elle donne lieu dans certains cas à du métissage. Suite au mélange de populations de multiples origines avec différents variants génétiques, les populations métissées peuvent présenter des niveaux élevés de variation génétique, reflétant les contributions de leurs multiples groupes ancestraux.

1.4.3 Consanguinité

La consanguinité survient lorsque des individus apparentés se reproduisent entre eux. Les génomes des individus apparentés contiennent de multiples segments chromosomiques IBD, c'est-à-dire hérités d'un ancêtre commun. Leur progéniture a ainsi plus de chance d'hériter des mêmes allèles venant de ses deux parents. Ce phénomène entraîne une augmentation de l'homozygotie. Cette augmentation peut avoir des effets néfastes lorsque les variants transmis sont récessives délétères et qu'elles expriment un mauvais phénotype lorsque présente en deux copies (homozygote) (115). Au niveau populationnel, la consanguinité diminue la taille effective de la population et donc la variabilité génétique celle-ci (116). Dans de petites populations subissant une forte dérive génétique, la consanguinité peut augmenter la fréquence de mutations légèrement délétères jusqu'à la possible fixation (117). Cependant, la consanguinité peut également permettre la purge de mutations récessives sévèrement délétères par le même processus (118).

1.4.4 Sélection naturelle

La sélection naturelle fait varier les fréquences alléliques des mutations, soit en augmentant la probabilité de transmission, donc la fréquence, en la diminuant ou en la maintenant à un niveau intermédiaire en fonction de l'influence de la mutation sur le succès reproducteur (64). Les trois différents type de pressions sélectives sont la sélection positive, la sélection négative et la sélection balancée. La première augmente la fréquence allélique d'un allèle qui aurait un effet positif sur le succès reproducteur des individus porteurs en augmentant sa chance d'être transmis aux générations suivantes (119). La deuxième, à l'inverse diminue la fréquence d'un allèle délétère

dans la population par le phénomène opposé (120). La dernière maintient les fréquences alléliques à un niveau intermédiaire pour les allèles présentant un avantage de l'hétérozygote, ce qui conserve la diversité génétique (121). L'avantage hétérozygote survient lorsque le succès reproducteur est plus élevé pour un individu porteur des deux allèles plutôt que de deux copies identiques (122). Ainsi, la sélection naturelle, contrairement à la dérive génétique qui est non-directionnelle et qui ne tient donc pas en compte l'impact fonctionnel de la mutation, est une force directionnelle.

1.4.5 Structure populationnelle

L'ensemble des facteurs énumérés ci-haut ont un impact sur les fréquences alléliques dans une population et sur sa diversité intra-populationnelle. La reproduction non-aléatoire des individus d'une même population peut aussi causer une différence au niveau des fréquences alléliques et créer une structure populationnelle (123). Cette structure peut aussi être appelée stratification de la population et est parfois causée par un isolat géographique (124). Elle peut entrainer un biais dans les études d'associations (125) c'est pourquoi il est important de la prendre en compte. Une des principales méthodes utilisées afin de visualiser et de tenir compte de la structure populationnelle est l'analyse en composantes principales (PCA) (126). La PCA est une technique non-supervisée permettant d'obtenir des composante principales (PC) qui expliquent le plus de variabilité dans les données génétiques (126). Souvent, un nombre restreint de composantes principales est considéré pour représenter les données, c'est alors une stratégie de réduction de dimensionnalité. Les composantes principales expriment en ordre décroissant la variabilité présente dans les données, de sorte que PC2 explique moins de variabilité PC1, mais plus de variabilité que PC3. Ainsi, les composantes principales illustreront la structure génétique présente dans la population étudiée. Il est donc important d'intégrer l'information des PCA dans les analyses de données génétiques afin de prendre en compte ce biais que peut entrainer la structure populationnelle. Lors d'études GWAS, les composantes principales (PC) sont intégrées comme co-variables afin de tenir compte de la structure populationnelle dans l'analyse.

1.4.6 Effet fondateur

L'effet fondateur survient lorsqu'un petit groupe de migrants ou de survivants s'établissent dans une nouvelle zone géographique. De manière générale, une population fondatrice voit le jour lorsqu'une population d'origine subit un goulot d'étranglement (*bottleneck*) suivi d'une expansion démographique lors de l'installation de colons sur un nouveau territoire. Diverses forces régissent les changements génétiques découlant d'un effet fondateur, on peut parler de dérive génétique, de sélection naturelle, de consanguinité ou encore de changements démographiques (104, 127). Tous ces phénomènes entraînent un changement dans les fréquences alléliques des variants dans la nouvelle population en comparaison à sa population ancestrale (128) en plus de diminuer la diversité populationnelle par rapport à population dont ils sont issus en raison du sous-échantillonnage (129) . Les patrons de déséquilibre de liaison sont par le fait même affectés et le taux de recombinaison ancestral peut être biaisé (104).

Pendant des années, les études des populations fondatrices ont représenté le courant dominant de la cartographie génétique dans l'effort de cibler les variants génétiques causant des troubles mendéliens (130). L'homogénéité génétique de ces populations ainsi que les expositions environnementales relativement homogènes sont également considérées comme des avantages dans les études sur les loci de susceptibilité génétique qui sous-tendent des maladies complexes. Effectivement, il a été démontré que les variants rares sont non seulement associés aux maladies mendéliennes mais aussi aux maladies complexes (131). Dans les populations issues d'un événement fondateur ou d'un isolat génétique, une mutation initialement à basse fréquence (rare) peut gagner en fréquence pour devenir plus facilement cartographiable (12, 132, 133).

1.4.7 Population fondatrice canadienne-française

La population canadienne-française est l'un des plus récents exemples d'une population fondatrice. Cette population est actuellement composée de plus de 6 millions d'individus descendant d'environ 8500 colons français (ou fondateurs) s'étant installés sur le territoire de la Nouvelle-France entre 1608 et 1759 (date de la fin de l'immigration française) (134). Il est donc

possible d'affirmer que la population canadienne-française provient d'un petit échantillon peu représentatif de la population ancestrale française (14). Il est important de noter que depuis la fin de l'immigration française en Nouvelle-France, la populations a évoluée de manière presque génétiquement isolée (14). Au cours du 19e et 20e siècle, des migrants d'origines diverses (acadiens, britanniques, irlandais, américains, juifs) se sont mêlés à la population canadienne-française, en ayant cependant un impact génétique très limité (124, 135). Il est aussi recensé un métissage génétique avec les populations des premières nations, toutefois les impacts de ce mélange demeurent mal caractérisés, malgré que l'importance de ce métissage demeure assez faible (136).

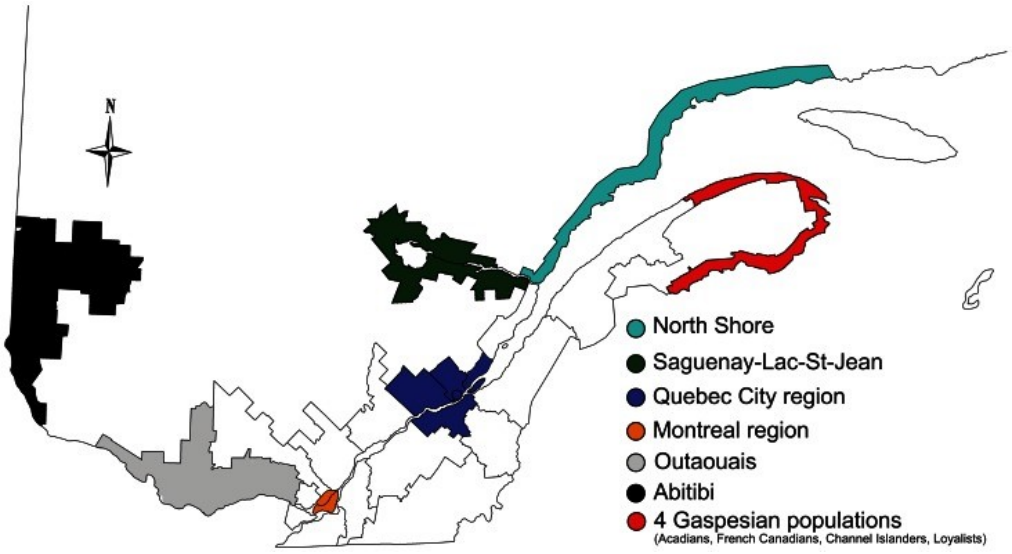


Figure 1.4. Carte des sous-populations canadienne-française du Québec (137).

Les premiers colons s'installèrent principalement sur les berges du fleuve Saint-Laurent, toutefois, la croissance démographique rapide de la population a conduit à la colonisation de nouvelles régions du Québec. Certains mouvements de colonisation se firent vers des régions éloignées et isolées, favorisant la stratification de la population (12). La colonisation a commencé en Gaspésie vers la fin du 18e siècle avec l'arrivée des Acadiens (descendants de pionniers français en Acadie) (138, 139). Ils furent rejoints ensuite par les loyalistes américains et par les Canadiens-français de la vallée du Saint-Laurent. Ces trois populations ethnoculturelles se sont reproduites

principalement entre elles. Par la suite, en 1840, un mouvement de migration vers la région du Saguenay fût entamé par les habitants de la région de Charlevoix, puis par les autres habitants de la vallée du Saint-Laurent. De 5 000 habitants en 1850, la population du Saguenay est aujourd'hui de 273 000 habitants en raison principalement d'un taux de natalité élevé. La région de la Côte-Nord fût elle aussi colonisée par des Canadiens-Français des régions de Charlevoix et du Bas-St-Laurent entre 1840 et 1920 (139). Deux grandes villes se développèrent rapidement dans la vallée du Saint-Laurent grâce aux descendants des premiers colons européens mais aussi des migrants des régions rurales qui se sont déplacés vers les zones urbaines dans le contexte des processus d'urbanisation et d'industrialisation (12). Chaque région a un effet fondateur distinct déterminé par le patrimoine génétique de la première génération de colons, qui dans de telles situations contribuent de manières disproportionnées à la structure de la population qui en résulte (12) (Figure 3).

Un avantage important de la population CF pour la recherche en génétique est la disponibilité de grands registres de population, comme le registre généalogique de population BALSAC (140) et le registre de la population du Québec (139). Un autre avantage de la population canadienne-française réside dans la corrélation des découvertes génétiques et de ces informations généalogiques pour chacune de ces sous-populations (12). Finalement, la population canadienne-française présente un excès de variants non-synonymes à faible fréquence, typiquement enrichis pour des variants fonctionnels, par rapport à sa population ancestrale, ce qui lui profère un attrait pour les études d'associations (141) (Figure 1.4).

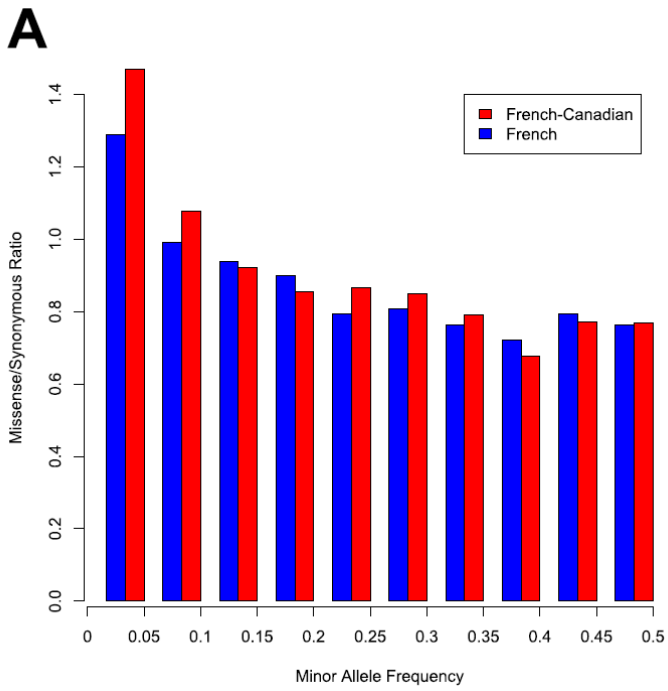


Figure 1.5. Excès de variants fonctionnels dans la population fondatrice canadienne-française. Figure tirée de (142). Ratio des changements non-synonymes et synonymes dans la population française et canadienne-française pour les variants regroupés par la fréquence de l'allèle mineur.

1.5 Jeux de données

Les données génomiques utilisées dans ce projet sont tirées de deux cohortes de la population québécoise, d'une cohorte de populations canadiennes ainsi que d'une cohorte de population mondiale. Dans cette section seront détaillés les jeux de données utilisés.

1.5.1 CARTaGENE

Le deuxième ensemble de données utilisé dans ce projet est la cohorte CARTaGENE (143). Il comprend à la fois des échantillons biologiques et des données sur la santé et le mode de vie

d'environ 43 000 Québécois âgés de 40 à 69 ans au moment du recrutement. Cette plate-forme publique de recherche du CHU Sainte-Justine vise à accélérer la recherche en santé. CaG est la plus grosse cohorte prospective de recherche en santé pour homme et femme au Québec (13). Sur ses 43 000 échantillons, 29 356 ont été génotypés à ce jour, desquels 937 ont été génotypés avec la puce Illumina Omni 2.5M, 988 sur la puce de la biobanque Affymetrix Axiom UK et les 27 429 autres sur le système de dépistage mondial Illumina Infinium avec quatre conceptions personnalisés détaillées dans les méthodes du chapitre 3 (143). Toutes ces puces de génotypage sont basées sur le génome de référence GRCh37 sauf pour la plus récente qui elle est en GRCh38. Plusieurs autres génotypes sont attendus d'ici la fin de l'année 2022. Les individus se répartissent entre six régions de la province : Montréal, Québec, Saguenay, Sherbrooke, Trois-Rivières et Gatineau (Outaouais). Les individus de cette cohorte ont été identifiée en fonction des villes dans lesquelles ils ont été recrutés, au lieu d'avoir une approche généalogique basée sur la génétique des populations ou la généalogie (origine des ancêtres).

1.5.2 Canadian Partnership for Tomorrow's Project (CanPath)

La cohorte *Canadian partnership for tomorrow's project* (CanPath, autrefois CPTP) est la plus grande étude sur la santé de la population au Canada (144). Cette initiative vise à étudier la biologie et les comportements des Canadiens afin d'en apprendre davantage sur les causes des maladies chroniques et du cancer. Ce projet d'envergure rassemblait en 2014, 300 000 participants inscrits. CanPath est divisé en plusieurs sous-cohortes régionales, elle est plus précisément composée de sept cohortes provinciales, soit : *BC Generations Project* (British Columbia, BCGP) (145), *Alberta's Tomorrow Project* (Alberta, ATP) (146), *Ontario Health Stud*y (Ontario, OHS) , *Atlantic Partnership for Tomorrow's Health* (Atlantic, ATL)(147), CARTaGENE (Québec, CaG), et les sous-cohorte non-publiées: *Healthy Future Sask* (Saskatchewan, HFS), *Manitoba Tomorrow's Project* (Manitoba, MTP) . Nous avons accès aux cinq premières sous-cohortes de ce projet, pour lesquelles nous disposons des données de génotypage d'environ 1 000 individus pour chacune d'entre elles qui ont toutes été obtenues à l'aide de la même puce de génotypage Affymetrix. Cette cohorte est utile car elle contient une sous-partie (990 individus) de CaG (13)

qui ont été génotypés par CanPath et renvoyés à CaG, nous permettant de comparer des caractéristiques non techniquement biaisées de cette population fondatrice.

1.5.3 Projet des 1000 Génomes (1000G)

Figure 1.6. Populations et sous-populations présentes dans le projet des 1000 Génomes (148). Les 26 sous-populations présentées dans la Figure 5 peuvent être rassemblés sous 5 populations Africaine (AFR, n=661), Américaine Centrale et du Sud (AMR, n=347), Asiatique de l'Est (EAS, n=504), Européenne (EUR, n=503) et Asiatique du Sud (SAS, n=489). Ce jeu de donné sert souvent de référence pour l'imputation, pour l'harmonisation des données génétiques ou encore pour l'inférence de la structure populationnelle dans un jeu de donné.

Le jeu de données public du projet des 1000 Génomes (*The 1000 Genomes Project*) (149) a pour but de permettre une meilleure compréhension de l'ensemble des variants génétiques présents dans le génome humain. La troisième phase de ce projet, composé du WGS de 2,504 individus issus de 26 populations différentes, contient un total de 88 millions de variants phasés en haplotypes de haute qualité. Un jeu de données plus récent du projet des 1000G est utilisé dans le Chapitre 3. Cette nouvelle version contient des données de séquençage à couverture élevée (30x)

alignées sur la référence du génome humain GRCh38. Il complémente la phase 3 du projet en y ajoutant le séquençage à haute couverture de 698 individus apparentés aux individus de la phase 3, ce qui permet d'avoir un total de 602 trios parents-enfant complets.

1.5.4 Panel de référence du Québec (QCRef)

Le panel de référence du Québec est un ensemble de données d'individus composé exclusivement de Canadiens-Français et se nomme le panel de référence du Québec (150). Chaque individu recruté dans cette cohorte devait répondre à un critère : être né dans la même région que ses quatre grands-parents. Il s'agit ici d'un moyen de s'assurer que le génotype reflète les particularités de la population, dans ce cas, dans les différentes régions de la province de Québec (3). Dans ce projet, 1058 individus ont été génotypés à quatre occasions différentes et sur quatre puces différentes. Le premier lot a été réalisé en 2008 et était composé de 118 personnes. Le deuxième lot génotypé en 2012 contenait 26 individus. Enfin, deux lots de 745 et 313 individus ont été générés respectivement sur la puce Neuro-X et une puce OMNI comprenant les individus des deux premières puces de génotypage. Le total de 1 058 échantillons a été réparti en 8 régions différentes (sous-populations) : Abitibi, Saguenay, Montréal, Québec, Côte-Nord, Beauce, Outaouais et Gaspésie. Comme la Gaspésie possède également une diversité d'origines au sein de sa population (voir section 1.4.5), les individus de cette région ont été plus précisément divisés en quatre sous-populations : les îles anglo-normandes, les Canadiens-Français, les Loyalistes et les Acadiens (138). Ces sous-régions présentent notamment des différences génétiques qui permettent de les séparer avec une PCA (Figure 6). Certains des individus ont été mis sur deux puces de génotypage différentes. Il est à noter que la puce de génotypage Neuro-X vise spécifiquement à caractériser les variants associés aux maladies neurodégénératives (151). Ceci signifie que les SNPs génotypés sont concentrés dans certaines régions spécifiques du génome qui ont été identifiés dans la littérature (GWAS, étude d'association) comme étant associés à ce type de maladies. Par le fait même, il est attendu que les performances d'imputation de ces individus soient plus faibles sur le génome entier (29, 35, 49).

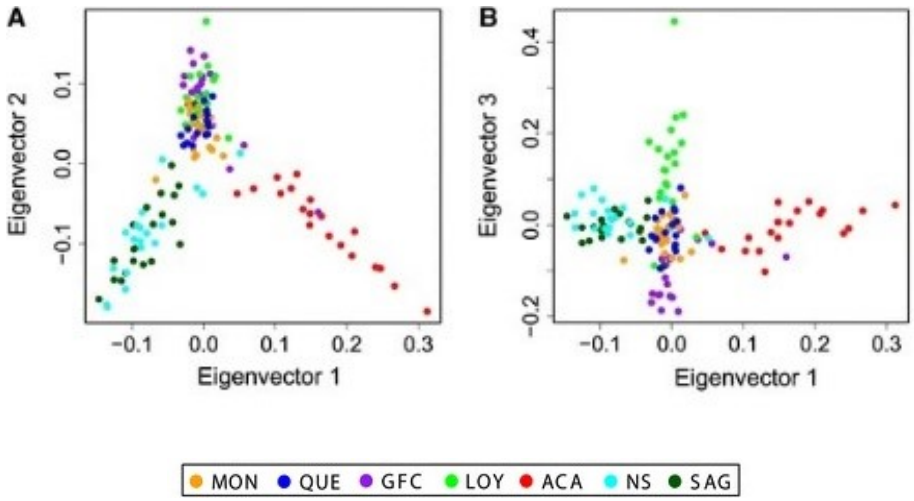


Figure 1.7. Structure de population du Québec capturée par les données génomiques. **A, B**. Graphique des trois premiers *eigenvectors* de l'analyse en composante principale (PCA) basée sur les données génomiques des sous-populations canadienne-française de la cohorte de référence du Québec (12, 141). Les sous-populations représentées par des couleurs sont Montréalais (MON), de la ville Québec (QUE), Gaspésiens canadiens-français (GFC), Loyalistes (LOY), Acadiens (ACA), de la Côte-nord (NS) et Saguenéens (SAG).

Chapitre 2 – Problématique (Hypothèses et Objectifs)

La composition génétique diffère entre les diverses populations humaines en fonction des contraintes évolutives, des événements démographiques et des mouvements migratoires. Plus spécifiquement, les populations fondatrices, bien qu'originaire de populations préexistantes, présentent une diversité génétique bien différente de celle de sa population mère. Ces différences confèrent un défi d'imputation en raison des différents patrons de diversité génétique et de déséquilibre de liaison entre les cohortes à imputer et les panels d'imputation qui sont principalement constituées d'individus d'origine européenne, n'étant pas représentatifs de la structure populationnelle mondiale. Malgré le défi que peut représenter son imputation génétique, la population Canadienne-Française demeure attrayante et utile à la découverte de variants associés aux traits et maladies complexes par les GWAS.

Des nouveaux panels d'imputation tels que TOPMed (93), avec plus de diversité génétique, prétendent pouvoir augmenter la quantité de sites imputables ainsi que d'en améliorer la qualité (63) pour la majorité des populations, même de celles qui ne sont pas inclues dans le panel de référence. Cette hypothèse, bien qu'intéressante, se doit d'être rigoureusement testée. Notre hypothèse est donc qu'un panel d'imputation plus diversifié et plus grand permettrait d'améliorer la qualité d'imputation des données génétiques Canadiennes-Françaises. Il se peut, cependant, qu'un panel de référence non-spécifique à la population fondatrice du Québec présente certaines lacunes qui doivent être caractérisées.

Afin de valider les hypothèses émises, nous comparerons l'imputation de CaG avec celle des autres populations canadiennes non-fondatrices de CanPath. Puis, nous comparerons la qualité d'imputation du jeu de donnée de CaG lorsque différents panels de référence sont utilisés. Nous développerons la stratégie optimale afin d'imputer un jeu de donné génotypé sur plusieurs architectures de puces. Finalement, il sera possible de vérifier l'exactitude de l'imputation en comparant les données génotypées puis imputés avec les données issues du WES de CaG disponibles (puisque les données de WGS ne seront disponibles qu'au courant de l'année 2022).

