UQAC
Université du Québec
à Chicoutimi

La génétique des populations à effet fondateur ; un miroir de la démographie et de l'histoire

par Laurence Gagnon

Mémoire présenté à l'Université du Québec à Chicoutimi en vue de l'obtention du grade de Maître ès sciences (M. Sc.) en sciences cliniques et biomédicales

Québec, Canada

Résumé

Les populations à effet fondateur ont été très utiles afin d'identifier des variants liés à des maladies rares, mais également afin de mieux comprendre l'impact des phénomènes démographiques sur la génétique de la population. Nous croyons que l'investigation approfondie de la structure fine présente au sein de ce type de population est cruciale pour l'étude et l'identification de nouveaux variants rares. En effet, une cohorte de plus petite taille, possédant une structure fine, permet de concentrer ce type de variant. Ainsi, leur fréquence est augmentée, ce qui faciliterait l'identification de nouveaux variants. Néanmoins, comprendre d'où provient cette structure aide également à bâtir de meilleures connaissances pour mieux étudier les maladies associées aux populations à effet fondateur. Avec l'aide de données généalogiques, il est possible de suivre la structure de la population québécoise qui est apparue dès 1750 jusqu'à aujourd'hui. De plus, ces mêmes données aident à comprendre l'impact de certains processus démographiques sur l'apparition de ces structures fines au Québec.

Ce mémoire dévoile donc l'utilité d'étudier les populations à effet fondateur et les structures fines de population. Nous croyons que la recherche sur des maladies débute par une bonne compréhension de la population à l'étude. Autant pour une population à effet fondateur ou non, cela commence par une investigation de la structure fine de cette population afin de tirer profit de cette structure unique. De plus, au Québec, nous sommes très choyés d'avoir accès à des données généalogiques complètes sur plus de 400 ans qui révèlent les processus démographiques vécus par la population, mais aussi pour suivre la transmission et comprendre la répartition des variants génétiques et leur impact sur la santé populationnelle.

Abstract

The populations with a founder effect have been extremely useful in identifying variants associated with rare diseases, as well as in better understanding the impact of demographic phenomena on population genetics. We believe that thorough investigation into the fine structure within this type of populations is crucial for studying and identifying new rare variants. Indeed, a smaller cohort with a fine structure allows the concentration of this type of variant. Thus, this increases their frequency, which would facilitate the identification of new variants. However, understanding the origin of this structure also contributes to build better knowledge for studying diseases associated with populations exhibiting a founder effect. With the help of genealogical data, it is possible to track the structure of the Quebec population from as early as 1750 to the present day. Furthermore, this same data helps understand the impact of certain demographic processes on the appearance of these fine structures in Quebec.

This thesis thus reveals the utility of studying populations with a founder effect and population fine structures. We believe that research on diseases begins with a thorough comprehension of the population under study. Whether for a population with a founder effect or not, this begins with an investigation into the fine structure of the population to take advantage of this unique structure. Moreover, in Quebec, we are privileged to have access to comprehensive genealogical data spanning