Malgré l'implémentation confirmée des pipelines d'imputation, aucun article scientifique n'adresse le sujet de la composition d'un jeu de données imputé constitué de différentes puces de génotypages. Deux stratégies peuvent être envisagées. La première consiste à imputer chacune des puces de données individuellement afin de garder le plus grand nombre de variants et donc d'information haplotypique pour l'imputation. L'imputation de ces données est ainsi faite séparément et ensuite les résultats imputés sont combinés. La deuxième stratégie effectue les étapes dans l'ordre inverse, c'est-à-dire qu'elle combine les différentes puces de génotypage en gardant l'intersection des variants de celles-ci avant de faire l'imputation du jeu de données en entier. Notre hypothèse est que l'imputation issue de la première méthode serait de plus haute qualité puisqu'elle conserve l'ensemble des informations mesurées par génotypage. Cependant, il se peut que cette technique amplifie des effets de lots. Nous tenterons de déterminer laquelle des deux stratégies d'imputation énumérées ci-haut permet d'obtenir la meilleure qualité d'imputation en effectuant l'imputation des données génétiques de CARTaGENE et de vérifier s'il y a amplification d'effets de lots.

Chapitre 3 – Article

Evaluation of genetic imputation in the French-Canadian founder population

By

Justin Pelletier[1,2,4], Jean-Christophe Grenier[1], Holly Trochet[1], Thibault de Malliard[3], Julie Hussin[1,2,4].

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

[2] Département de biochimie et médecine moléculaire, Université de Montréal, Montreal, Qc, Canada

[3] CHU Saint-Justine, Research Center, Montreal, Qc, Canada

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

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

Author contributions:

- Justin Pelletier conceived and performed all experiments on CaG datasets, interpreted the results and wrote the manuscript and figures;
- Jean-Christophe Grenier pre-processed genotyping data and post-processed imputed data from the QCRef, re-calculated R^2 scores with Beagle and assisted in the preparation of the manuscript and figures;
- Holly Trochet assisted in the post-processed imputed data;
- Thibault de Malliard pre-processed genetic data from CARTaGENE cohort.
- Julie Hussin supervised the project, assisted in the conception of analyses to test hypotheses and assisted in the preparation of the manuscript and figures.

3.1 Abstract

The French-Canadian (FC) founder population has been useful in genetic association studies in genetic diseases because of its unique diversity and excess of rare variants. Here, we investigate the performance of the imputation of a founder population not represented in the available reference panels. We characterize the imputation of 29,356 individuals from the province of Quebec, genotyped on six genotyping arrays, from the CARTaGENE cohort (CaG). We establish that the newest and more diverse reference panel Trans-Omics for Precision Medicine (TOPMed) outperformed the latest Haplotype Reference Consortium (HRC) reference panel in the CARTaGENE dataset. We assessed the precision of imputation with the widely used quality score (R^2) and the accuracy calculated based on sequenced variants available in CARTaGENE Whole Exome Sequencing data (WES). With more high-quality imputed variants and better imputation in the rare variants, imputation with TOPMed reaches higher imputation performance in the French-Canadians. However, we observed that the R^2 score tend to reflect minor allele frequency rather than imputation accuracy, which calls for new ways of assessing imputation quality in rare variants. The standard GWAS R^2 threshold at 0.3 excluded 19.3% of the exonic variants imputed in CaG, of which 98.5% achieved an accuracy of imputation over 98%, counting 75.1% SNPs reaching 100% of accuracy. We also compare and establish an optimized strategy to increase imputation quality on heterogenous datasets by merging every sub-dataset after imputing them

individually. These results highlight the impact of the lack of representation of specific founder populations' haplotype diversity in reference panels in human genetics today, even if they are of European descent, demonstrating the importance of including even more populations in these panels.

3.2 Introduction

Identifying and characterizing genetic variation that impacts human traits such as diseases, response to drugs and other phenotypic traits is the central objective in human genetics. In due time, this would be achieved by sequencing personal genomes, however this is not yet feasible due to genotyping costs, storage capacity and computing capacities (152). However, significant progress can be made by measuring only a modest number of genetic variants in each individual using genotyping arrays. The idea that data on a limited set of genetic variants can provide useful information about other genetic variants in the same genome forms the theorical basis of genetic linkage (153): these genotyped markers are used to identify haplotypes, the set of variations that tends to be inherited together (combination of alleles found on a same chromosome), using linkage disequilibrium (LD). Genotype imputation uses the information from these shared haplotypes to infer the genotypes at many variants that are not directly genotyped (88). This methodology imputes untyped variants in a genome using information on Linkage Disequilibrium (LD), the correlation between nearby variants such that the alleles at neighboring polymorphisms are associated within a population more often than if they were unlinked, from a haplotype reference panel as well as the observed genotyped variants from that specific genome. The current state-of-the-art reference panels are built from Next Generation Sequencing (NGS) data and contain detailed haplotype diversity from its sequenced individuals (154).

Most available reference panels for imputation are geared towards European ancestry, which usually leads to substantially poorer imputation quality for non-European individuals (1) because of haplotype diversity. Until 2021, the largest panel available was the Haplotype Reference Consortium (HRC) (90) with a total of 64,976 haplotypes available for genotype imputation. The individuals included in the HRC panel were predominantly from populations of European descent, except for the 661 individuals of African descent who were part of the

individuals sequenced in the 1000 Genomes Project (1000G) (155). Some smaller panels focused on African ancestry like the African Genome Resources (AGR) panel (156) and the Consortium on Asthma among African-ancestry Populations in the Americas (CAAPA) (157) but were not widely used. Since 2021, the Trans-Omics for Precision Medicine R2 (TOPMed) panel (2), a larger and more diverse reference panel for imputation, has been made available to the community. With 97,256 individuals from Asian, European American, Caribbean, South American, Central American, European, African, and Oceanian ancestries, TOPMed is the largest and most inclusive panel available to date (1). Used in modern imputation pipelines, it has been shown to provide significant improvement in the power and reach of genome-wide association studies (GWAS), even for rare variants (frequency down to 0.01%) (2). TOPMed also provide haplotypes from different founder population such as Finnish (158), Amish (n=255) (159) and Samoan (n=1,198) (160, 161).

In the populations arising from a genetic isolate and/or a founder event followed by a rapid increase in effective population size, rare variants from the mother population can be lost quickly or see their frequency increase (158, 162, 163). Since mapping rare variants is more challenging due to their implied rarity, founder populations are praised for the gain of frequency of these variants (133, 164). Multiple founder populations are born as a consequence of the colonization of the Americas by the Europeans. One of these founder population is the descending settlers of Nouvelle-France, currently forming most of the population of the province of Quebec in Canada. Historically, the Europeans founders of Nouvelle-France were mainly of French origin, who settled on the territory in the beginning of the 17th century (165). Because of limited number of European founders (14), the presumably low level of Native American admixture (137) and demographic expansions in the 19th century (165, 166), a number of endemic recessive diseases are seen at higher prevalence in the province, due to the increased frequency of specific mutations (14). Therefore, despite being a founder population of European descent, genetic diversity in French-Canadian (FC) does not entirely reflects its ancestral population's diversity (France) (142). In fact, the haplotypes present in the French-Canadians population differ from their ancestral population (12, 142, 162, 164, 167). Furthermore, the French-Canadian founder population is not genetically homogeneous due to unequal genetic contribution of the founders in different regions of Nouvelle France, and to serial demographic expansions (12, 167). These regional populations, as well as the French Canadian population as a whole, has shown to be of interest for discovery of

causal rare variants for single gene Mendelian disorders (168, 169) and for complex traits and diseases in GWAS through genotype imputation (155).

Although some founder population's haplotypes can be found in the current reference panels, the French Canadians are not well represented. It is unclear whether the presence of few French Canadians haplotypes is enough to significantly represent the haplotype diversity and to generate imputation accuracy across the frequency spectrum that is comparable to non-founder European populations (1, 12, 142, 162, 164, 167).

The aim of this study is to evaluate the improvement in imputation quality achieved using the larger and more diverse reference panel TOPMed provides in the French-Canadian founder population using the CARTaGENE (CaG) cohort (13), a population-based biobank of Quebec. CaG is the largest ongoing health study in Quebec with close to 43 000 participants aged between 40 and 69 years of age recruited between 2009 and 2013, of which 29 356 have been genotyped on six different genotyping arrays. The multiple genotyping arrays, each including various numbers of individuals and different sets of genotyped markers, complicated the imputation process of such a cohort. We thus compared two imputation strategies, being wary of minimizing potential batch effects: (1) imputation of each genotyping array separately, followed by merging of the imputed datasets into a unified cohort; (2) merging of the genotyping arrays into a pre-imputation unified dataset with shared variants, followed by imputation. We also aim at comparing the imputation quality of the French Canadians with non-founder Canadian populations of European descent from other provinces present in the CanPath consortium (144). We also took advantage of sequenced individuals in CARTaGENE to validate the imputation results by directly comparing imputed to sequenced variants using whole exome sequencing data from CaG.

3.3 Methods

3.3.1 Genetic datasets

Genotyping data of 27,429 individuals from CARTaGENE cohort (143) were used to study the performance of genomic imputation in French-Canadian population of Quebec. Genotyping

data were obtained using six different genotyping arrays described in Table 3.1. of which, the GSA_17K was lifted over from GRCh38 to GRCh37 using CrossMap lift tool (170). Seven individuals were removed from the smallest of the two genotyping arrays they were genotyped on (Supp. Tab. 3.1). In addition, Whole Exome Sequencing (WES) data for 90 genotyped French-Canadian individuals is used to assess imputation accuracy using the TOPMed imputation reference panel (62). The Canadian Partnership for Tomorrow's Health (CanPath) cohort (144) provides a comparative cohort including Canadian populations of European descent without founder effects. CanPath includes individuals from five regional cohorts: the BC Generation Project (BCGP), the Alberta's Tomorrow Project (ATP), the Ontario Health Study (OHS), the Atlantic PATH (ATL) and CARTaGENE cohort (CaG). CanPath genotyping data were generated using the Affymetrix Axiom UK Biobank genotyping array and the genotyped sub-cohorts contain respectively 979, 967, 952, 937 and 988 genotyped samples. The Quebec Reference Panel (QCRef) (12) includes genotyped individuals that represent the genetic diversity of French-Canadians of Quebec. A total of 1,022 samples are divided into 8 different regions: Abitibi, Saguenay, Montreal, Quebec, Côte-Nord, Beauce, Outaouais and Gaspésie, with the latter regional population divided into four subpopulations: Channel Islands, French-Canadians, Loyalists, and Acadians. This cohort was genotyped on two genotyping arrays, a custom Illumina's Omni, and the Illumina NeuroX array (171), the latter targeting variants associated with neurodegenerative diseases. The high coverage whole-genome sequencing data from the 1,000 Genomes project (1000G) Phase III dataset (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/) (172) was used as a reference for identifying individuals of European descent in the cohorts mentioned above. This dataset has genomic variants of 2,504 individuals across five ancestral populations: Africans (AFR, n = 661), Europeans (EUR, n = 503), East Asians (EAS, n = 504), South Asians (SAS, n = 489), and Americans (AMR, n = 347) (44).

Table 3.1. Description of CARTaGENE (CaG) genomic data by genotyping array, reported number of markers (SNPs+indels), number of samples, number of samples included in Whole Exome Sequencing (WES) dataset and genomic assembly build.

Genotyping array	Genotyped markers	Genotyped samples	WES samples	Reference Genome
Affymetrix Axiom UK Biobank	773 164	990	-	GRCh37
Illumina Omni 2.5M	2 381 000	937	-	GRCh37
Custom Illumina Infinium Global Screening Array (GSA_760)	626 377	726	1	GRCh37
Custom Illumina Infinium Global Screening Array (GSA_4224)	728 919	4 180	61	GRCh37
Custom Illumina Infinium Global Screening Array (GSA_5300)	658 296	5 237	16	GRCh37
Custom Illumina Infinium Global Screening Array (GSA_17K)	713 087	17 286	12	GRCh38

3.3.2 Pre-processing of genotyping data

The genetic datasets from a wide variety of genotyping arrays were processed through quality-control (QC) independently using the same filters. A pre-imputation step was conducted on all datasets keeping only genotypes passing MAF of 1%, 1 % of missing data per site and Hardy-Weinberg-Equilibrium (HWE) p-value > 1e-5 using VCFTools (173). Harmonization to the GRCh37 reference genome was done using Genotype Harmonizer and BCFTools with the fixref plugin (-m flip option) (174). Sex chromosomes were removed from all analyses since the imputation quality of these chromosomes is beyond the scope of this study (175). All individuals were kept in from the datasets since TOPMed showed an increased imputation quality for samples from every ethnic background.

3.3.3 Comparison between French-Canadians and other Canadians of European descent

The imputation of CanPath's sub-cohorts was done on individually pre-processed genotyping data with the TIS using TOPMed reference panel. The number of high-quality sites (R^2>=0.8) in all sub-cohort except for one where the same site had a R^2 value under the good quality threshold (R^2<0.3) was counted and represented the variants that are challenging to impute for each Canadian sub-population.

3.3.4 Strategies of imputation for multiple genotyping array datasets

Imputation strategies

CaG imputation with TOPMed reference panel was performed using multiple genotyping arrays (*Tab. 3.1*.). We defined two possible strategies to impute such heterogeneously genotyped datasets using TOPMed reference panel, the largest panel available (2). TOPMed Imputation Server (TIS) first performs a pre-phasing step using Eagle v2.4 (77) and the imputation using Minimac4 (80). In the first strategy, Impute-Merge, QC, and imputation was done on every genotyping array individually (6 batches). The imputed data of every genotyping array was then merged. In the second strategy, Merge-Impute, all genotyping arrays were merged using BCFTools merge, keeping only sites that were shared between every imputed genotyping array using BCFTools merge, leaving 401,097 variants post QC. The QC was done on the merged dataset according to the filters described in section 3.3.2. The samples of the merged and QC-processed CaG dataset were randomly splitted in two somewhat balanced batches of 15,000 and 14,369 samples (2 batches) because of the TIS sample size limits of 25,000 samples per imputation job. The two imputed batches were then merged back together using BCFTools merge.

Post-imputation quality is assessed by the estimated Minimac R^2 score outputted by the TIS that ranges from 0 to 1, 0 being the worse and 1 the best quality. Every imputation output an R^2 score for each imputed variant. As in all cases, the data was submitted to TIS in batches,

multiple R^2 values were obtained for the same variants. When merging batches or when sub-selecting sets of individuals in our analyses, we recalculated the R^2 score using Beagle v5.1 algorithm (176), an alternative to Minimac4. Beagle v5.1 bases its R^2 score calculation on genotyping dosages outputted by Minimac4 in imputed VCF files.

Accuracy of imputation using Whole Exome Sequencing data

We used 90 samples in CaG on which WES and genotyping was performed (*Tab. 3.1.*) to develop an alternative measure of imputation quality, which we call sequenced-based accuracy according to the following formula:

$$Accuracy = 100 \times \left(\frac{Number\ of\ correctly\ imputed\ samples}{Total\ number\ of\ imputed\ samples}\right)$$

This score is reported in percentage and is calculated for every single nucleotide polymorphism (SNP) of the exome that was imputed and sequenced. The WES data that we have access to is from a published study (142) and includes individuals from confirmed FC origins. The accuracy score was calculated on CaG Impute-Merge and Merge-Impute strategies on the 633,283 imputed SNPs that were previously called in WES.

Principal Component Analysis (PCA) on imputed and genotyped data

CaG datasets were evaluated on QCRef and 1000G, two population datasets, for QC, population stratification and batch effect evaluation purposes. The two genotyping arrays of QCRef were imputed using the TIS with the same filters as applied to CaG and CanPath. The imputed data of both imputation strategies of CaG (section 3.2.2) of QCRef and of CanPath were post-processed to keep high R^2 variants (≥0.8) in each imputed dataset. The shared variants left after the post-processing were kept and extracted from 1000G project WGS data and the resulting dataset was filtered for MAF of 1%, 1 % of missing data per site and Hardy-Weinberg-Equilibrium (HWE) p-value > 1e-6 using VCFTools (173). We used the --indep-pairwise function in Plink v1.9b 5.2 (177) to prune the SNPs using the suggested parameters of a 1000 kilobase window, a step size of 50 variants, and an r^2 of 0.05 (48,745 SNPs lefts).

We used FlashPCA2 (178) to calculate the first 10 PCs on the 1000G project individuals and projected CaG Impute-Merge, CaG Merge-Impute and CanPath datasets onto the 1000G 10

PCs. We extracted the samples that clustered within the bounds established by the Europeans in 1000G PCA: of -0.09424291 ≤ PC1 ≤ -0.02937275 and 0.1618124 ≤ PC2 ≤ 0.2226286 (Supp. Fig. 3.2.). The 20,589 clustering with 1000G individuals of European descent identified in CaG were kept for the rest of the analysis. We also performed a PCA on the two imputed genotyping arrays of QCRef dataset using FlashPCA2 (Supp. Fig. 3.3.). The identified individuals of European descent in CaG Impute-Merge and CaG Merge-Impute were projected onto the QCRef PCs ($QCRef_{imp}$ PCs). To get comparable principal component for both datasets. A final PCA was performed with the QCRef individuals from initial pre-processed genotyped variants shared between QCRef and the genotyped variants shared by all genotyping arrays of CaG. CaG genotyped data was projected onto the QCRef genotyped PCs ($QCRef_{geno}$ PCs).

Detection of batch effect in CARTaGENE imputations

Detection of possible batch effects, introduced by either the Impute-Merge or Merge-Impute strategy for CaG, were investigated using the distribution of Principal Component (PC) values from the PCA of QCRef in sets from both strategies. Kolmogorov-Smirnov test (KS test) were performed on the projected values for CaG individuals on the 10 $QCRef_{imp}$ PCs for pairwise combinations of imputed genotyping arrays in both CaG Impute-Merge and CaG Merge-Impute separately to investigate differences in distributions that could reflect batch effects. D statistics from the KS test from pair of arrays for the Impute-Merge dataset were compared to the D statistic for the same pair of arrays in the Merge-Impute CaG dataset.

To compare batch effects across arrays between genotyped and imputed data, we computed Pearson correlation coefficients between the first 10 genotyped $QCRef_{geno}$ and imputed $QCRef_{imp}$ PCs. KS tests were performed on the projected values for the 3 $QCRef_{geno}$ PCs for pairwise combinations of genotyping arrays in genotyped CaG data. D statistics from the KS test from pair of arrays for the genotyping dataset was compared to the D statistic for the same pair of arrays in the Impute-Merge CaG dataset. The same comparison was done for the Merge-Impute CaG dataset.

3.3.5 Statistical Analyses, Code and Source Data

Every analysis is summarized in Supp. Tab. 3.2. Except when stated, all statistical analyses and figure generation were performed in R (v.3.6.3). Data represented as a graph in main figures, as well as the code to reproduce figures and analyses can be found here: https://github.com/JustinPelletier/FC-imputation. The CARTaGENE biobank was accessed through data access approval under the project number #406713. Information to apply for data access can be found here: https://www.cartagene.qc.ca/en/researchers/access-request. CanPath Biobank data was accessed through data access approval under project number DAO-240237 is available here https://canpath.ca/access-process/. The Quebec Reference Panel was accessed via a collaboration with the authors of this paper: doi: 10.1007/s00439-010-0945-x (12). The 1000 Genomes Project is publicly available here https://www.internationalgenome.org/data.

3.4 Results

3.4.1 Imputation strategies for datasets with multiple genotyping arrays

No clear imputation strategy exists for datasets genotyped using multiples genotyping array designs. We here implemented two distinct imputation strategies applied to the imputation of CARTaGENE cohort using TOPMed reference panel (2). The first strategy imputes every pre-processed genotyping array individually and merges the results of the imputed data to retain the most genetic information out of every sub-dataset (between 450,975 and 490,084 variants, see *Methods*). The second strategy consists in merging all genotyping arrays before imputation, keeping only shared genotyped variants (401,097 SNPs), pre-processing the dataset, and imputing everything together. Because of the limitation of 25,000 samples per imputation job on the TIS, the merged dataset is split in two balanced random batches of 15,000 samples, imputed and merged once again (*Methods*). The MIS and TIS use the R^2 score calculated by Minimac4 to assess the quality of imputation per site for each imputed batch of samples. When merging datasets post-imputation, R^2 scores were recomputed using Beagle (*Methods*). We also assessed imputation quality using the accuracy percent (Methods section 3.3.3.) based on WES data from CaG. We

defined that a variant had a high accuracy of imputation when the variant was correctly imputed in all individual or only incorrectly imputed in 1 individual (accuracy over 98%).

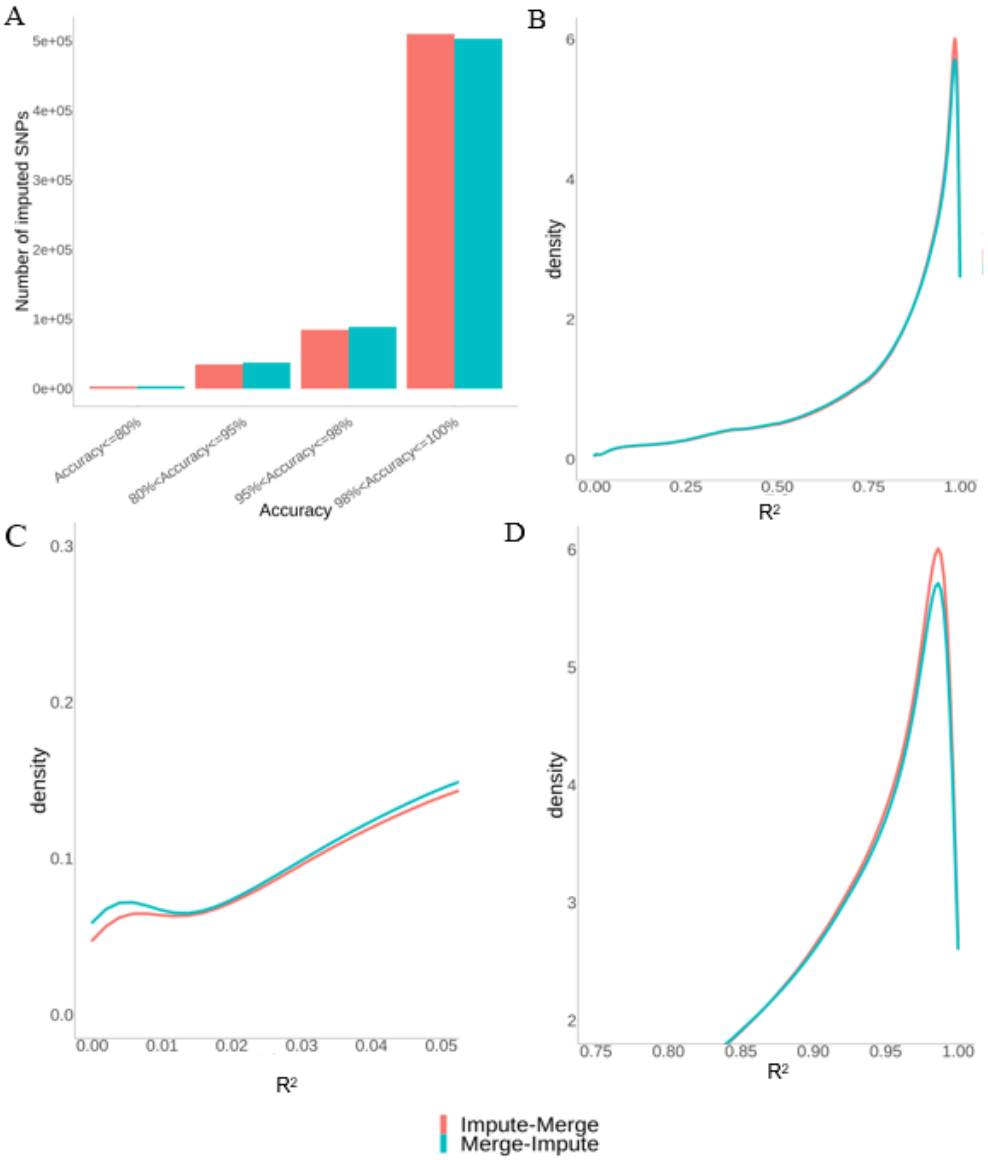


Figure. 3.1. Distribution of TOPMed imputation quality in CaG cohort using R^2 scores and accuracy derived from WES data for the Impute-Merge and Merge-Impute strategies for Variants with a MAF over 0. **A.** The accuracy, measured in percentage, is computed on 90 individuals and 633,283 imputed SNPs overlapping the WES data. **B.** R^2 scores on all markers genome-wide are recomputed from the per-array R^2 values (from Minimac4) on the full CaG cohort using Beagle. **C.** Zoom in the lowest part of the R^2 score density plot shown in panel B. **D.** Zoom in the highest part of the R^2 score density plot shown in panel B.

The number of variants that have a high accuracy is higher in the results from the Impute-Merge method with 510,527 high-accuracy exonic SNPs compared with 503,314 for the Merge-Impute (Fig. 3.1. A.). In fact, the Merge-Impute strategy gets an accuracy over 98% for 79.48% of its imputed SNPs while the Impute-Merge methods for 80.62%. The same phenomenon is observed with R^2 score values, with a higher density of high R^2 scores for markers (SNPs and indels) imputed with the Impute-Merge strategy over the Merge-Impute (Fig. 3.1. C.). Low quality markers ($R^2<0.3$) are observed in a higher density with the Merge-Impute method. Thus, both quality measures suggest a better imputation quality when utilizing the most information possible of every genotyping array for imputation with TOPMed.

Imputation is known to be more challenging for rare variants because of their implied rarity in the reference panels (179, 180). To see throughout the frequency spectrum where the gain in quality was attained with the Impute-Merge method, we performed Pearson's Chi-squared tests on the different classes of variants, comparing the numbers of high-quality vs low-quality markers in each MAF categories between the two strategies. We see significantly more exonic SNPs in high-accuracy categories for every MAF category with p-values ranging from p=0.03819 to p< 2.2e-16 (Table 3.2) in Impute-Merge strategy. The rare SNPs category (MAF<0.01), spanning the biggest part of the imputed data with 42% of the total set of exonic SNPs, shows the most subtle difference between the two strategies. Similarly, we performed Pearson's Chi-squared tests for R^2 scores in genome-wide imputed markers with a MAF over 0 (62,672,708 SNPs and indels investigated), on three categories: $R^2 < 0.3$; $0.3 <= R^2 < 0.8$; $R^2 > 0.8$, for low quality, good quality and high-quality markers, respectively. In every MAF category, the Impute-Merge gets significantly more high-quality markers (p< 2.2e-16, Table 3.2). Therefore, the gain in imputation quality is significantly higher in rare, low frequency and common markers when imputing before merging. This result could be explained by the retained genotype information from every genotyping array instead of keeping only the shared genotyped sites.

Table 3.2. Comparison of imputation quality between the Impute-Merge and Merge-Impute strategies**.** Pearson's Chi-squared test p-values are computed for accuracy percentage from WES data and R^2 scores, stratified by MAF for MAF>0. **A**. Number of SNPs at an accuracy over and under 98% for the 2x2 contingency table and Chi-square p-value for SNPs shared between the imputation datasets and WES data (633,283 SNPs), stratified by MAF. **B**. Number of markers (SNPs and indels) at different R^2 score thresholds for the 2x3 contingency table and Chi-square p-value for all markers shared between the two imputation strategies, stratified by MAF.

A					
Groups	**Number of SNPs with accuracy over 98%**				**P-value**
	Merge-Impute		**Impute-Merge**		
	Accuracy≤98%	**Accuracy>98%**	**Accuracy≤98%**	**Accuracy>98%**	
MAF < 0.01	4,227	261,762	4,040	261,949	0.0382
0.01 <= MAF < 0.05	18,310	109,194	16,993	110,511	4.30E-14
0.05 <= MAF < 0.1	17,025	36,765	15,706	38,084	< 2.2E-16
0.1 <= MAF < 0.25	38,260	49,641	35,836	52,065	< 2.2E-16
0.25 <= MAF < 0.5	52,147	45,952	50,181	47,918	< 2.2E-16

B							
Groups	**Number of SNPs in R2 groups**						**P-value**
	Merge-Impute			**Impute-Merge**			
	R^2<0.3	**0.3≤R^2<0.8**	**0.8≤R^2**	**R^2<0.3**	**0.3≤R^2<0.8**	**0.8≤R^2**	
MAF < 0.01	4,033,533	21,714,003	27,618,100	3,941,464	21,268,559	43,839,412	< 2.2E-16
0.01 <= MAF < 0.05	115,948	244,357	2,583,992	106,812	229,086	2,605,891	< 2.2E-16
0.05 <= MAF < 0.1	7,072	69,675	1,232,628	4,934	59,617	1,243,241	< 2.2E-16
0.1 <= MAF < 0.25	5,319	60,436	2,236,169	3,866	47,321	2,251,117	< 2.2E-16
0.25 <= MAF < 0.5	6,283	37,652	2,647,032	3,999	32,378	2,653,290	< 2.2E-16

The distribution of imputation quality is similar in exonic the SNPs compared to genome-wide variants (Fig. 3.1.A-B). For the accuracy measure in the exome, the highest density is observed for high-accuracy SNPs, with less than half of SNPs having an accuracy below 98% (49.69%). When keeping only R^2 scores in the set of the 633,283 exonic SNPs, most variants get a high R^2 score with fewer low R^2 variants (Fig. 3.2.), similar to the genome wide R^2 score distribution (Fig. 3.1.B). When stratifying by MAF calculated in the WES samples (Fig. 3.2), the rare variants (MAF<0.01) exhibit a higher density of low R^2 SNPs, with only few low R^2 variants present in low frequency and common variants. This result confirms that rare variants systematically get lower R^2 values, in line with what has been reported in the literature. The average R^2 in the exonic SNPs (mean R^2 =0.7783) is higher compared to the genome-wide variants

(mean R^2 =0.7659). This result can be explained by the fact that the proportion of rare, imputed SNPs in the exome kept in the accuracy analysis is only 42% while the proportion of rare markers in the variants kept for Fig. 3.1 is 82.25%. This difference is expected since the sample size differs greatly in the two analyses with only 90 samples for the WES, meaning that there are statistically more chances of characterizing rare variants in a larger dataset.

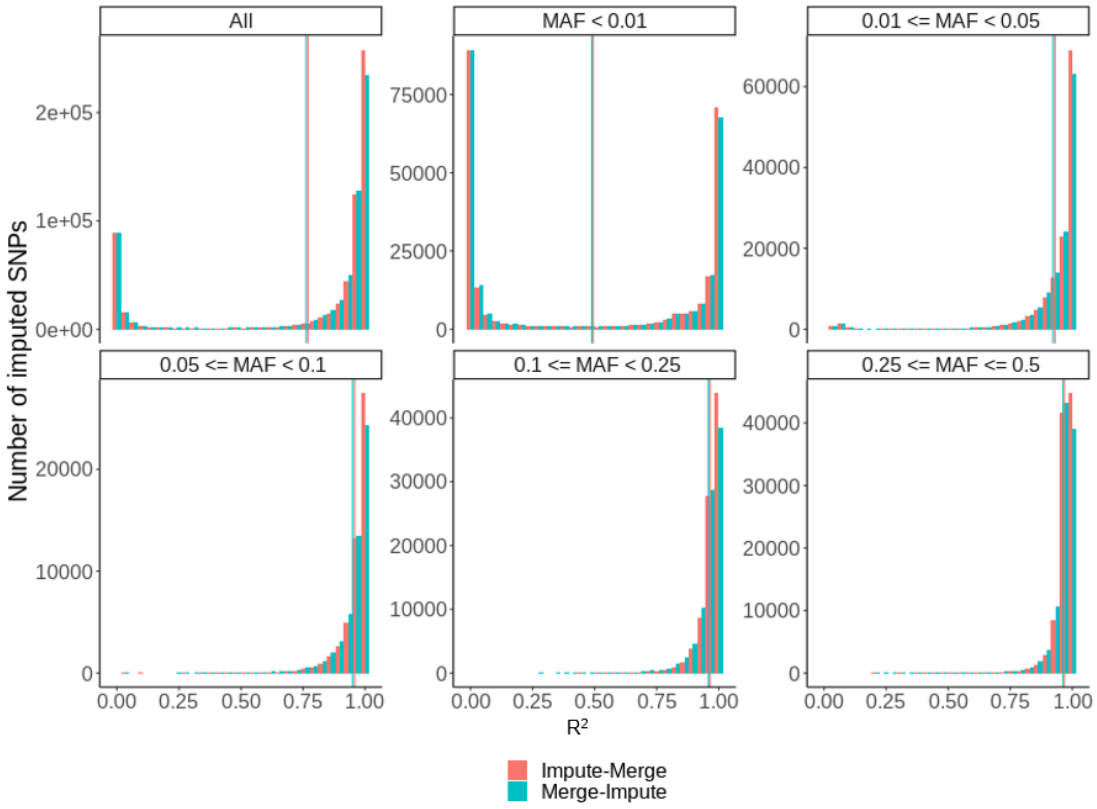


Figure 3.2. Distribution of the R2 scores in CaG TOPMed imputation for the Impute-Merge and Merge-Impute strategies, for 633,283 SNPs shared with the WES dataset, stratified by MAF computed in the WES. Vertical lines represent the mean R2 score per imputation strategy for each MAF category. The actual mean values for R2 score for the two methods are shown in Supp. Tab. 3.3.

Altogether, theses results show that regardless of MAF and exonic/genomic subsets, more high-quality markers are obtained when using the Impute-Merge method (Supp. Tab. 3.3.), supporting the use of the Impute-Merge strategy to impute heterogenous datasets such as CARTaGENE.

3.4.2 Detection of batch effect in CaG imputed data

Heterogenous genotyped datasets are prone to batch effects. CaG is genotyped on six different genotyping arrays, increasing the chances of seeing structure in the genetic data due to technical differences instead of real population structure. We investigated the possible batch effects in the Europeans in CaG data by comparing the two imputation strategies on TOPMed by using Principal Component Analysis (PCA) and investigating differences in PC values for individuals from different genotyping arrays with Kolmogorov-Smirnov tests (KS tests). To do so, CaG individuals were projected onto the Quebec Reference Panel (QCRef) PCA (*Methods*). The distribution of KS D statistics in PCs assessed between pairs of genotyping arrays in the same imputation strategy exhibits differences in distributions for both methods. However, these D values in the 10 first PCs are highly correlated between the two strategies, only showing a slight deviation towards higher difference between genotyping arrays in the Merge-Impute strategy (*Fig. 3.3. **A***.). The patterns of D-values are similar in both methods (Supp. Fig. 3.4.), demonstrating either a replicated batch effect by both imputation strategies or population structure that differs between the samples included in genotyping arrays, especially in PC1 and PC2. To test the latter, we compared the KS test D statistics between arrays for the genotyped data with D statistics obtained for the two imputation strategies for the first 3 PCs since only PC1, PC2 and PC3 of imputed and genotyped PCAs were highly correlated ($r \geq 0.95$) (Supp. Tab. 3.4.). These PCs were selected because they are highly correlated between genotypes and imputed data ($r > 0.95$). The analysis confirms that higher D values between arrays (hence array effects) in PC1 and PC2 were already present in genotyping data. Slightly higher D statistics are observed for the imputed data compared to genotyping data, regardless of the imputation strategy, suggesting that imputation minimally increases the differences between genotyping arrays. The absence of systematic bias observed between the imputation strategies suggest once again that neither strategy amplified batch effects beyond the initial genotyping array biases. Since it shows better imputation quality and no amplification of batch effects, the imputed data with the Impute-Merge strategy only is considered in further analysis.

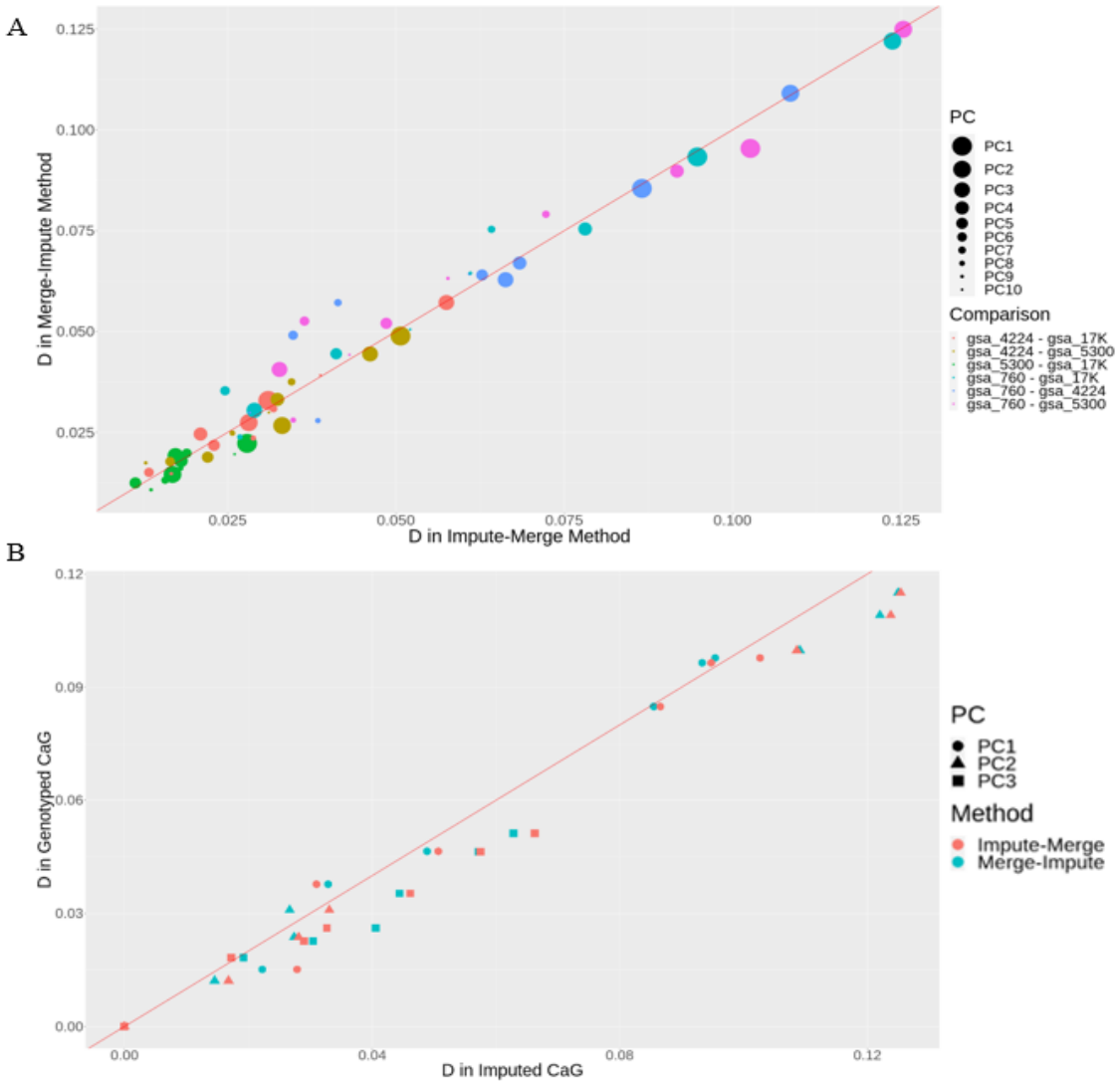


Figure 3.3. Evaluation of differences between genotyping arrays on principal components (PC) to evaluate the presence of batch effects due to the imputation process. PCs are computed on the QCRef Panel (Supp. Fig. 3.2). CaG genotyped and TOPMed imputed data are projected onto the PCs to obtain values for each CaG individuals. D statistics from Kolmogorov Smirnoff test are computed between PC values distributions for CaG individuals for pairs of genotyping arrays (see Methods). The red diagonal serves as a reference for perfect equality. A. Comparison of D values between arrays for the 10 first PCs for imputed data resulting from the Impute-Merge (x-axis) and Merge-Impute strategies (y-axis). B. Comparison of D values between arrays for the 3 first PCs for imputed data (x-axis) and genotyped data (y-axis), for both imputation strategies. The large D values seen between arrays on PC1-2 in the imputation datasets (A) are also observed in the genotyping data (B).

3.4.3 Comparison of R^2 scores and sequencing-based accuracy

We next compared the R^2 scores of the Impute-Merge imputed dataset to imputation accuracy computed based on WES data from CaG (*Methods*), stratified by MAF for exonic SNPs. The overall exonic distribution of R^2 scores exhibits lower quality of imputation compared to the calculated accuracy (Fig. 3.4.). In fact, the mean accuracy percent is 98.74% for all SNPs shared in the imputation and WES data, whereas the average R^2 for the same set of imputed SNPs is 0.7694. Although the metrics are not directly comparable, the mean accuracy is above the 98% threshold, whereas the mean R^2 score is below 0.8, generally considered as the high-quality imputation threshold. Furthermore, the number of high-accuracy SNPs is significantly higher than the number of high R^2 variants for every MAF group, although the difference in number is reduced as the MAF rises. Rare SNPs have the biggest proportion of high accuracy variants based on sequenced data, with 75.1% reaching an accuracy of 100%, when they have the lowest R^2 score of all MAF groups. This result suggests that rare variants are in reality well imputed (based on "ground truth" WES data) even if most of them would be excluded using the minimum R^2 threshold of 0.3. Many common variants, on the contrary, show high R^2 scores while being of lower accuracy (<98%, meaning that it is wrongly imputed in one or less individual). This means that a considerable fraction of genotypes at common variants sites are wrongly imputed in some individuals while being classified as high-quality variants based on the R^2 threshold of 0.8. The calculated accuracy score repudiates the tendency of the R^2 metric in classifying rare variants as badly imputed and common one as well imputed, which suggest the R^2 values may be driven by MAF than by actual imputation quality, which needs to be taken into account by the community when processing post-imputation datasets with state-of-the art methodologies.

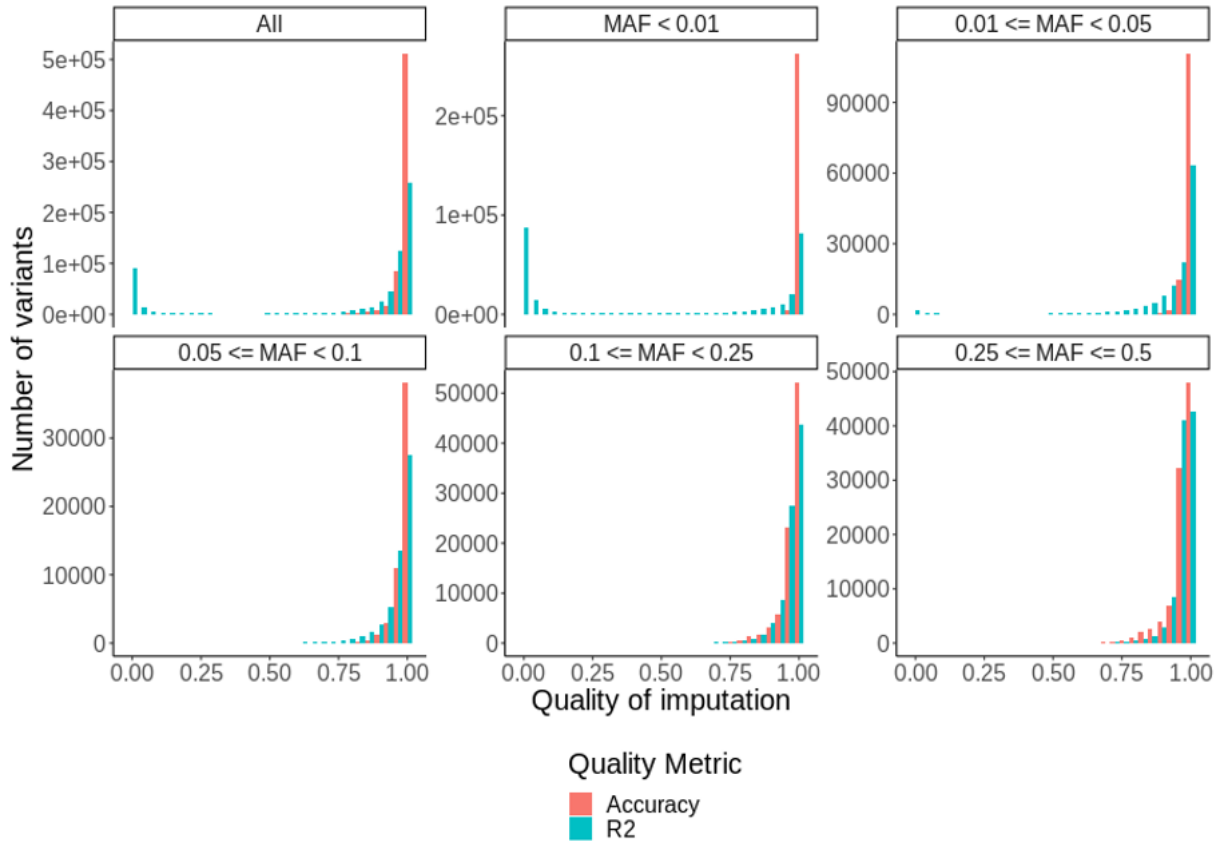


Figure 3.4. Distribution of R2 scores and accuracy of TOPMed imputation with the Impute-Merge strategy according to MAF. R^2 scores are recomputed from the per-array R^2 values using Beagle on the subset of 90 CaG individuals for the 633,283 SNPs from the WES, on which the accuracy, measured in percentage, is computed. MAF is calculated in the WES data.

Next, we investigated the relationship that exists within the sequenced-based imputation accuracy and the R^2 score for the exonic SNPs, split by the R^2 scores: low R^2 (under 0.3), good R^2 (between 0.3 and 0.8) and high R^2 (over 0.8) (*Fig. 3.5.*). First, we see that low R^2 SNPs reached an accuracy close to 100%. More specifically, 14.30% of the variants with a R^2 under 0.3, which would be excluded from the final dataset, have an accuracy of 100%. Second, SNPs that have a high R^2 are not well imputed. In fact, 35.69% of the SNPs with a R^2 equal or greater than 0.3 are not perfectly imputed based on their accuracy (<100%). According to the standard R^2 threshold, these SNPs would be kept even if they are some of the times wrongly imputed and can lead to false discovery on association studies made on the imputed dataset.

We also observed that genomic regions on specific chromosomes have lower imputation accuracy than the rest of the genome (Fig. 3.5, Supp. Fig. 3.6.). In both low and good R^2 scores, Chromosome 15 has regions with the lowest accuracy, reaching 11.11%. Chromosomes 1 and 22 also have regions reaching mean accuracy as low as 45.55 and 47.77%, respectively. Isolated variants on chromosomes 2, 11 and 17 reach accuracy as low as 0%. These observations highlight that the imputation performs differently across genomic loci due to haplotype diversity.

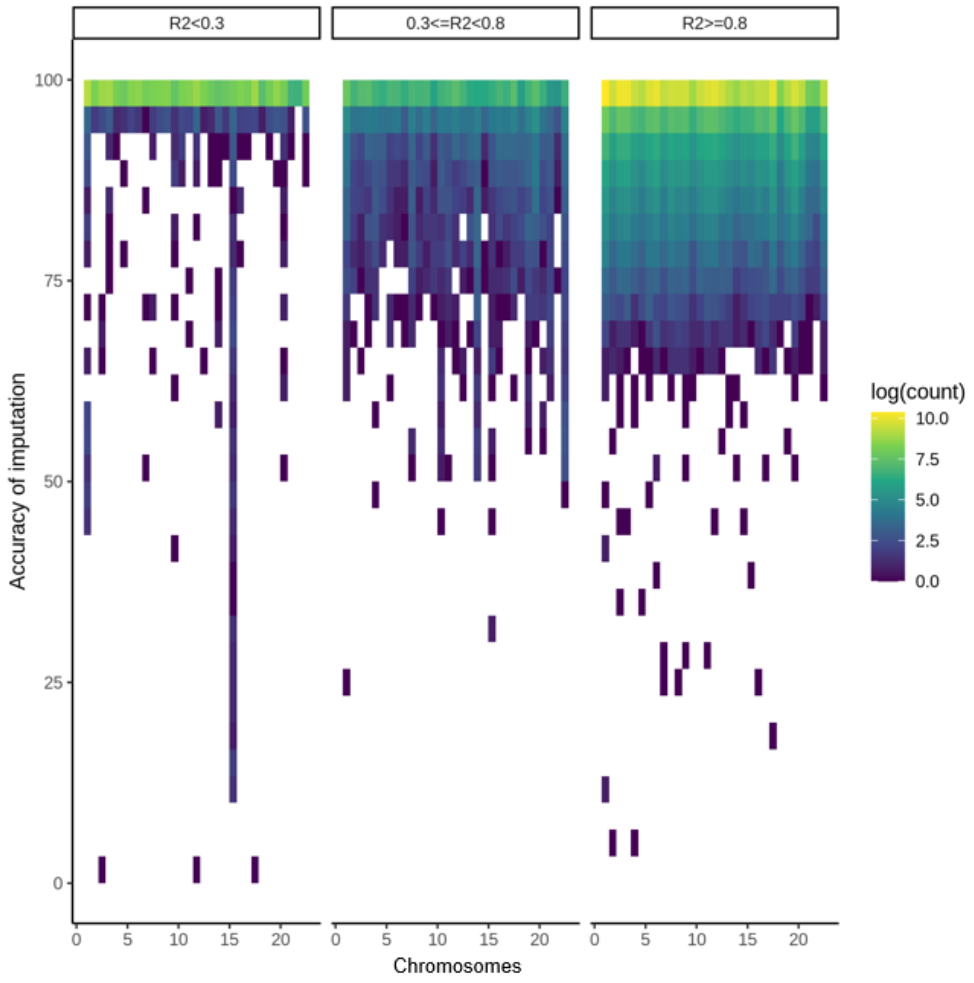


Figure 3.5. Relationship between R2 scores and accuracy of TOPMed imputation with the Impute-Merge strategy. R^2 scores are recomputed using Beagle on the 90 individuals matching the ones obtained with WES. The accuracy, measured in percentage, is computed on the 90 individuals, and reported on the y-axis in 30 bins of 3.33%, for SNPs on each chromosome split in 30 bins (x-axis). The 633,283 SNPs from the WES dataset are split in three R^2 groups and the color scale represents the density of in each bin.

3.4.4 Performance of imputation in French-Canadians compared to other Canadian populations

We have established that TOPMed increases the number of markers with R^2 higher than 0.3 over the previous reference panel HRC when imputing CARTaGENE, enriched for individuals from the FC founder population (Supp Material section 3.5.1). We utilized CanPath to compare imputation performances in FC founder population with other populations of European descent across Canada (*Methods*), to test whether TOPMed indistinguishably imputes individuals from founder and non-founder populations of the same global ethnicity. For five populations from different regions of Canada (Alberta, British-Columbia, Ontario, Atlantic and Quebec), the CanPath project provides similar number of samples genotyped on the same genotyping Affymetrix Axiom UK Biobank array. We excluded, from the five datasets, individuals showing non-European ancestry based on a PCA analysis with 1000G (*Methods*) and recomputed R^2 using Beagle for the remaining individuals. We counted the number of SNPs that have a low imputed R^2 score ($R^2 < 0.3$) in a specific cohort while having a high imputed R^2 score in all four others ($R^2 \geq 0.8$), that we termed "population-specific challenging variants".

The result exhibits a higher number of population-specific challenging variants in CaG sub-cohort than in the others. In fact, CaG has 330,389 low R^2 SNPs that have a high R^2 in all 4 others sub-cohorts while the second cohort in line is ATL, with a number of 232,245 SNPs. ATP comes third with 190 978 SNPs while OHS and BCGP comes respectively penultimate and last with 187 442 and 176 433 SNPs. The fact that CaG has higher numbers of population-specific challenging variants imputed compared to other Canadian-European samples supports the hypothesis that the FC founder population remains harder to impute than other populations of European descent using TOPMed.

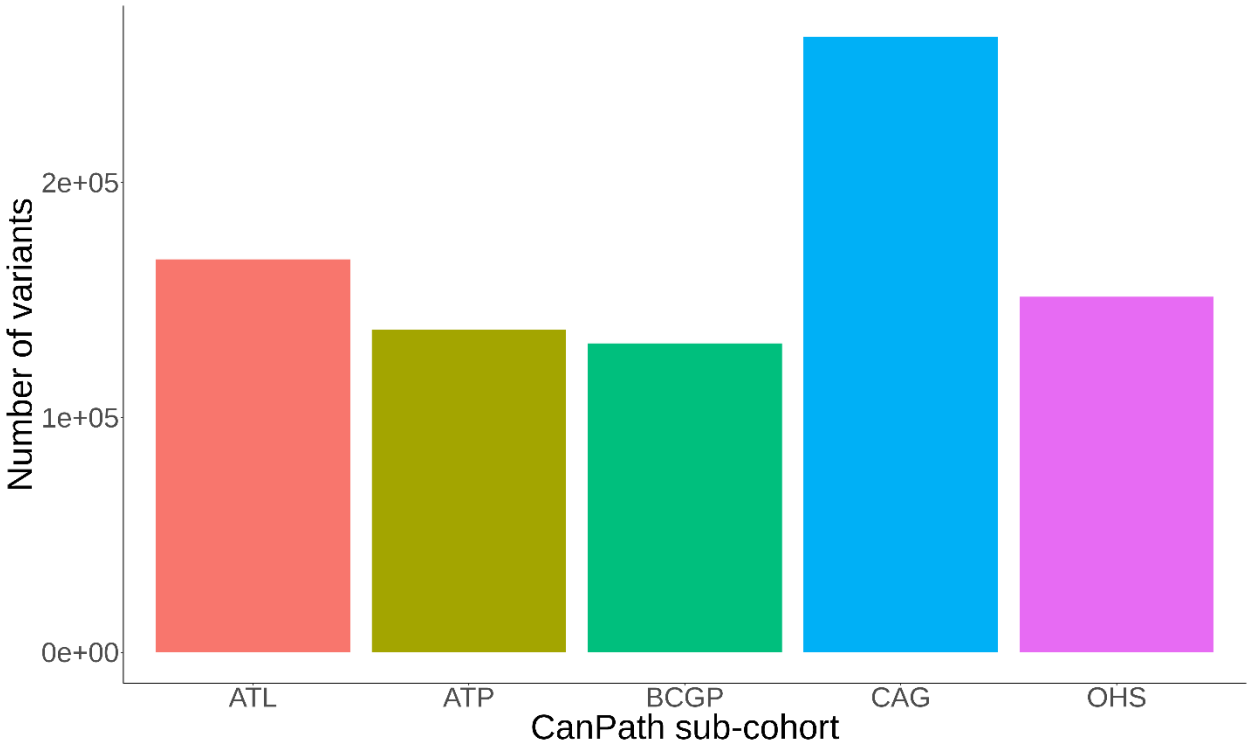


Figure 3.6. Comparison of CanPath sub-cohorts for the number of imputed markers (SNPs and indels) using TOPMed imputation identifying markers that do not pass the $R^2 \geq 0.3$ in the mentioned CanPath sub-cohort on x-axis while having an imputed R^2 score ≥ 0.8 in all four others CanPath sub-cohorts.

3.5 Discussion

We demonstrate that the use of a more diverse reference panel such as TOPMed increases imputation quality for the French-Canadian individuals, even if the quality could be yet improved by including more French-Canadian haplotype diversity in reference panels. We also suggest complementing the current Quality score (R^2) with a real measure of accuracy based on sequencing data for populations not well represented in the reference panel. Finally, we demonstrate that keeping the most information as possible from the genotyping data yields a better imputation quality without introducing significant batch effects.

Extensive analysis of imputation using the two most recent, diverse, and larger reference panels available for genotype imputation, HRC and TOPMed, unveiled an increase in good R^2 and high R^2 variants in FC founder population using TOPMed reference panel compared to its predecessor HRC. This replicates what has been found in the Amish founder population (2). TOPMed imputation outperforms HRC with a higher number of rare variants passing R^2 threshold of 0.3. However, the proportion of good R^2 rare variants is higher with HRC imputation. This result might be explained by the lower number of rare variants in the HRC reference panel. Since the proportion of rare variants is higher in HRC, the proportion of high R^2 imputed variants reflects the average in MAF of HRC variants. In total, TOPMed gets a higher number of good, rare variants, it is probable that the rare variants in HRC also pass the R^2 threshold in TOPMed and that the newest characterizable variants that offers TOPMed are still harder to accurately impute due to lower MAF. These general improvements have been documented while using the wider and more diverse TOPMed reference panel compared to its predecessors (1, 2). Furthermore, TOPMed contains indels, claiming the ability to impute variants other than SNPs, which HRC can't do. This explains part of the increase in number of well imputed variants, although it has been reported that imputing indels is less performant than imputing SNPs due to lack of characterization and rarity of indels in the panel (1). Despite the increase in quality of low frequency variants with TOPMed imputation, 68.3% of the imputed rare variants (MAF<1%) do not pass the quality threshold of $R^2 \geq 0.3$. Even with its difficulty to improve rarer variants, TOPMed still increased the total number of well imputed rare variants. An augmentation of haplotyped diversity and number of haplotype samples could help increase the performance of imputation for these variants.

We demonstrated that, compared to other Canadian populations of European descent included in CanPath, the imputation performance in a founder population was lower.

The number of variants with a R^2 value under 0.3 in a sub-cohort that have a high R^2 value in all others sub-cohort of CanPath is higher in CaG exhibiting the challenges that the FC haplotype diversity represent for imputation. Since every sub-cohort in CanPath have been genotyped using the exact same genotyping array (Axiom), this result represents the challenges of imputation caused by genetic diversity in the FC. Considering the importance that founder populations have in genetic studies, our results reiterate the need for population specific reference panel for accurate genotype imputation in these populations (1, 4). Our results also highlight a higher number of

difficult SNPs to impute in the Atlantic cohort (ATL) compared to other CanPath cohorts, which might be explained by their demographic history with small founder effects and the presence of Acadians, although this hypothesis remains to be investigated.

In most GWAS, the quality of imputation of a variant is assessed by the coefficient of correlation R^2 calculated between imputed allele dosages and masked genotypes. This metric is calculated on posterior probabilities and stands for a fast and accurate way of appraising imputation quality. However, our results suggest that R^2 reflects rarity of imputed variants rather than real imputation accuracy. When measuring imputation quality with an accuracy metric based on "ground truth" obtained from WES in CaG, we observe important discordances with R^2. Indeed, in all MAF categories, we observe higher number of SNPs with high accuracy (over 98%) than SNPs with high R^2 (>0.8), with the effect being amplified as MAF decreases towards rare variants. Rare variants indeed show a higher level of accuracy than what is reported according to the R^2 score distribution, meaning that the R^2 score underestimates the general accuracy of imputation. The standard GWAS R^2 threshold of 0.3 would exclude 19.3% of imputed exonic variants in CaG, of which 98.05% achieved an accuracy over 98% (75.1% achieved 100% accuracy) based on sequencing data. Excluding that many variants would significantly decrease the breath of discovery in association studies, false positive rates need to be considered when including them and corrected with proper multiple testing methods (181). The opposite phenomenon can also be observed with 80.7% of imputed exonic variants passing the R^2 threshold of 0.3 of which 23.55% have an accuracy under 98%. Keeping a lot of false genotypes for further studies and analysis can lead to false discoveries. These results can be explained by the reference panel containing almost no samples from the studied population, meaning that the assessment of quality is biased towards what would be expected in a non-founder European population.

Different genomic regions have been identified as challenging to impute due to high polymorphism levels, such as the HLA region (182-184). Our analysis suggest that the exome is gets better imputation quality compared to the rest of the genome, based on average R^2 scores. The distribution of R^2 scores is on average higher than the overall distribution and the number of low-R^2 SNPs imputed in the exome is lower than the rest of the genome. This shift in quality can be explained by the number of low frequency variants that are present our WES dataset with 41.83% of variants having a MAF under 1% compared to CaG imputation with TOPMed having 82.25%

of its imputed variants being rare. This can be explained by the subsampling of individual used in the WES analysis, with more rare variants not imputed in the samples used in WES compared to the whole cohort. Another possibility that could explain this observation is the distribution of initially genotyped markers on the arrays. Since coding variants are well characterized and mapped in the literature, they are more frequently integrated in the designs of genotyping arrays (185). The initial density of SNPs in these regions is higher in genotyping data and may result in an ease to impute these regions compared to the other chunks. Both factors mentioned impact the quality of imputation in the exome but could be extended to any higher density of SNPs in the genotyping arrays or in the genome. However, current available arrays are mostly design for GWAS purposes and tend to cover most haplotypes blocks in the genome, resulting in an equal imputation quality no matter the initial number of SNPs (28).

Imputation pipelines have been well defined for genotyping data and the availability of the MIS makes their use accessible to all researchers (62). However, no study to date have evaluated the optimal strategy to impute heterogenous genotyped datasets. CaG cohort has been genotyped in batches according to researchers needs and, more recently as a concerted effort, using a total of six different genotyping microarrays available to the research community (13). Here, we compared two potential solutions to generate a single unified dataset while optimizing the imputation of the data, the first one being to impute all genotyping arrays independently and merge the imputation results into one dataset (Impute-Merge method) and the second being to merge all arrays prior to imputing the dataset (Merge-Impute method). The Impute-Merge method shows a significant increase in imputation quality for genomic variants compared to the Merge-Impute methods based R^2 scores from TOPMed imputation. Indeed, the Merge-Impute method shows significantly more high-quality imputed variants, passing the $R^2 > 0.8$ threshold, for every MAF class. The surrogate accuracy measure based on the WES ground truth demonstrates the same result, with significantly more sites having an excellent accuracy (over 98%, meaning that the imputation was erroneous for a maximum of 1 individual out of the samples that were sequenced), and a smaller number of sites having a lower accuracy (under 98%). We therefore conclude that the Merge-Impute method is better suited for the imputation of a disparate dataset. These results demonstrate a clear advantage in conserving as much observed data as possible for the imputation task. Since the Merge-Impute method only keeps common variants within all genotyping arrays, it eliminates a large number of sites that could benefit the imputation algorithm in its inferences. Indeed, the more

observed genotyped variants are phased into haplotypes, the more information can be utilized by the imputation algorithm to infer the correct imputed haplotypes (1, 62, 63, 95). However, batch effects can be observed in heterogenous datasets (102, 103) due to technical bias and any downstream analyses may amplify these effects. We here demonstrate that none of the imputation strategies for the CaG dataset introduce noticeable batch effects in the data, when compared to the genotyping data. Both resulting datasets show no clear differentiation in their principal component's distribution between arrays, with a slight increase in differentiation between arrays for the smaller PCs (PC4-10) in the Merge-Impute method. The largest array discrepancies were seen with both methods, whether the imputation was done before or after the merge as well as in the genotyping data in PC1-3 for the GSA_760, GSA_4224, GSA_5300, GSA_17K arrays. These apparent "array" effects are in fact due to true genetic discrepancies between subsets of samples genotyped in each array because of differing population structure subsampling from the whole CaG cohort. Nevertheless, differences between genotyping arrays post-imputation, for both methods, are slightly higher, however we cannot exclude that this result is due to the first three PCs of the genotyping data and the imputed data being not entirely identical, despite being highly correlated ($r > 0.95$). We therefor recommend using pre-imputation PCs in GWAS. Overall, these results show that no clear additional batch effects were created by either of the imputation strategies.

Although this study demonstrates the importance of genetic diversity and representation of the studied population in reference panels for imputation, there are limitations that remain to be considered. The results presented in this paper are reported for individuals of CaG that cluster with European populations on the PCA with 1000G samples. Except for a subset of the QCRef samples, that are known to have their four grand-parents born in the province of Quebec, confirmed by genealogical analysis in BALSAC population register and the Early Quebec Population Register, evaluating the imputation quality strictly for French-Canadians individuals is a difficult task since we still lack appropriate methods to differentiate them from the rest of the Europeans descents individuals and because of recent admixture with other European populations (specifically for individuals in urban centers) (13). Our dataset is therefore enriched for FC given sampling location but may include a non-negligible proportion of non-FC individuals. The surrogate measure of imputation quality (accuracy) based on WES data maybe not be a perfect "ground truth" as sequencing data can also contain errors (186, 187) and the sample size available for WES data is

only 90 individuals. This accuracy score was also computed on SNPs that are strictly imputed in the exome, more specifically on the coding variants, which we have shown, are easier to impute, meaning that the accuracy is not representative of the genome-wide distribution. Finally, in both cases we had to recompute R^2 score for merged datasets after imputation using Beagle on outputted data by the MIS. We observed a slight shift in higher R^2 in comparison with the MIS output R^2, that we attribute to a dosage value being rounded up in the VCF files.

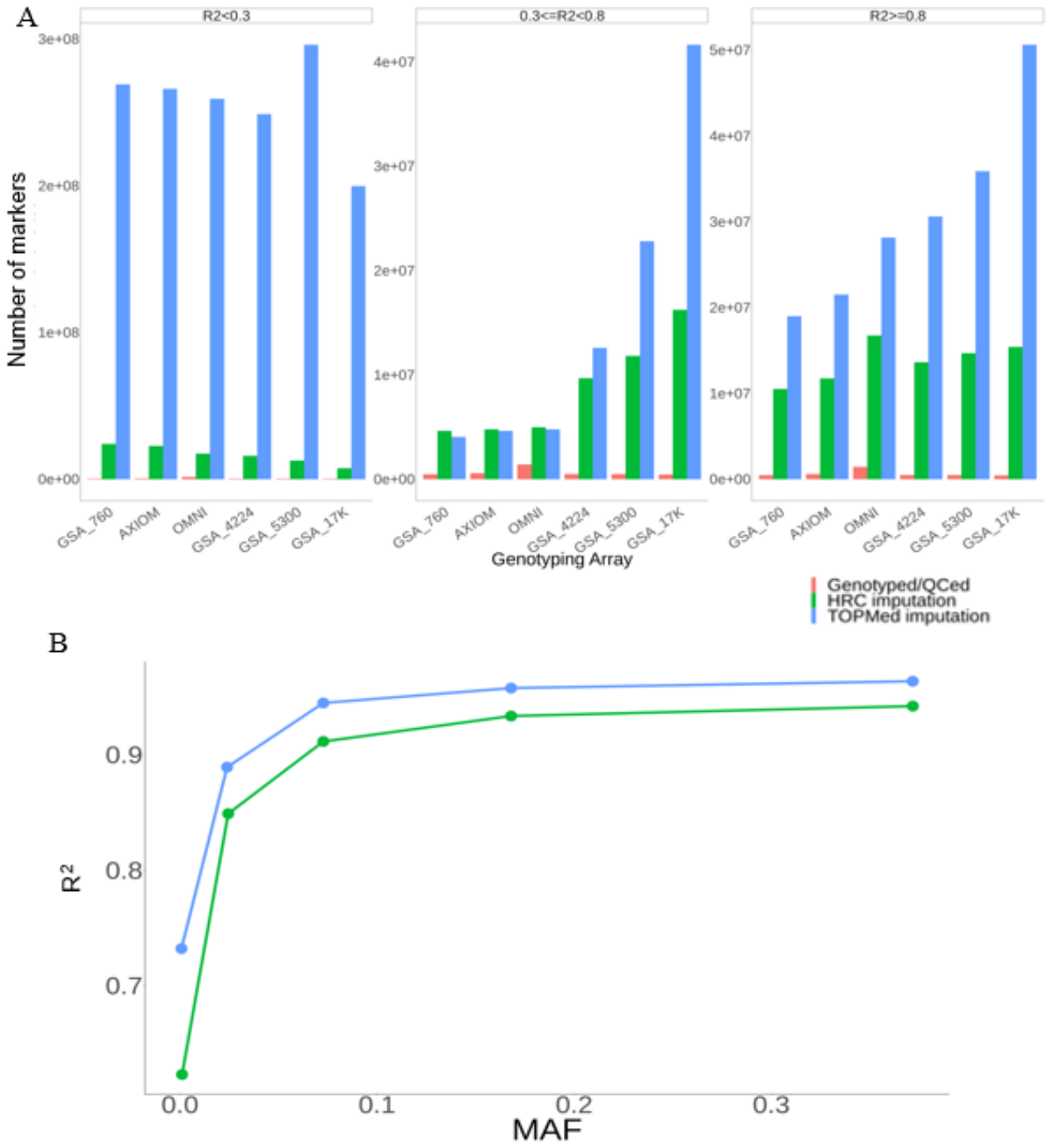
In conclusion, our results demonstrate an improvement in genomic imputation for the founder population of Quebec thanks to the new and more diverse reference panel TOPMed. This dataset promises to increase the bulk of high-quality variants available for future GWAS studies, unveiling the potential to discover genetic loci responsible of complex genetic diseases in CARTaGENE cohort. We also compared imputation strategies to impute and merge datasets genotyped on multiple genotyping arrays and raised questions about the validity of only using R^2 scores to assess imputation quality, a measure that relies more on the MAF of the imputed variant rather than the accuracy of the imputed genotypes. Future work will hopefully include the release of close to 2500 Whole Genome Sequence individuals by the CARTaGENE project, which will be instrumental to assess genome-wide accuracy of imputation, and to build a French-Canadian specific reference panel for imputation in the highly interesting French-Canadian population.

3.6 Supplementary Material

CaG imputation with HRC and TOPMed reference panels

The CaG Impute-Merge imputed dataset with TOPMed was compared to the pre-processed genotyping arrays of CaG using the Haplotype Reference Consortium (HRC) reference panel (90). We used the Michigan imputation server (MIS) (62) for the imputation with HRC reference panel. The Mis uses the same pre-phasing step (Eagle v2.4) (77) and imputation step (Minimac4) (80) as the TIS. HRC is the second largest reference panel available for imputation with 64,976 haplotypes and 39,235,157 SNPs from worldwide populations but mainly Africans and Europeans (83). R^2 score is recalculated after the merge of imputed genotyping array after HRC imputation of CaG with Beagle v5.1 (according to the Impute-Merge method). Imputation performances with both reference panels were compared using the number of sites that were above the R^2 (calculated by Minimac4) thresholds: 0.3 (good quality) and 0.8 (high quality) on each CaG genotyping array separately.

We aimed to demonstrate the improvement that a more diverse and bigger reference panel makes on imputation quality. Compared to what was available before, we imputed every genotyping array available (*Methods*) in CaG with both the HRC and TOPMed reference panels using the Impute-Merge strategy, to determine whether the increase in imputation quality described previously with the TOPMed panel (2) would apply to the French-Canadian founder population. Overall, we see a major improvement using the TOPMed reference panel over HRC (Supp. Fig 3.1.). The number of TOPMed good-quality and high-quality imputed markers, with an R^2 score over 0.3 and 0.8 respectively, is higher than with HRC reference panel. The increase obtained while imputing CaG datasets with TOPMed compared to HRC is about 101.03% and 124.63% for good and high R^2 imputed variants respectively. We note, however, that indels are not present in the HRC panel, diminishing the number of markers available to impute, although it has been reported that TOPMed indel imputation still remains weak (1). Those unequivocal results show a major increase in number of high-quality variants in CaG when using TOPMed as a reference panel for imputation compared to less diverse and smaller sized HRC reference panel.



Supplementary Figure 3.1. Comparison of HRC and TOPMed imputation with the Impute-Merge strategy on CaG genotyping arrays. **A**. Number of imputed markers (SNPs and indels), split in three R^2 scores groups, for HRC and TOPMed imputation reference panels. Total number of pre-processed SNPs, identical in all three graphs, for each SNPs are also shown for each array for scaling purposes. **B**. Distribution of the mean R^2 scores (recomputed using Beagle on the full CaG cohort) according to MAF. Markers with an overall MAF of 0 in CaG cohort are excluded.

We observed a variation in the number of good $R^2 > 0.8$ imputed sites depending on the used array, with both reference panels. All genotyping arrays had initially an average of 497,146 markers left after pre-processing (*Methods*), except for the Omni array with 1,428,380 markers, the number of good R^2 imputed sites varying between the genotyping arrays. Using TOPMed reference panel looking at good R^2 imputed markers, the worse imputation in CaG is observed in the GSA_760 array with 23,055,550 imputed variants, followed by the Axiom array with 26,152,582 imputed variants, Omni array with 32,931,823 imputed variants, GSA_4224 with 43,218,655 imputed variants, GSA_5300 with 53,409,866 imputed variants, and best imputation in the GSA_17K with 92,274,836 good R^2 imputed sites. The same order of imputation quality is observed for HRC good R^2 imputation. In the high R^2 variants, the Omni array dataset outperformed all the others with 16,742,609 SNPs imputed using HRC reference panel. However, the TOPMed imputation reflects the same pattern has the good R^2 variants. Except for the Omni array HRC' imputation numbers for high R^2 variants, the imputation performances increase with sample size in the genotyping array datasets (Tab. 3.1.). The number of low R^2 variants follows the inverted order in terms of number of sites for both imputations, meaning that the proportion of variants above the threshold of 0.3 is higher for the GSA_17K than the GSA_760 with the GSA_5300 being the only anomaly, showing the highest level of low R^2 variants with TOPMed imputation. These results suggest that sample size could increase the estimation of imputation quality calculated by the R^2 score. The proportion of imputed markers that passes the 0.3 R^2 threshold in each genotyping array for TOPMed imputation is smaller than the number of low R^2 variants. This means that most imputed variants have a low quality of imputation.

It has been reported in literature that rare variants are harder to impute because of their lack of representation in the reference panels (63). TOPMed, the biggest and more diverse reference panel to date has unveiled the possibility to improve and to increase the imputation quality of common but also rare variants by having more haplotype diversity (2). Based on Chromosome 1, we evaluated the imputation performance using the R^2 score for TOPMed and HRC in the better, Impute-Merged CARTaGENE dataset. Imputed genotypes were binned by minor allele frequency taken from their respective imputations. As expected, TOPMed panel produces the strongest results for rare, low frequency and common variants. For example, consider SNPs with a MAF of less than 1%, TOPMed achieves an R^2 of 0.73 and HRC which achieves a R^2 of 0.62. The MAF and R^2 score follow an inverse exponential curve, meaning that the R^2 increases rapidly as the

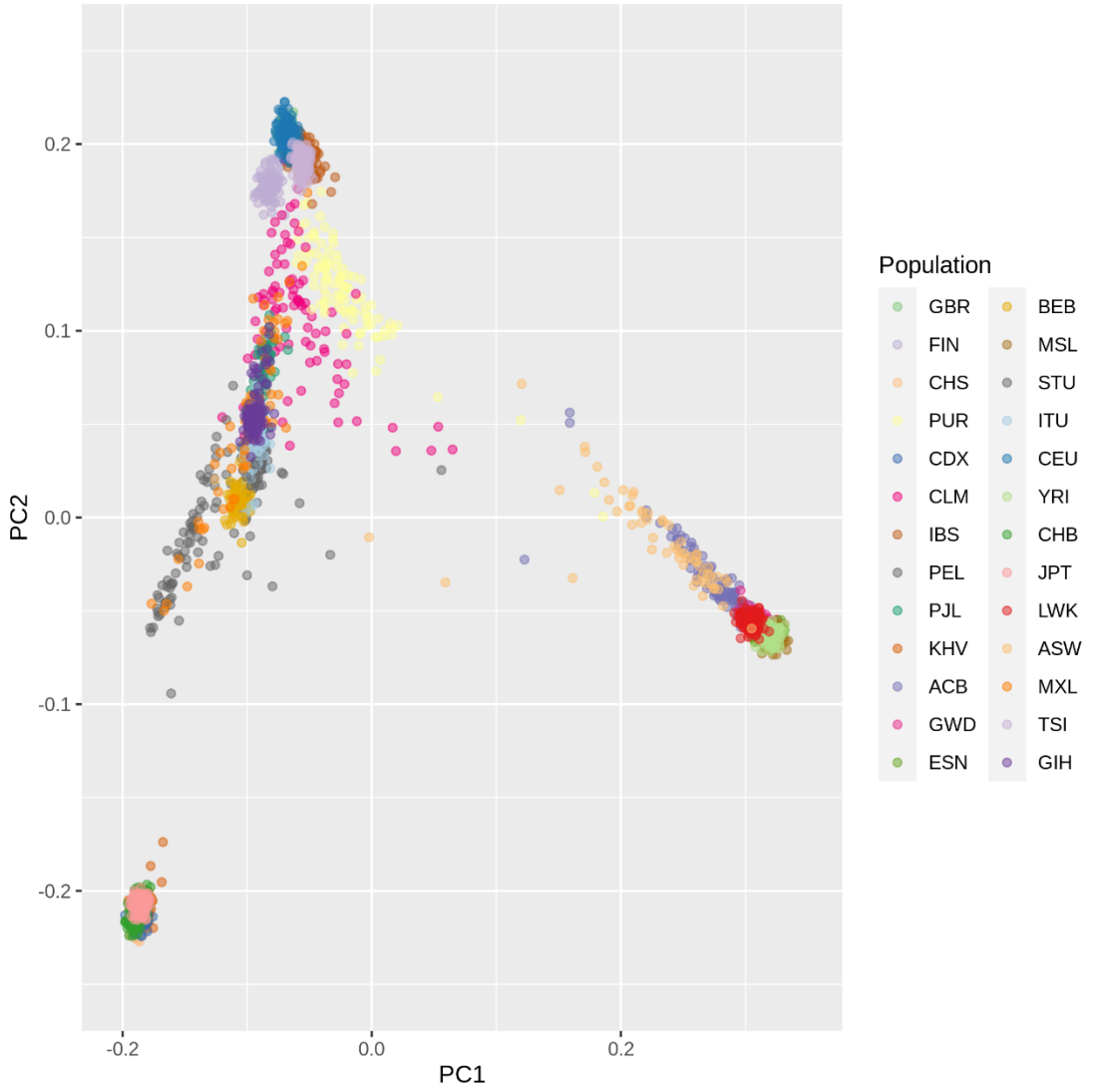
MAF goes higher for rare and low frequency variants (MAF under 5%) and then, increases more slowly with MAF increase amongst common variants. Rare variants in both HRC and TOPMed's imputation have an average R^2 over 0.3 (Supp. Fig. 3.1. B.), stating that they are almost all well imputed. This result is expected since variants with an imputed MAF of 0% are removed from this analysis. The rarer variants are less present in the reference panel's haplotypes meaning that they are less likely to be imputed correctly. Nevertheless, their mean R^2 score is in the good R^2 range when imputing with both imputation panels if they are present in the imputed genotypes.

Supplementary Table 3.1. Description of CaG samples that were genotyped twice on two different genotyping arrays, and the array in which the samples were kept (removed from the other genotyping array).

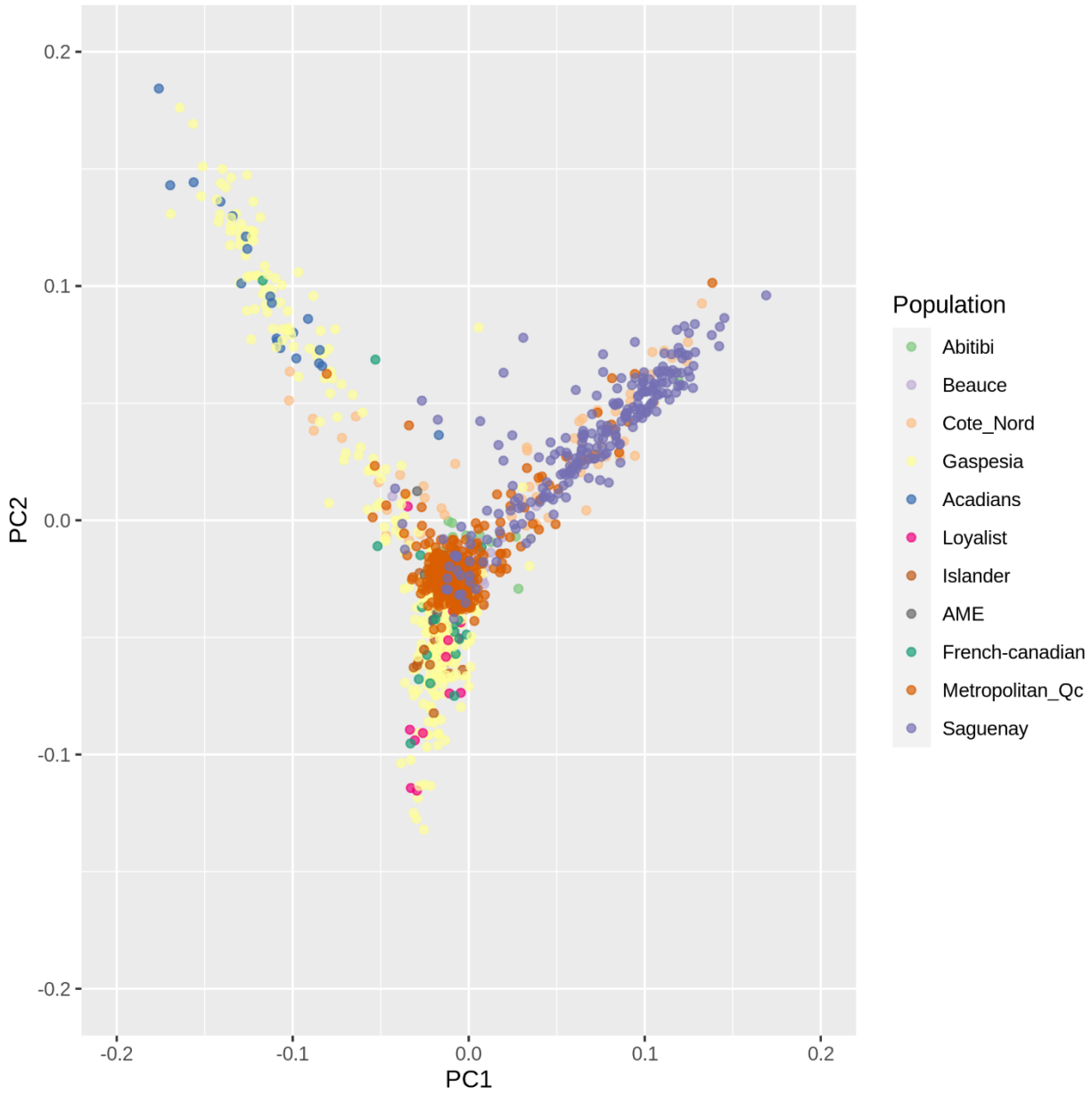
CaG sample ID	Genotyping array 1	Genotyping array 2	Kept genotyping array
11101656	GSA 5300	GSA 760	GSA 5300
11107016	Omni	GSA 4224	Omni
11107238	Omni	GSA 4224	Omni
11110814	GSA 5300	GSA 760	GSA 5300
11125043	Axiom	GSA 4224	Axiom
11134013	Axiom	GSA 4224	Axiom
11137274	GSA 5300	GSA 760	GSA 5300

Supplementary Table 3.2. Summarized description of analysis performed in this study. Detailed by the title and section, datasets used, pre-processing steps and the imputation reference panel.

Analysis	Dataset	Pre-processing	Imputation
Imputation strategies for datasets with multiple genotyping arrays (Section 3.4.1.)	CARTaGENE (GSA_760, GSA_4224, GSA_5300 and GSA_17K)	Pre-processing/QC and imputation per genotyping array individually followed by a merging of all imputed genotyping arrays. Recalculation of the R2 score using Beagle. Merging of the genotyping arrays keeping common variants followed by the Pre-processing/QC. Imputation was done on randomly splitted 15,000 and 14,369 samples. Merging of the two imputation and recalculation of the R2 score using Beagle.	TOPMed imputation server
Detection of genomic bias in CaG imputed data (Section 3.4.1.)	1000G Project (WGS) QCRef Impute-Merge CARTaGENE Impute-Merge CARTaGENE Merge-Impute QCRef genotyped CARTaGENE Genotyped	Imputation using TOPMed imputation server, followed by a filtering for high R2 (≥0.8). Variants in common between all four datasets were merged using Plink. PCA was done with FlashPCA on 1000G individuals and a projection of CARTaGENE Impute-Merge and Merge-Impute samples on the PCs was done to keep European samples only. PCA was done with FlashPCA on the QCRef dataset and a projection of CARTaGENE Impute-Merge and Merge-Impute samples on the PCs. Ks-test were performed between the distribution of CARTaGENE samples in both methods Variants in common between all genotyping arrays datasets were merged using Plink. PCA was performed on the QCRef genotyped samples and CARTaGENE genotyped samples were projected onto these PCs	TOPMed imputation server
Comparison of R^2 quality score and real calculated accuracy (Section 3.4.2)	CARTaGENE Impute-Merge (n=90) CARTaGENE WES (n=90)	Merging and recalculation of imputed genotyping arrays with Beagle. Keeping the common SNPs between CARTaGENE Impute-Merge and CARTaGENE WES datasets and calculation of Accuracy score.	TOPMed imputation server
CaG imputation with HRC and TOPMed reference panels (Section 3.4.3)	CARTaGENE (Axiom, Omni, GSA_760, GSA_4224, GSA_5300 and GSA_17K)	Imputation of every available CARTaGENE genotyping array with HRC and TOPMed. Filtering done for good R2 variants (≥0.3) and high R2 variants (≥0.8).	HRC Michigan imputation server TOPMed imputation server
Performance of imputation in French-Canadians compared to other Canadian populations (Section 3.4.4)	CanPath	Imputation of every available CanPath sub-cohort separately. Calculation of the number of low R2 variants in each cohort that has high R2 score in all other sub-cohorts.	TOPMed imputation server



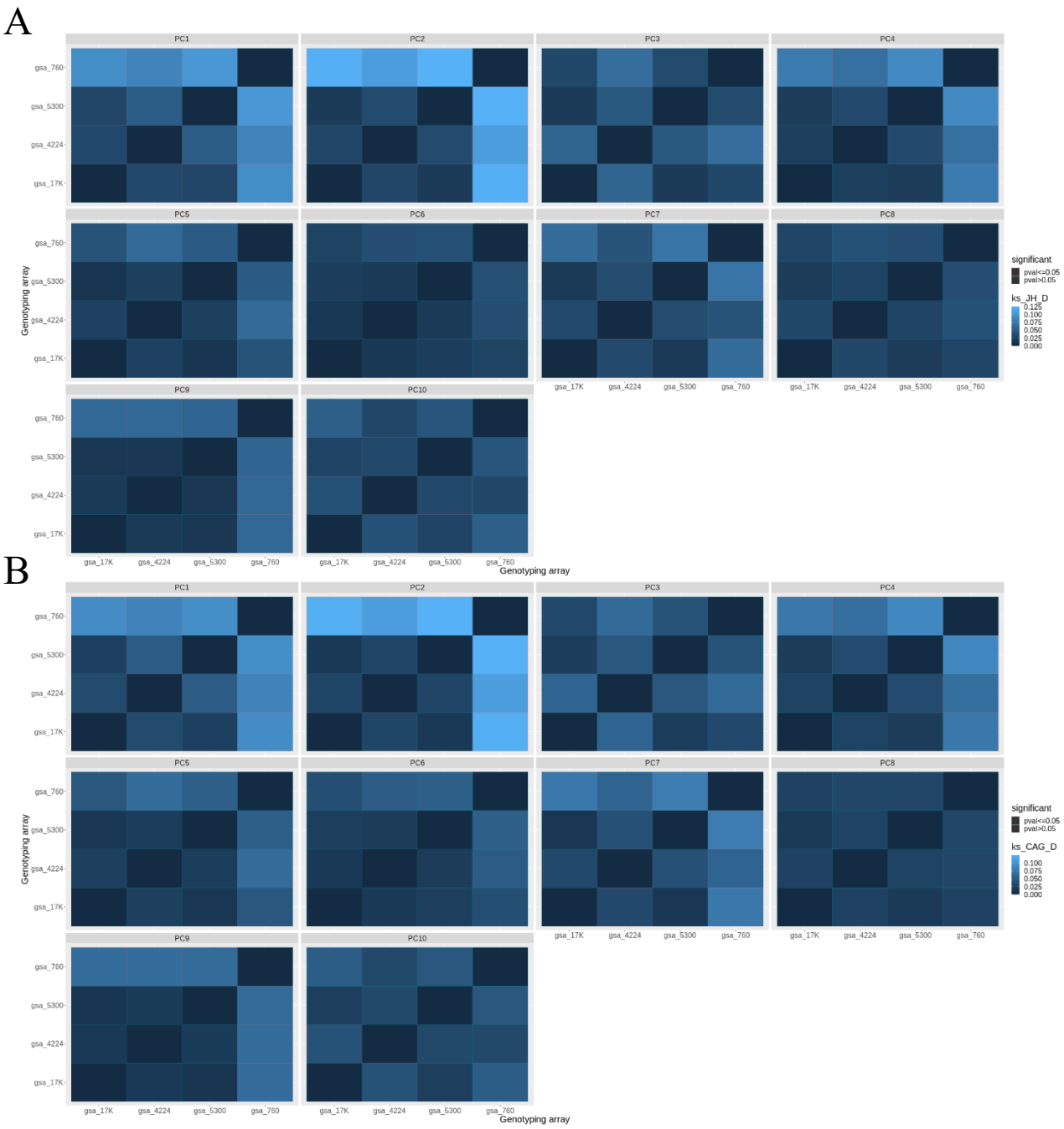
Supplementary Figure 3.2. First two principal components of the 1000G Project PCA on WGS data. Each point is an individual, colored by his or her country of origin.



Supplementary Figure 3.3. First two principal components of the QCRef panel PCA using TOPMed's imputed data. Each point is an individual, colored by his or her sub-region of origin.

Supplementary Table 3.3. Average R2 scores in CaG TOPMed imputation for the Impute-Merge and Merge-Impute strategies, for 633,283 SNPs shared with the WES dataset, stratified by MAF computed in the WES.

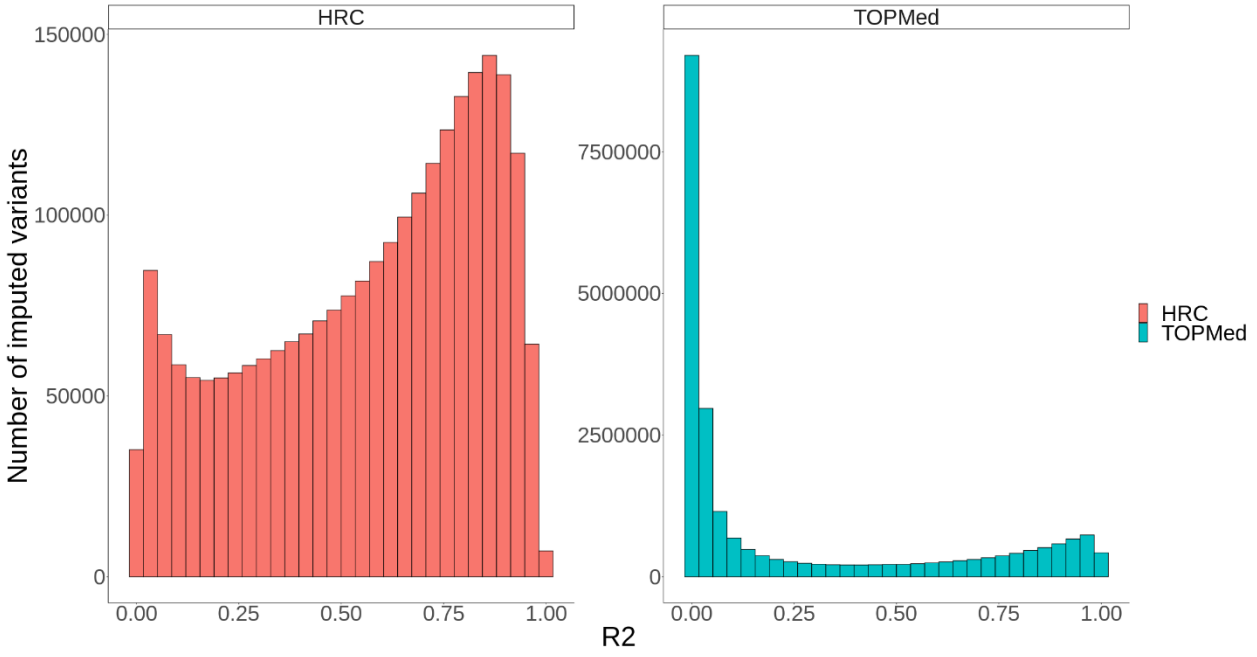
Method	Impute-Merge	Merge-Impute
All	0.769	0.762
MAF<0.01	0.494	0.489
0.01<=MAF<0.05	0.93	0.922
0.05<=MAF<0.1	0.959	0.95
0.1<=MAF<0.25	0.965	0.958
0.25<=MAF<=0.5	0.968	0.962



Supplementary Figure 3.4. Evaluation of differences between genotyping arrays on principal components (PC) to evaluate the presence of batch effects due to the imputation process. PCs are computed on the QCRef Panel. CaG TOPMed imputed data is projected onto the PCs to obtain values for each CaG individuals. D statistics from Kolmogorov Smirnoff test are computed between PC values distributions for CaG individuals for pairs of genotyping arrays (see Methods). **A**. Values of D for CaG Impute-Merge strategy in the first 10 PCs. **B**. Values of D for CaG Merge-Impute strategy in the first 10 PCs.

Supplementary Table 3.4. Coefficient of correlation r between QCRef genotyped and TOPMed imputed PCs.

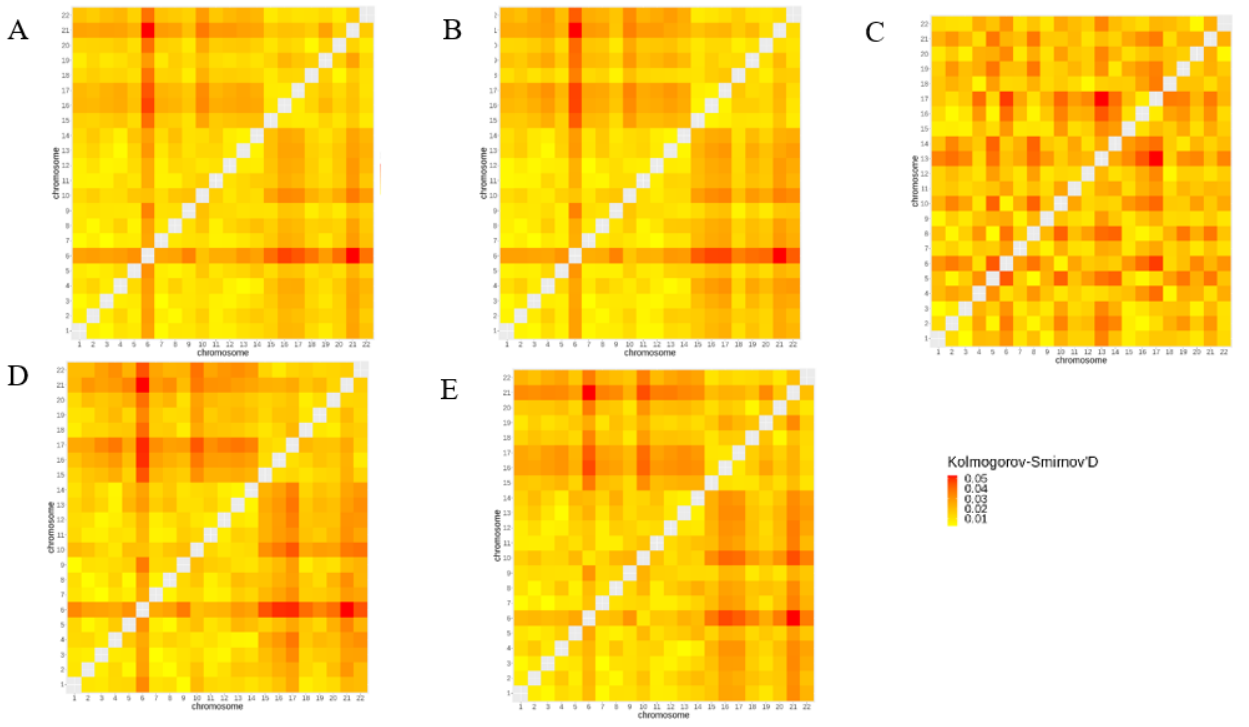
Genotyped PC	Coefficient of correlation	Imputed PC
PC1	0.964919	PC1
PC2	0.9594311	PC2
PC3	0.9594089	PC3
PC4	0.7118778	PC4
PC5	0.6805888	PC5
PC6	0.7761398	PC7
PC7	0.6047984	PC8
PC8	0.8462585	PC9
PC9	0.6207502	PC8
PC10	0.5670008	PC10



Supplementary Figure 3.5. Distribution of the R2 scores in CaG TOPMed imputation for the Impute-Merge markers on chromosome 1. R^2 scores are using Beagle on the merged dataset having 27,429 CaG samples.

Distribution of TOPMed imputation R^2 score along chromosomes

The Imputation performed by TOPMed and HRC are computed by Chunks that are parts of single chromosomes (*Methods*). The number of typed variants in each chunk will help determine the haplotype in the reference panel that resembles the most missing variants to impute. We were interested in seeing if the quality of imputation differs between chromosomes by comparing the distribution of imputation quality with the axiom array while imputing with TOPMed and HRC. Surprisingly, we detected a chromosome that has a significantly different distribution to almost all the other chromosome while performing a Kolmogorov-Smirnov (ks) test on the distribution of R^2. The chromosome 6 appeared to have a different quality of imputation (*Fig. sup. 3.4*.). By investigating the different possibilities that could affect the imputation quality, we removed the repeated elements in the imputed dataset (*Fig. sup. 3.4. B.*). The results stayed the same with chromosome 6 being significantly more different in its imputation. The Human Leukocyte Antigen located on chromosome 6 is a major challenge in genetic and is located on chromosome 6 (188, 189). We removed the region but did not see the effect disappeared (*Fig. sup. 3.4. D.*). After seeing that the imputation quality is affected by the Minor Allele Frequency of the variants, we compared the distribution of MAF in each chromosome, seeing significant difference in almost every chromosome, without replicating the pattern seen in the R^2 distribution. Since Imputation based its inference on genotyped variants (*Methods*) in haplotype chunks, the hypothesis for these results stands in the initial density coverage of genotyped variants in each chromosome. Different genotyping microarrays have specific conception targeting variants that sometimes aren't uniformly distributed across the genome, possibly causing the Imputation to perform differently on various chromosomes.



Supplementary Figure 3.6. Distribution of imputed R2 values on each chromosome for the imputation of the Axiom microarray in CaG. D statistics from Kolmogorov Smirnoff test are computed between R^2 values distributions for CaG individuals for pairs of imputed chromosomes. **A**. Variants imputed with TOPMed. **B**. Variants imputed with TOPMed without repeated regions. **C**. MAF distribution of variants imputed with TOPMed. **D**. Variants imputed with HRC. **E**. Variants imputed with TOPMed without *Human Leukocyte Antigen* (HLA) region.

3.7 Acknowledgements

Most importantly, we are grateful to all participants of the CARTaGENE project, CanPath project, Quebec Reference Panel and the 1000 Genome Project who have kindly provided information and samples. We would also like to thank our collaborators and the scientific team behind CARTaGENE. A special thank you to Dr. Philip Awadalla for sharing CanPath data and to Dr Catherine Laprise for sharing data of the Quebec Reference Panel, as well as the scientific team behind CanPath and the Quebec Reference Panel. Finally, we'd like to thank Dr. Guillaume Lettre and Dr. Simon Gravel, scientific directors of CaG, for supervision of co-author TM who provided datasets and analyses for this study. We thank members of the Hussin group for constructive discussions throughout this project. This work was completed thanks to computational resources provided by Compute Canada clusters Graham, Beluga and Narval. This work was funded by the Institut de Valorisation des Données (IVADO), BioTalent scholarships and the Montreal Heart Institute Foundation.

Chapitre 4 – Synthèse

4.1 Discussion

Depuis que l'impact des variants rares sur les traits et maladies complexes a été démontré (10, 11, 190), le focus des études génomiques s'est tourné vers des variants à plus faible fréquence en plus des variants plus communs. Les événements démographiques complexes récents de l'histoire de la population CF ont entraîné des changements dans les patrons de déséquilibre de liaison par rapport à leurs ancêtres européens (132, 150). Une homogénéité génétique au sein des variants communs, due au goulot d'étranglement, une modification de la fréquence allélique des variants rares, due à la dérive génétique, et un gain de variants rares dû à l'expansion démographique rapide, forme la composition génétique actuelle de la population fondatrice CF. Autant de facteurs qui suscitent un intérêt pour la population fondatrice CF dans la cartographie des variants et l'évaluation des facteurs de risques génétiques des maladies rares (142), mais soulignent davantage la nécessité d'une plus grande représentation de sa diversité dans les panels de référence utilisés pour l'imputation. Dans ce projet, nous avons caractérisé les différentes solutions existantes dans l'objectif d'améliorer l'imputation des données génomiques CF, plus spécifiquement de la cohorte québécoise CARTaGENE (13).

Amélioration grâce à TOPMed

Une analyse approfondie de l'imputation à l'aide des deux panels de référence les plus récents, les plus diversifiés et les plus grands disponibles pour l'imputation des génotypes humains, HRC et TOPMed, nous a dévoilé une augmentation significative des variants de bonne ($R^2 \geq 0.3$) et de haute qualité ($R^2 \geq 0.8$) dans la population fondatrice CF en utilisant le panel de référence TOPMed par rapport à son prédécesseur. L'augmentation des performances d'imputation sur la cohorte CaG s'accordent à ce qui a été observé pour plusieurs populations lors de l'imputation avec TOPMed (1, 2). Ceci réplique les résultats rapportés dans la littérature démontrant l'amélioration

de la qualité de l'imputation dans la population fondatrice Amish grâce à l'utilisation TOPMed (2). En comparant les deux panels de référence, l'imputation TOPMed surpasse HRC avec un nombre plus élevé de variants rares et communs bien imputés. Ces améliorations sont attribuables à la plus grande diversité haplotypique contenu dans le panel de référence TOPMed qui permet d'imputer des variants plus rares et d'étendre les bloc haplotypiques utilisés par les logiciels d'imputation (2). L'éventail de variété des haplotypes présents dans les panels de référence étant le facteur le plus important dans l'imputation des génotypes, la diversité des populations non européennes et métissées présentes dans TOPMed lui octroie l'avantage sur les panels de référence plus petits qui l'ont précédé en augmentant la diversité haplotypique. Sachant que la représentation génétique de la population imputée dans le panel de référence a démontré une amélioration considérable de la qualité de l'imputation (1, 191-193), une partie de l'augmentation de la qualité de l'imputation pourrait être expliquée par la présence d'un petit nombre de génomes CF. Effectivement, la présence d'haplotypes directement dérivés de la population à l'étude, permet de mieux imputer les génotypes des individus qui partagent ces haplotypes. TOPMed contient également un ensemble d'indels en plus des SNPs, revendiquant ainsi la capacité d'imputer des variants autres que des polymorphismes d'un seul nucléotide, ce que HRC ne peut pas faire. Une analyse détaillée de la différence entre indels et SNPs pourrait être intéressante.

Malgré l'amélioration moyenne du score R^2 et du nombre des variants à basse fréquence avec l'imputation TOPMed, la majorité des variants avec une fréquence de l'allèle mineur inférieure à 1 % ne dépassent généralement pas le seuil de qualité de R2 ≥ 0,3 (68.3%). Comme les variants plus rares ont moins de chance d'être observés sur les haplotypes présents dans le panel de référence, on s'attend à ce qu'ils obtiennent une qualité d'imputation inférieure aux variants communs. L'utilisation de TOPMed pour l'imputation démontre une amélioration concrète de la qualité d'imputation des CF. Grâce à l'étendue de sa diversité haplotypique recensée ainsi qu'à son intégration des indels en plus des SNPs, il permet d'étendre l'imputation à de nouveaux variants et de perfectionner l'imputation de variants rares et communs. Néanmoins, le panel actuel n'est possiblement pas suffisamment complet et diversifié pour imputer avec précision l'ensemble des variants plus rares et spécifiques à des populations fondatrices. En effet, nous démontrons ici que l'imputation de la population CF présente encore des défis en comparaison avec d'autres populations canadiennes de descendance européenne. La comparaison du nombre de sites imputés de basse qualité ($R^2 < 0.3$) qui ont un haut R^2 (> 0.8) pour toutes les autres sous-cohortes

canadiennes de CanPath est plus élevé dans l'imputation de CaG. Toutefois, les variants présentés dans cette analyse comptent les variants imputés à une MAF de 0. Puisque nous savons que les variants avec une MAF faible sont souvent très mal imputés, il est possible que ce résultat reflète en soit le nombre de variants absent des CF mais présent dans les autres sous-populations. Cet excès s'expliquerait partiellement par le plus faible taux de diversité présent dans la population fondatrice CF.

De plus, des traces de métissage avec les populations autochtones présentes sur le territoire du Québec a été répertorié (132). Ce métissage entraîne la création de nouveaux haplotypes et une nouvelle diversité génétique qui est très peu documentée en raison de la difficulté d'obtention d'échantillon d'ADN pour ces populations américaine natives (193). Le manque de caractérisation des CF en plus de l'absence de populations autochtones dans TOPMed peut expliquer ce nombre plus élevé de variants difficilement imputables dans CaG. Compte tenu de l'importance de ces populations pour les études génomiques, notre plus récente démonstration des difficultés d'imputation chez les Canadiens français réitère la nécessité d'un panel de référence spécifique à la population pour une imputation précise du génotype (1, 4).

Méthodes d'évaluation de qualité d'imputation

Avec Minimac4 étant le logiciel utilisé par le serveur d'imputation MIS et celui de TIS, la qualité d'imputation d'un variant est quantifiée par le coefficient de corrélation entre les dosages d'allèles imputés et les vrais génotypes non-observés, le score d'imputation R^2 (80, 81). Cette métrique est calculée selon des probabilités *a posteriori* des génotypes et représente un moyen rapide et précis d'évaluer la qualité de l'imputation. Nous montrons dans cette étude que les scores R^2 tendent à refléter le MAF du variant imputé plus que l'exactitude d'imputation réelle du génotype. Lorsque l'on mesure la qualité d'imputation avec une métrique d'exactitude basée sur des données séquencées obtenues à partir de WES dans CaG que l'on considère comme étant la « vérité absolue », nous observons des discordances avec le score R^2. Le nombre de variants rares (MAF < 1%) qui ont un haut taux d'exactitude (bien imputés dans plus de 98% des échantillons) est plus élevé que le nombre de variants possédant un haut R^2. Un phénomène similaire est observé pour les variants à faibles fréquences (1% < MAF < 5%). À l'opposé, une surestimation du nombre de variants de haute exactitude est reportée par le nombre de variants à haut R^2. Cela signifie que ce score sous-estime la précision générale de l'imputation et tend à catégoriser les variants rares

comme étant mal imputées en raison de l'importance de la fréquence des allèles mineurs dans le calcul du score R^2.

Le seuil standard GWAS de R^2 à 0.3 exclut 19,3 % des variants exoniques imputés dans CaG, dont 98,05 % ont atteint une précision supérieure à 98 % et 75,1 % ont atteint une précision de 100 %, sur la base des données de séquençage WES. L'exclusion de nombreux variants d'une analyse basé sur le seuil établi du R^2 réduit donc considérablement la puissance de découverte dans les études GWAS, toutefois, leur inclusion exigerait l'application de tests multiples et de méthodes de correction pour le taux de faux-positif également (181). Le nombre de variants communs n'ayant pas de taux d'exactitude très élevé (<98%) peut lui mener à de fausses découvertes lors d'études GWAS. Ces résultats peuvent être associés au fait que le panel de référence ne contient presque pas d'échantillons de la population étudiée, ce qui signifie que l'évaluation de la qualité est biaisée par rapport à ce qui serait attendu dans une population européenne non fondatrice dans le calcul du R^2. Un score de qualité basé directement sur les dosages, n'incluant pas la fréquence allélique pourrait permettre de représenter de manière exacte la qualité d'imputation.

Un GWAS sur les sites partagés entre les données de génotypage filtrées et sur ces mêmes positions post-imputation de CaG révèle une discordance importante dans les fréquences alléliques pré et post imputation révélée par les hautes p-values entre les données pré et post imputation pour les mêmes individus dans les deux stratégies d'imputation (Figure 4.1.). Ces résultats suggèrent que le serveur d'imputation de TOPMed exclut certains variants génotypés pour les réimputer selon une fréquence se rapprochant de celle présente dans les populations européenne non-finlandaise (en comparant avec gnomAD (194)). Ce phénomène se produit dans 0.21% des variants génotypés donnés en entrée pour l'imputation communs dans les deux méthodes. La cause de ce phénomène est actuellement sous investigation. Cet uniformisation des données faites par le serveur d'imputation de TOPMed pourrait nuire à l'identification des variants spécifiques à la population dans ces régions en changeant l'haplotype canadien-français qui possède des fréquences différentes des populations européennes en un haplotype européen. Ces résultats viennent soutenir l'importance d'utiliser un panel de référence pour l'imputation représentatif de la population étudiée (1, 5, 179, 192).

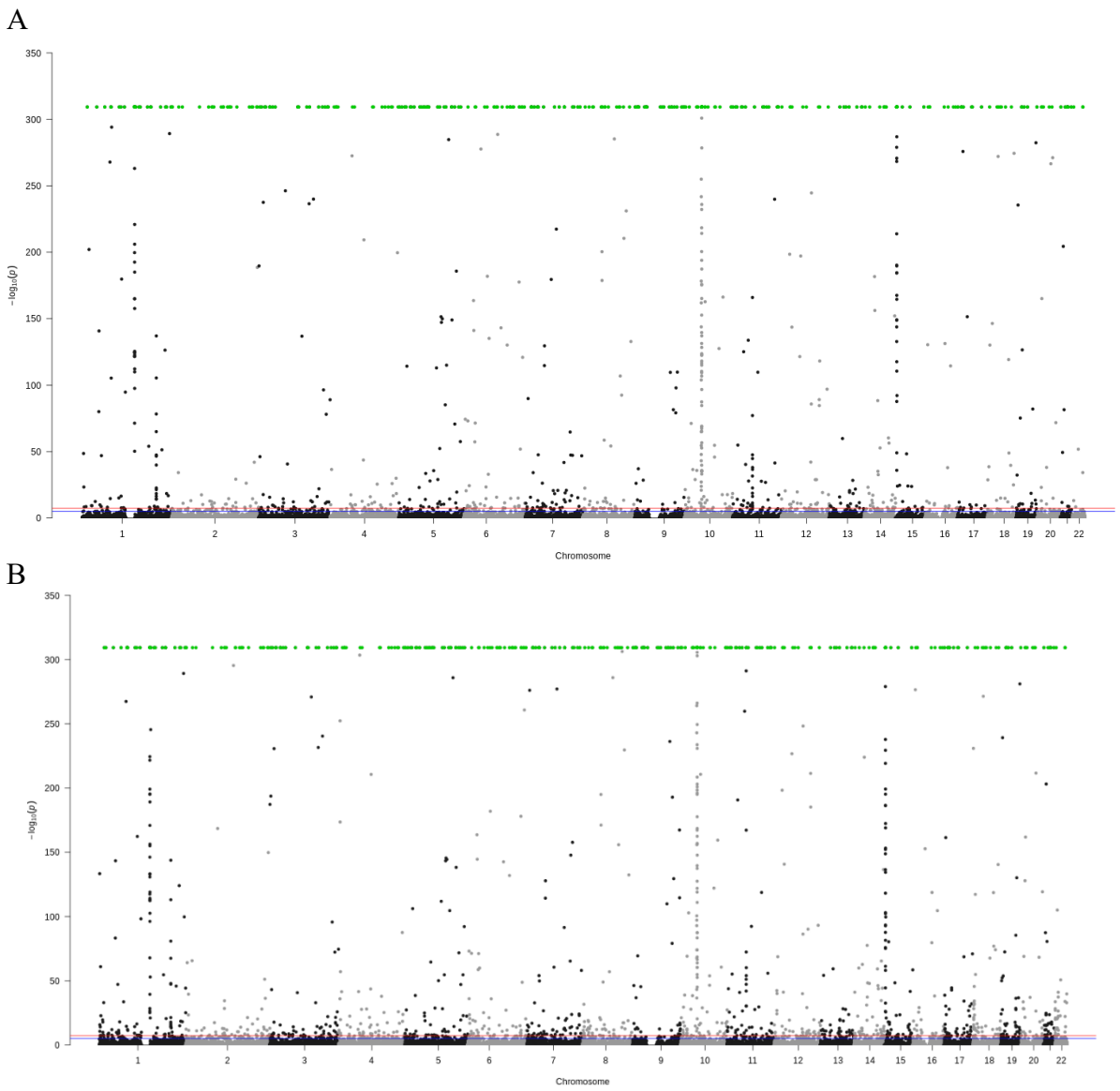


Figure 4.1. Comparaison des SNP et des indels génotypés et imputés à l'aide de l'imputation TOPMed pour la stratégie Impute-Merge et Merge-Impute. Manhattan plot des résultats de l'analyse GWAS des SNPs génotypés pré et post imputation à l'aide du serveur du TOPMed imputation server. L'axe y représente -$\log_{10}$ (valeurs P) pour l'association des variants avec l'imputation, à partir de l'analyse GWAS contrôles. Les variants avec une valeur p de 0 (-$\log_{10}$(P)=Infini) sont mis en évidence sous forme de points verts et leur valeur a été fixée à $4{,}768^{e-310}$. Une taille d'échantillon totale de 27,422 individus dupliqués comme 27,422 contrôles (génotypés) et 27,422 cas (imputés). La ligne rouge horizontale représente le seuil de signification à l'échelle du génome. **A.** Manhattan plot de la stratégie Impute-Merge. **B.** Manhattan plot de la stratégie Merge-Impute.

Spécificité de l'exome

Différentes régions génomiques ont été identifiées comme étant difficiles à imputer en raison des niveaux élevés de polymorphisme tels que la région HLA (182-184). Notre analyse suggère que l'exome est plus facile à imputer que le reste du génome. En effet, la distribution des scores R^2 est en moyenne supérieure à la distribution globale et le nombre de SNPs de mauvaise qualité imputés dans l'exome est inférieur au reste du génome. Cette différence de qualité peut s'expliquer par le nombre plus faible de variants à basse fréquence présents dans notre jeu de donné WES, dans lequel 41,83% des variants ont un MAF sous 1% alors que 82.25% des variants imputés par TOPMed dans l'ensemble du génome ont un MAF sous 1%. Ceci s'explique par le sous-échantillonnage des individus qui ont été génotypés et WES, un nombre plus élevé de variants imputés mais absents dans les 90 échantillons WES (MAF = 0) en comparaison avec la cohorte complète de CaG. Une autre possibilité qui pourrait expliquer cette observation est la distribution de marqueurs initialement génotypés sur les puces de génotypages. Étant donné que les variants codants sont bien répertoriés et cartographiées dans la littérature, ils sont plus fréquemment intégrés dans les conceptions de puces à ADN (185). La densité de SNPs dans ces régions est plus élevée et se traduit par une facilité d'imputation en comparaison avec certains autres tronçons (*chunks)*. Les deux facteurs mentionnés ci-haut affectent la qualité de l'imputation dans l'exome mais leur impact pourrait être étendus à toute région génomique où la densité de SNPs génotypés est plus élevée, que ce soit dans puces à ADN ou dans le génome. Toutefois, une récente étude a révélé que la plupart des puces de génotypage étaient conçues pour les GWAS et couvraient donc la majorité des bloc haplotypiques du génome humain, faisant en sorte que le nombre de variants initialement génotypé avait un plus faible impact sur la qualité d'imputation (28).

Stratégies d'imputation sur jeux de données hétérogènes.

Les pipelines d'imputation ont été bien définis pour les données de génotypage grâce à la disponibilité du MIT (62). Cependant, aucune étude à ce jour n'a évalué la stratégie optimale pour imputer des jeux de données de génotypage hétérogènes. La cohorte CARTaGENE a été génotypée en six lots à l'aide de six puces de génotypage différentes (13). Nous avons comparé les deux solutions envisageables afin d'optimiser la qualité de l'imputation des données. La première est d'imputer toutes les puces indépendamment et de fusionner les résultats d'imputation en une seule cohorte (méthode Impute-Merge), la seconde étant de fusionner les puces avant de les imputer

(méthode de Merge-Impute). La méthode Impute-Merge montre une augmentation significative de la qualité d'imputation pour les variants génomiques, basée sur la qualité d'imputation déterminée par le score R^2 de Minimac4. Avec une densité de variants un peu plus élevée dans la partie supérieure de la distribution du score et une concentration plus faible dans les faibles R^2, la méthode Impute-Merge augmente significativement les chances qu'un variant obtienne une bonne imputation (peu importe la fréquence allélique de ce variant). Le score d'exactitude se basant sur les données séquencées démontre la même tendance, comptant plus de sites ayant une excellente précision (plus de 98%) pour la méthode Impute-Merge, et moins de sites ayant une mauvaise précision (moins de 98%). La réplication de l'amélioration de la qualité d'imputation mesurée par R^2 et par l'exactitude démontre un net avantage de la méthode Impute-Merge. Ce phénomène s'explique par la conservation de la plupart des données disponibles (génotypées) pour l'imputation. En effet, le nombre de variants et la densité de sites génotypés affecte directement le la construction des haplotypes phasés. Ainsi, un plus grand nombre de variants caractérisés permet d'étendre les haplotypes et facilite ainsi l'inférence des génotypes manquants avec l'algorithme d'imputation qui se base sur des haplotypes de référence. Cela signifie qu'une augmentation de la couverture initiale du génome dans les données de génotypages entraîne une augmentation de l'imputation de sites de qualité (1, 62, 63, 95). Or, la méthode Merge-Impute ne conserve que les variants communs dans toutes les puces de génotypage et élimine donc de nombreux sites qui pourraient bénéficier à l'algorithme d'imputation dans ses inférences sur les génotypes manquants. Nous concluons donc que la méthode Impute-Merge est mieux adaptée à l'imputation d'un ensemble de données disparates. Ces résultats démontrent un net avantage à conserver autant de données génotypées que possible pour l'imputation.

Nous démontrons aussi qu'aucune des stratégies d'imputation pour l'ensemble de données CaG n'introduit d'effets de lot (102, 103) notables dans les données, au-delà de celui pouvant exister au sein des données de génotypage. Les deux ensembles de données résultants ne montrent aucune différenciation claire dans la distribution de leur composantes principales (PCs) entre les puces de génotypage, avec une légère augmentation de la différenciation entre les puces pour les plus petits PCs (PC4-10) dans la méthode Merge-Impute. Les écarts de puces de génotypage les plus importants ont été observés dans les deux stratégies d'imputation ainsi que dans les données de génotypage dans PC1-3 pour les tableaux GSA_760, GSA_4224, GSA_5300, GSA_17K. Ces effets apparents de "puces" sont en fait dus à de véritables divergences génétiques entre des sous-

ensembles d'échantillons génotypés dans chaque puce en raison de critères de sélection différents des individus pour des lots de génotypage différents dans la cohorte de CARTaGENE (13). Néanmoins, nous observons que les différences entre les puces de génotypage post-imputation, pour les deux méthodes, sont légèrement plus élevées que celles observées dans les données génotypés. Nous ne pouvons pas exclure que ce résultat soit dû au fait que les trois premiers PC des données de génotypage et les données imputées ne sont pas entièrement identiques, bien qu'elles soient fortement corrélées ($r > 0.95$). Une méthode d'apprentissage automatique pour faire la prédiction de la puce de génotypage (tache de classification), par un *Diet Network* (DietNet) par exemple (195), permettrait de détecter des effets cryptiques dans les données au-delà des PCs. En comparant les exactitudes des prédictions dépendamment des données d'entrées (données génotypées ou imputées, pour la méthode Impute-Merge ou Merge-Impute). Dans l'ensemble, ces résultats montrent qu'aucun effet de lot supplémentaire clair n'a été créé par l'une ou l'autre des stratégies d'imputation.

Limites de l'étude

Bien que les résultats de ce projet démontrent l'importance de la diversité génétique et de la représentation de la population étudiée dans les panels de référence pour l'imputation, certaines limites doivent être prises en compte. L'évaluation de la qualité d'imputation pour les individus strictement CF est difficile à faire puisque nous n'avons pas encore de méthode satisfaisante pour les différencier du reste des individus de descendance européenne. L'ensemble des Caucasiens de la cohorte CARTaGENE (CF et Européens) s'agglomèrent sur les populations européennes du projet des 1000 Génomes. Au sein de la population CF uniquement, il existe des individus indistinguables des individus Européens, en raison notamment du métissage récent avec des descendants européens (spécifiquement pour les individus de centres urbains) (12). Les résultats présentés dans ce mémoire sont donc effectués sur la population Québécoise d'origine européenne dans son ensemble, qui est enrichie pour les Canadiens-français. De plus, nous n'avons pas spécifiquement investigué la région HLA dans notre population à l'étude en raison du manque de caractérisation des fréquences chez les Canadiens-français. Toutefois, une étude plus approfondie de l'imputation de cette région dans CaG serait informative.

Une autre limitation de ce projet a trait à la mesure alternative d'exactitude de l'imputation (basée sur les génotypes séquencés) que nous avons utilisée. Cette méthode se base sur des données de WES établissant la « vérité ». Or, il est possible que ces données contiennent des erreurs liés aux technologies de séquençage et au traitement des données (control de qualité) (186). Par exemple, le taux d'erreur des technologies de séquençage de nouvelle génération se situe entre 0.1 et 0.6%, donc certains variants ont pu être mal séquencés (196). De plus, ce score d'exactitude a également été calculé sur les SNPs qui sont strictement imputés dans l'exome, ce qui signifie que la précision n'est pas représentative de la qualité globale de l'imputation, mais plus précisément sur les variants codants, dont nous avons démontré une meilleure imputation que les variants non-codants. Également, la taille de l'échantillon disponible au moment de l'étude pour les séquences WES n'est que de 90 individus, ce qui fait en sorte que le score d'exactitude de tous les variants n'est calculé que sur 0.30% de la cohorte totale. Cet échantillon pourrait être non-représentatif de l'ensemble des données et pourrait apporter un biais dans l'interprétation des résultats.

La dernière limitation réside dans le score de qualité R^2. Puisqu'il est recalculé pour chacune des cohortes, il n'est pas directement pris des fichiers VCF sortant du serveur d'imputation (MIS) car il est impossible de soumettre plus de 15,000 séquences à la fois. Tel que mentionné dans les figures supplémentaires de l'article au Chapitre 3, nous avons dû recalculer le score R^2 après avoir fusionné les différents lots de CARTaGENE, pour les deux méthodes. Nous avons remarqué que le score R^2 de Beagle démontre une légère surestimation par rapport à celui calculé par le MIS, qui pourrait s'expliquer par l'arrondissement à la hausse des valeurs de dosages dans les fichiers VCF. Toutefois, puisque cette étape est répliquée dans toutes nos analyses, nos comparaisons demeurent valides.

4.2 Perspectives

Pour conclure, les résultats de ce projet de maîtrise ont démontrés que l'augmentation de la diversité et de la représentation des populations non-européennes dans les jeux de données génomiques servant à l'imputation étaient essentiels afin d'augmenter la qualité des jeux de donnés imputés. Ces données permettront d'augmenter le potentiel de découverte des variants associés aux traits et maladies complexes (52). Malgré les améliorations récentes dans le domaine, les populations à diversité génétique particulière, et plus spécifiquement les populations fondatrices telles que la population Canadienne-Française présentent toujours un défi d'imputation en raison du manque de représentation de leur diversité haplotypique dans les panels de référence. Nous avons également soulevé la question quant à l'interprétation du score de qualité d'imputation R^2 qui semble être biaisé vers une représentation de la fréquence de l'allèle mineur plutôt que de la véritable exactitude. Finalement, nous avons établi une stratégie de fusion des différentes puces de génotypage composant une même cohorte disparate en termes de technique de génotypage pour l'acquisition de données. L'imputation individuelle de chaque puce suivi de la fusion atteint meilleurs performances grâce à la conservation des données initiales permettant d'exploiter l'ensemble de la diversité caractérisée.

Ce projet est le premier à caractériser le défi que représente la population fondatrice Canadienne-française lors de l'imputation à partir des données de génotypage. Les données disponibles pour la cohorte québécoise CARTaGENE nous ont permis d'utiliser le séquençage de l'exome de 90 individus pour mesurer précisément la qualité d'imputation des données avec TOPMed. Cette stratégie nous as permis de constater que le R^2 utilisé dans la littérature pour qualifier l'imputation ne reflète pas toujours la véritable exactitude des génotypes imputés et nécessiterait d'être complémentée. Il serait intéressant de développer un nouveau score de qualité qui reflète plus précisément la véritable qualité d'imputation afin d'augmenter l'exactitude des études utilisant ces données imputées. Par exemple un panel de référence spécifique à la population imputée qui serait utilisé pour calculer les différences de fréquences obtenues dans les données imputées en comparaison avec les fréquences attendues dans la population. De surcroit, l'utilisation des données de séquençage du génome entier (WGS) plutôt que de l'exome. CARTaGENE publiera très bientôt les données de WGS pour environ 2500 individus qui sont

aussi génotypés. La mesure de l'exactitude pourra donc être calculées sur ces données et sera plus représentative de la qualité d'imputation pour tous les types de variants en plus d'avoir une plus grande taille d'échantillon.

Nos résultats démontrant l'importance de la représentation de la diversité haplotypique de la population CF, ils ouvrent la porte à la nécessité de développer un futur d'un panel de référence CF pour l'imputation. La future publication des données de WGS de CARTaGENE mentionnés ci-haut ainsi que l'apport de l'ensemble des données de séquençage des biobanques québécoises existantes telles que celles de l'Institut de Cardiologie de Montréal, de Génome Québec (Genizon) ainsi que des nombreuses séquences produites durant les deux dernières années d'épidémie de SARS-COV-2 faisant partie de la BQC19, permettront d'améliorer grandement la qualité de l'imputation de la population fondatrice. Il serait également intéressant de comparer l'utilisation d'un panel de référence spécifique à celui d'un panel générique auquel on intègrerait le panel spécifique, de manière à vérifier si une trop grande diversité haplotypique et une sous-représentation de la population à l'étude affecte la qualité d'imputation. Ces avancées permettront ultimement d'exploiter tous les avantages que procurent une population fondatrice dans les études d'association génétiques.

Références bibliographiques

1. O'Connell J, Yun T, Moreno M, Li H, Litterman N, Kolesnikov A, et al. A population-specific reference panel for improved genotype imputation in African Americans. Commun Biol. 2021;4(1):1269-.

2. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature. 2021;590(7845):290-9.

3. Vergara C, Parker MM, Franco L, Cho MH, Valencia-Duarte AV, Beaty TH, et al. Genotype imputation performance of three reference panels using African ancestry individuals. Human genetics. 2018;137(4):281-92.

4. Ritari J, Hyvärinen K, Clancy J, Partanen J, Koskela S. Increasing accuracy of HLA imputation by a population-specific reference panel in a FinnGen biobank cohort. NAR Genom Bioinform. 2020;2(2):lqaa030.

5. Hou L, Kember RL, Roach JC, O'Connell JR, Craig DW, Bucan M, et al. A population-specific reference panel empowers genetic studies of Anabaptist populations. Sci Rep. 2017;7(1):6079.

6. Bai W-Y, Zhu X-W, Cong P-K, Zhang X-J, Richards JB, Zheng H-F. Genotype imputation and reference panel: a systematic evaluation on haplotype size and diversity. Briefings in Bioinformatics. 2019;21(5):1806-17.

7. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nature Genetics. 2016;48(10):1284-7.

8. Agarwala V, Flannick J, Sunyaev S, Altshuler D. Evaluating empirical bounds on complex disease genetic architecture. Nat Genet. 2013;45(12):1418-27.

9. Speed D, Hemani G, Johnson MR, Balding DJ. Improved heritability estimation from genome-wide SNPs. Am J Hum Genet. 2012;91(6):1011-21.

10. Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI. Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. Am J Hum Genet. 2008;82(1):100-12.

11. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet. 2010;11(6):415-25.

12. Roy-Gagnon MH, Moreau C, Bherer C, St-Onge P, Sinnett D, Laprise C, et al. Genomic and genealogical investigation of the French Canadian founder population structure. Hum Genet. 2011;129(5):521-31.

13. Awadalla P, Boileau C, Payette Y, Idaghdour Y, Goulet JP, Knoppers B, et al. Cohort profile of the CARTaGENE study: Quebec's population-based biobank for public health and personalized genomics. Int J Epidemiol. 2013;42(5):1285-99.

14. Laberge AM, Michaud J, Richter A, Lemyre E, Lambert M, Brais B, et al. Population history and its impact on medical genetics in Quebec. Clin Genet. 2005;68(4):287-301.

15. Bodmer W. Human Genome Project. In: Maloy S, Hughes K, editors. Brenner's Encyclopedia of Genetics (Second Edition). San Diego: Academic Press; 2013. p. 552-4.

16. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977;74(12):5463-7.

17. Shendure J, Mitra RD, Varma C, Church GM. Advanced sequencing technologies: methods and goals. Nature Reviews Genetics. 2004;5(5):335-44.

18. Ronaghi M, Uhlén M, Nyrén P. A sequencing method based on real-time pyrophosphate. Science. 1998;281(5375):363, 5.

19. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, et al. Accurate multiplex polony sequencing of an evolved bacterial genome. Science. 2005;309(5741):1728-32.

20. Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, et al. An integrated semiconductor device enabling non-optical genome sequencing. Nature. 2011;475(7356):348-52.

21. Muzzey D, Evans EA, Lieber C. Understanding the Basics of NGS: From Mechanism to Variant Calling. Curr Genet Med Rep. 2015;3(4):158-65.

22. Bansal V, Boucher C. Sequencing Technologies and Analyses: Where Have We Been and Where Are We Going? iScience. 2019;18:37-41.

23. Teer JK, Mullikin JC. Exome sequencing: the sweet spot before whole genomes. Hum Mol Genet. 2010;19(R2):R145-51.

24. Bashiardes S, Veile R, Helms C, Mardis ER, Bowcock AM, Lovett M. Direct genomic selection. Nat Methods. 2005;2(1):63-9.

25. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, et al. Target-enrichment strategies for next-generation sequencing. Nat Methods. 2010;7(2):111-8.

26. Ulintz PJ, Wu W, Gates CM. Bioinformatics Analysis of Whole Exome Sequencing Data. Methods Mol Biol. 2019;1881:277-318.

27. Petersen B-S, Fredrich B, Hoeppner MP, Ellinghaus D, Franke A. Opportunities and challenges of whole-genome and -exome sequencing. BMC genetics. 2017;18(1):14-.

28. Verlouw JAM, Clemens E, de Vries JH, Zolk O, Verkerk AJMH, am Zehnhoff-Dinnesen A, et al. A comparison of genotyping arrays. European Journal of Human Genetics. 2021;29(11):1611-24.

29. LaFramboise T. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. Nucleic Acids Res. 2009;37(13):4181-93.

30. Ha NT, Freytag S, Bickeboeller H. Coverage and efficiency in current SNP chips. Eur J Hum Genet. 2014;22(9):1124-30.

31. Hotchkiss RD. Models of genetic recombination. Annu Rev Microbiol. 1974;28(0):445-68.

32. Neil A Campbell JBR. Campbell Biologie. 4e. : Pearson ERPI; 2012.

33. Butlin RK. Recombination and speciation. Mol Ecol. 2005;14(9):2621-35.

34. Lichten M, Goldman ASH. MEIOTIC RECOMBINATION HOTSPOTS. 1995;29(1):423-44.

35. Slatkin M. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. Nat Rev Genet. 2008;9(6):477-85.

36. Choi K, Henderson IR. Meiotic recombination hotspots - a comparative view. Plant J. 2015;83(1):52-61.

37. Hedrick PW. Assortative Mating and Linkage Disequilibrium. G3 (Bethesda). 2017;7(1):55-62.

38. Williams KL. Gene Mapping. In: Ranganathan S, Gribskov M, Nakai K, Schönbach C, editors. Encyclopedia of Bioinformatics and Computational Biology. Oxford: Academic Press; 2019. p. 242-50.

39. Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu F, Yang H, et al. The International HapMap Project. Nature. 2003;426(6968):789-96.

40. Altshuler D, Donnelly P, The International HapMap C. A haplotype map of the human genome. Nature. 2005;437(7063):1299-320.

41. International HapMap C, Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, et al. Integrating common and rare genetic variation in diverse human populations. Nature. 2010;467(7311):52-8.

42. Bergström A, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecek P, et al. Insights into human genetic variation and population history from 929 diverse genomes. 2020;367(6484):eaay5012.

43. Nelson MR, Bryc K, King KS, Indap A, Boyko AR, Novembre J, et al. The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. Am J Hum Genet. 2008;83(3):347-58.

44. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74.

45. Armstrong RA. When to use the Bonferroni correction. Ophthalmic Physiol Opt. 2014;34(5):502-8.

46. Chen Z, Boehnke M, Wen X, Mukherjee B. Revisiting the genome-wide significance threshold for common variant GWAS. G3 Genes|Genomes|Genetics. 2021;11(2).

47. Giri P, Mohapatra B. Candidate Gene. In: Vonk J, Shackelford T, editors. Encyclopedia of Animal Cognition and Behavior. Cham: Springer International Publishing; 2017. p. 1-4.

48. Tabor HK, Risch NJ, Myers RM. Candidate-gene approaches for studying complex genetic traits: practical considerations. Nature Reviews Genetics. 2002;3(5):391-7.

49. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nature Genetics. 2007;39(7):906-13.

50. Carlson CS, Eberle MA, Kruglyak L, Nickerson DA. Mapping complex disease loci in whole-genome association studies. Nature. 2004;429(6990):446-52.

51. Wang WYS, Barratt BJ, Clayton DG, Todd JA. Genome-wide association studies: theoretical and practical concerns. Nature Reviews Genetics. 2005;6(2):109-18.

52. Sirugo G, Williams SM, Tishkoff SA. The Missing Diversity in Human Genetic Studies. Cell. 2019;177(1):26-31.

53. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. American journal of human genetics. 2017;101(1):5-22.

54. de Bakker PIW, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D. Efficiency and power in genetic association studies. Nature Genetics. 2005;37(11):1217-23.

55. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007;449(7164):851-61.

56. Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, et al. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. Am J Hum Genet. 2017;100(4):635-49.

57. Bustamante CD, Burchard EG, De la Vega FM. Genomics for the world. Nature. 2011;475(7355):163-5.

58. Popejoy AB, Fullerton SM. Genomics is failing on diversity. Nature. 2016;538(7624):161-4.

59. Lin M, Park DS, Zaitlen NA, Henn BM, Gignoux CR. Admixed Populations Improve Power for Variant Discovery and Portability in Genome-Wide Association Studies. Front Genet. 2021;12:673167-.

60. Rosenberg NA, Huang L, Jewett EM, Szpiech ZA, Jankovic I, Boehnke M. Genome-wide association studies in diverse populations. Nat Rev Genet. 2010;11(5):356-66.

61. Li N, Stephens M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics. 2003;165(4):2213-33.

62. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016;48(10):1284-7.

63. Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. Annu Rev Genomics Hum Genet. 2009;10:387-406.

64. Lefevre T RM, Thomas F. . Biologie évolutive: . Boeck supérieur2016.

65. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nature Reviews Genetics. 2010;11(7):499-511.

66. Howie BN, Donnelly P, Marchini J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. PLOS Genetics. 2009;5(6):e1000529.

67. Deelen P, Bonder MJ, van der Velde KJ, Westra H-J, Winder E, Hendriksen D, et al. Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration. BMC Research Notes. 2014;7(1):901.

68. Roshyara NR, Kirsten H, Horn K, Ahnert P, Scholz M. Impact of pre-imputation SNP-filtering on genotype imputation results. BMC Genet. 2014;15:88.

69. Browning SR, Browning BL. Haplotype phasing: existing methods and new developments. Nature Reviews Genetics. 2011;12(10):703-14.

70. Tewhey R, Bansal V, Torkamani A, Topol EJ, Schork NJ. The importance of phase information for human genomics. Nature Reviews Genetics. 2011;12(3):215-23.

71. Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. American journal of human genetics. 2001;68(4):978-89.

72. Scheet P, Stephens M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. American journal of human genetics. 2006;78(4):629-44.

73. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. American journal of human genetics. 2007;81(5):1084-97.

74. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol. 2010;34(8):816-34.

75. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. Nat Methods. 2011;9(2):179-81.

76. Delaneau O, Marchini J, Zagury J-F. A linear complexity phasing method for thousands of genomes. Nature Methods. 2012;9(2):179-81.

77. Loh P-R, Danecek P, Palamara PF, Fuchsberger C, A Reshef Y, K Finucane H, et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nature genetics. 2016;48(11):1443-8.

78. Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, Clarke R, et al. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat Genet. 2008;40(2):161-9.

79. Browning SR. Multilocus Association Mapping Using Variable-Length Markov Chains. The American Journal of Human Genetics. 2006;78(6):903-13.

80. Fuchsberger C, Abecasis GR, Hinds DA. minimac2: faster genotype imputation. Bioinformatics. 2014;31(5):782-4.

81. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nature Genetics. 2012;44(8):955-9.

82. Howie BN, Donnelly P, Marchini JJPg. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. 2009;5(6):e1000529.

83. Loh PR, Danecek P, Palamara PF, Fuchsberger C, Y AR, H KF, et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet. 2016;48(11):1443-8.

84. Degenhardt F, Wendorff M, Wittig M, Ellinghaus E, Datta LW, Schembri J, et al. Construction and benchmarking of a multi-ethnic reference panel for the imputation of HLA class I and II alleles. Human Molecular Genetics. 2018;28(12):2078-92.

85. The GenomeAsia 100K Project enables genetic discoveries across Asia. Nature. 2019;576(7785):106-11.

86. Vergara C, Parker MM, Franco L, Cho MH, Valencia-Duarte AV, Beaty TH, et al. Genotype imputation performance of three reference panels using African ancestry individuals. Hum Genet. 2018;137(4):281-92.

87. Skipper M. HapMap Phase II unveiled. Nature Reviews Genetics. 2007;8(11):827-.

88. Li Y, Willer C, Sanna S, Abecasis G. Genotype Imputation. 2009;10(1):387-406.

89. Cavalli-Sforza LL. The Human Genome Diversity Project: past, present and future. Nature Reviews Genetics. 2005;6(4):333-40.

90. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet. 2016;48(10):1279-83.

91. Walter K, Min JL, Huang J, Crooks L, Memari Y, McCarthy S, et al. The UK10K project identifies rare variants in health and disease. Nature. 2015;526(7571):82-90.

92. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015;526(7571):75-81.

93. Kowalski MH, Qian H, Hou Z, Rosen JD, Tapia AL, Shan Y, et al. Use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. PLoS Genet. 2019;15(12):e1008500.

94. Mayo O. A century of Hardy-Weinberg equilibrium. Twin Res Hum Genet. 2008;11(3):249-56.

95. Roshyara NR, Horn K, Kirsten H, Ahnert P, Scholz M. Comparing performance of modern genotype imputation methods in different ethnicities. Scientific Reports. 2016;6(1):34386.

96. Wojcik GL, Fuchsberger C, Taliun D, Welch R, Martin AR, Shringarpure S, et al. Imputation-Aware Tag SNP Selection To Improve Power for Large-Scale, Multi-ethnic Association Studies. G3 (Bethesda). 2018;8(10):3255-67.

97. Flores-Espinoza R, Paz-Cruz E, Ruiz-Pozo VA, Lopez-Carrera M, Cabrera-Andrade A, Gusmão L, et al. Investigating genetic diversity in admixed populations from Ecuador. Am J Phys Anthropol. 2021;176(1):109-19.

98. Liu EY, Li M, Wang W, Li Y. MaCH-admix: genotype imputation for admixed populations. Genet Epidemiol. 2013;37(1):25-37.

99. Browning BL, Browning SR. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet. 2009;84(2):210-23.

100. Vlachopoulou E, Lahtela E, Wennerström A, Havulinna AS, Salo P, Perola M, et al. Evaluation of HLA-DRB1 imputation using a Finnish dataset. Tissue Antigens. 2014;83(5):350-5.

101. Benito M, Parker J, Du Q, Wu J, Xiang D, Perou CM, et al. Adjustment of systematic microarray data biases. Bioinformatics. 2004;20(1):105-14.

102. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8(1):118-27.

103. Reese SE, Archer KJ, Therneau TM, Atkinson EJ, Vachon CM, de Andrade M, et al. A new statistic for identifying batch effects in high-throughput genomic data that uses guided principal component analysis. Bioinformatics (Oxford, England). 2013;29(22):2877-83.

104. Okasha S. Population Genetics. The Stanford Encyclopedia of Philosophy (Winter 2016 Edition)2016.

105. Wang CM, Lin CJ, Feng HY. How to teach genetic drift. Yi Chuan. 2020;42(12):1211-20.

106. Lynch M, Ackerman MS, Gout JF, Long H, Sung W, Thomas WK, et al. Genetic drift, selection and the evolution of the mutation rate. Nat Rev Genet. 2016;17(11):704-14.

107. Ohta T. Role of random genetic drift in the evolution of interactive systems. J Mol Evol. 1997;44 Suppl 1:S9-14.

108. Relethford JH, Crawford MH. Genetic drift and the population history of the Irish travellers. Am J Phys Anthropol. 2013;150(2):184-9.

109. Thierry Lefevre MR, Frédéric Thomas. Biologie évolutive: de boeck; 2016.

110. Gazave E, Chang D, Clark AG, Keinan A. Population growth inflates the per-individual number of deleterious mutations and reduces their mean effect. Genetics. 2013;195(3):969-78.

111. Keinan A, Clark AG. Recent explosive human population growth has resulted in an excess of rare genetic variants. Science. 2012;336(6082):740-3.

112. Magurran AE, May RM, Amos W, Harwood J. Factors affecting levels of genetic diversity in natural populations. 1998;353(1366):177-86.

113. E M Wijsman a, Cavalli-Sforza LL. Migration and Genetic Population Structure with Special Reference to Humans. 1984;15(1):279-301.

114. Li W-H. Effect of Migration on Genetic Distance. 1976;110(975):841-7.

115. Gao Z, Waggoner D, Stephens M, Ober C, Przeworski M. An estimate of the average number of recessive lethal mutations carried by humans. Genetics. 2015;199(4):1243-54.

116. Charlesworth D. Effects of inbreeding on the genetic diversity of populations. Philos Trans R Soc Lond B Biol Sci. 2003;358(1434):1051-70.

117. Marchi N, Mennecier P, Georges M, Lafosse S, Hegay T, Dorzhu C, et al. Close inbreeding and low genetic diversity in Inner Asian human populations despite geographical exogamy. Scientific reports. 2018;8(1):9397-.

118. Moreno E, Pérez-González J, Carranza J, Moya-Laraño J. Better Fitness in Captive Cuvier's Gazelle despite Inbreeding Increase: Evidence of Purging? PLoS One. 2015;10(12):e0145111.

119. Booker TR, Jackson BC, Keightley PD. Detecting positive selection in the genome. BMC Biology. 2017;15(1):98.

120. Cvijović I, Good BH, Desai MM. The Effect of Strong Purifying Selection on Genetic Diversity. Genetics. 2018;209(4):1235-78.

121. Charlesworth D. Balancing selection and its effects on sequences in nearby genome regions. PLoS Genet. 2006;2(4):e64.

122. Hedrick PW. What is the evidence for heterozygote advantage selection? Trends Ecol Evol. 2012;27(12):698-704.

123. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, et al. Genes mirror geography within Europe. Nature. 2008;456(7218):98-101.

124. Bherer C, Labuda D, Roy-Gagnon M-H, Houde L, Tremblay M, Vézina H. Admixed ancestry and stratification of Quebec regional populations. 2011;144(3):432-41.

125. Hellwege JN, Keaton JM, Giri A, Gao X, Velez Edwards DR, Edwards TL. Population Stratification in Genetic Association Studies. Curr Protoc Hum Genet. 2017;95:1.22.1-1..3.

126. Menozzi P, Piazza A, Cavalli-Sforza L. Synthetic maps of human gene frequencies in Europeans. Science. 1978;201(4358):786-92.

127. Templeton AR. The theory of speciation via the founder principle. Genetics. 1980;94(4):1011-38.

128. Kere J. Human population genetics: lessons from Finland. Annu Rev Genomics Hum Genet. 2001;2:103-28.

129. Cavalli-Sforza LL, Feldman MW. The application of molecular genetic approaches to the study of human evolution. Nat Genet. 2003;33 Suppl:266-75.

130. McKusick VA. Mendelian Inheritance in Man and its online version, OMIM. Am J Hum Genet. 2007;80(4):588-604.

131. Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI. Shifting Paradigm of Association Studies: Value of Rare Single-Nucleotide Polymorphisms. The American Journal of Human Genetics. 2008;82(1):100-12.

132. Moreau C, Lefebvre JF, Jomphe M, Bhérer C, Ruiz-Linares A, Vézina H, et al. Native American admixture in the Quebec founder population. PLoS One. 2013;8(6):e65507.

133. Peltonen L, Palotie A, Lange K. Use of population isolates for mapping complex traits. Nat Rev Genet. 2000;1(3):182-90.

134. Charbonneau H DB, Légaré J, Denis H The population of the St-Lawrence Valley, 1608–1760. Haines MR SR, editor. Cambridge University Press, New York2000.

135. Vézina H, Tremblay M. Origines et contributions génétiques des fondatrices et des fondateurs de la population québécoise. Cahiers québécois de démographie. 2005;34.

136. Moreau C, Vézina H, Yotova V, Hamon R, de Knijff P, Sinnett D, et al. Genetic heterogeneity in regional populations of Quebec--parental lineages in the Gaspe Peninsula. Am J Phys Anthropol. 2009;139(4):512-22.

137. Moreau C, Lefebvre J-F, Jomphe M, Bhérer C, Ruiz-Linares A, Vézina H, et al. Native American admixture in the Quebec founder population. PLoS One. 2013;8(6):e65507-e.

138. Desjardins M FY, Bélanger J, Hétu B. Histoire de la Gaspésie. Les Presses de l'Université Laval. 1999.

139. B D. Le Registre de la population du Québec ancien1998.

140. Bouchard G VH. Projet BALSAC—Rapport annuel 2008–2009. Université du Québec à Chicoutimi2009.

141. Casals F, Hodgkinson A, Hussin J, Idaghdour Y, Bruat V, de Maillard T, et al. Whole-exome sequencing reveals a rapid change in the frequency of rare functional variants in a founding population of humans. PLoS genetics. 2013;9(9):e1003815-e.

142. Casals F, Hodgkinson A, Hussin J, Idaghdour Y, Bruat V, de Maillard T, et al. Whole-Exome Sequencing Reveals a Rapid Change in the Frequency of Rare Functional Variants in a Founding Population of Humans. PLOS Genetics. 2013;9(9):e1003815.

143. Awadalla P, Boileau C, Payette Y, Idaghdour Y, Goulet J-P, Knoppers B, et al. Cohort profile of the CARTaGENE study: Quebec's population-based biobank for public health and personalized genomics. International Journal of Epidemiology. 2012;42(5):1285-99.

144. Dummer TJB, Awadalla P, Boileau C, Craig C, Fortier I, Goel V, et al. The Canadian Partnership for Tomorrow Project: a pan-Canadian platform for research on chronic disease prevention. Cmaj. 2018;190(23):E710-e7.

145. Dhalla A, McDonald TE, Gallagher RP, Spinelli JJ, Brooks-Wilson AR, Lee TK, et al. Cohort Profile: The British Columbia Generations Project (BCGP). Int J Epidemiol. 2019;48(2):377-8k.

146. Robson PJ, Solbak NM, Haig TR, Whelan HK, Vena JE, Akawung AK, et al. Design, methods and demographics from phase I of Alberta's Tomorrow Project cohort: a prospective cohort profile. CMAJ Open. 2016;4(3):E515-e27.

147. Sweeney E, Cui Y, DeClercq V, Devichand P, Forbes C, Grandy S, et al. Cohort Profile: The Atlantic Partnership for Tomorrow's Health (Atlantic PATH) Study. Int J Epidemiol. 2017;46(6):1762-3i.

148. Chen S, Ghandikota S, Gautam Y, Mersha T. MI-MAAP: marker informativeness for multi-ancestry admixed populations. BMC Bioinformatics. 2020;21.

149. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74.

150. Roy-Gagnon M-H, Moreau C, Bherer C, St-Onge P, Sinnett D, Laprise C, et al. Genomic and genealogical investigation of the French Canadian founder population structure. Human Genetics. 2011;129(5):521-31.

151. Nalls MA, Bras J, Hernandez DG, Keller MF, Majounie E, Renton AE, et al. NeuroX, a fast and efficient genotyping platform for investigation of neurodegenerative diseases. Neurobiol Aging. 2015;36(3):1605.e7-12.

152. Daber R, Sukhadia S, Morrissette JJ. Understanding the limitations of next generation sequencing informatics, an approach to clinical pipeline validation using artificial data sets. Cancer Genet. 2013;206(12):441-8.

153. Human gene mapping, genetic linkage, and clinical applications. Ann Intern Med. 1980;93(3):469-79.

154. Das S, Abecasis GR, Browning BL. Genotype Imputation from Large Reference Panels. Annu Rev Genomics Hum Genet. 2018;19:73-96.

155. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat Rev Genet. 2010;11(7):499-511.

156. Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K, et al. The African Genome Variation Project shapes medical genetics in Africa. Nature. 2015;517(7534):327-32.

157. Mathias RA, Taub MA, Gignoux CR, Fu W, Musharoff S, O'Connor TD, et al. A continuum of admixture in the Western Hemisphere revealed by the African Diaspora genome. Nat Commun. 2016;7:12522.

158. Peltonen L, Jalanko A, Varilo T. Molecular genetics of the Finnish disease heritage. Hum Mol Genet. 1999;8(10):1913-23.

159. Goodman RM. Medical genetic studies of the amish. American Journal of Human Genetics. 1979;31(1):86-7.

160. Wollstein A, Lao O, Becker C, Brauer S, Trent RJ, Nürnberg P, et al. Demographic history of Oceania inferred from genome-wide data. Curr Biol. 2010;20(22):1983-92.

161. Harris DN, Kessler MD, Shetty AC, Weeks DE, Minster RL, Browning S, et al. Evolutionary history of modern Samoans. Proc Natl Acad Sci U S A. 2020;117(17):9458-65.

162. Labuda D, Zietkiewicz E, Labuda M. The genetic clock and the age of the founder effect in growing populations: a lesson from French Canadians and Ashkenazim. Am J Hum Genet. 1997;61(3):768-71.

163. Ostrer H. A genetic profile of contemporary Jewish populations. Nat Rev Genet. 2001;2(11):891-8.

164. Service S, DeYoung J, Karayiorgou M, Roos JL, Pretorious H, Bedoya G, et al. Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. Nat Genet. 2006;38(5):556-60.

165. Charbonneau HJLAN. Les francophones du Québec de 1608 à 1960. 1988;78:220-32.

166. Bouchard G, De Braekeleer MJHsSH. Homogénéité ou diversité? L'histoire de la population du Québec revue à travers ses gènes. 1990.

167. Bherer C, Labuda D, Roy-Gagnon MH, Houde L, Tremblay M, Vézina H. Admixed ancestry and stratification of Quebec regional populations. Am J Phys Anthropol. 2011;144(3):432-41.

168. Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat Genet. 2003;33 Suppl:228-37.

169. de la Chapelle A. Disease gene mapping in isolated human populations: the example of Finland. J Med Genet. 1993;30(10):857-65.

170. Zhao H, Sun Z, Wang J, Huang H, Kocher JP, Wang L. CrossMap: a versatile tool for coordinate conversion between genome assemblies. Bioinformatics. 2014;30(7):1006-7.

171. Nalls MA, Bras J, Hernandez DG, Keller MF, Majounie E, Renton AE, et al. NeuroX, a fast and efficient genotyping platform for investigation of neurodegenerative diseases. Neurobiology of aging. 2015;36(3):1605.e7-.e1.605E12.

172. Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, et al. High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. 2021:2021.02.06.430068.

173. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics (Oxford, England). 2011;27(15):2156-8.

174. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10(2):giab008.

175. Laurie CC, Doheny KF, Mirel DB, Pugh EW, Bierut LJ, Bhangale T, et al. Quality control and quality assurance in genotypic data for genome-wide association studies. Genet Epidemiol. 2010;34(6):591-602.

176. Browning BL, Zhou Y, Browning SR. A One-Penny Imputed Genome from Next-Generation Reference Panels. Am J Hum Genet. 2018;103(3):338-48.

177. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7.

178. Abraham G, Qiu Y, Inouye M. FlashPCA2: principal component analysis of Biobank-scale genotype datasets. Bioinformatics. 2017;33(17):2776-8.

179. Mitt M, Kals M, Pärn K, Gabriel SB, Lander ES, Palotie A, et al. Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. European Journal of Human Genetics. 2017;25(7):869-76.

180. Sariya S, Lee JH, Mayeux R, Vardarajan BN, Reyes-Dumeyer D, Manly JJ, et al. Rare Variants Imputation in Admixed Populations: Comparison Across Reference Panels and Bioinformatics Tools. 2019;10.

181. Kaler AS, Purcell LC. Estimation of a significance threshold for genome-wide association studies. BMC Genomics. 2019;20(1):618.

182. Khor SS, Yang W, Kawashima M, Kamitsuji S, Zheng X, Nishida N, et al. High-accuracy imputation for HLA class I and II genes based on high-resolution SNP data of population-specific references. Pharmacogenomics J. 2015;15(6):530-7.

183. Kim K, Bang SY, Lee HS, Bae SC. Construction and application of a Korean reference panel for imputing classical alleles and amino acids of human leukocyte antigen genes. PLoS One. 2014;9(11):e112546.

184. Pappas DJ, Lizee A, Paunic V, Beutner KR, Motyer A, Vukcevic D, et al. Significant variation between SNP-based HLA imputations in diverse populations: the last mile is the hardest. Pharmacogenomics J. 2018;18(3):367-76.

185. Manzardo AM, Gunewardena S, Wang K, Butler MG. Exon microarray analysis of human dorsolateral prefrontal cortex in alcoholism. Alcohol Clin Exp Res. 2014;38(6):1594-601.

186. Ma X, Shao Y, Tian L, Flasch DA, Mulder HL, Edmonson MN, et al. Analysis of error profiles in deep next-generation sequencing data. Genome Biology. 2019;20(1):50.

187. Shringarpure SS, Mathias RA, Hernandez RD, O'Connor TD, Szpiech ZA, Torres R, et al. Using genotype array data to compare multi- and single-sample variant calls and improve variant call sets from deep coverage whole-genome sequencing data. Bioinformatics. 2017;33(8):1147-53.

188. Stewart CA, Horton R, Allcock RJ, Ashurst JL, Atrazhev AM, Coggill P, et al. Complete MHC haplotype sequencing for common disease gene mapping. Genome Res. 2004;14(6):1176-87.

189. Mungall AJ, Palmer SA, Sims SK, Edwards CA, Ashurst JL, Wilming L, et al. The DNA sequence and analysis of human chromosome 6. Nature. 2003;425(6960):805-11.

190. Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet. 2008;40(6):695-701.

191. Liu CT, Deng X, Fisher V, Heard-Costa N, Xu H, Zhou Y, et al. Revisit Population-based and Family-based Genotype Imputation. Sci Rep. 2019;9(1):1800.

192. Ahmad M, Sinha A, Ghosh S, Kumar V, Davila S, Yajnik CS, et al. Inclusion of Population-specific Reference Panel from India to the 1000 Genomes Phase 3 Panel Improves Imputation Accuracy. Sci Rep. 2017;7(1):6733.

193. Jiménez-Kaufmann A, Chong AY, Cortés A, Quinto-Cortés CD, Fernandez-Valverde SL, Ferreyra-Reyes L, et al. Imputation Performance in Latin American Populations: Improving Rare Variants Representation With the Inclusion of Native American Genomes. Front Genet. 2022;12:719791-.

194. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434-43.

195. Romero A, Carrier PL, Erraqabi A, Sylvain T, Auvolat A, Dejoie E, et al. Diet Networks: Thin Parameters for Fat Genomic. 2016;abs/1611.09340.

196. Wall JD, Tang LF, Zerbe B, Kvale MN, Kwok P-Y, Schaefer C, et al. Estimating genotype error rates from high-coverage next-generation sequence data. Genome Res. 2014;24(11):1734-9.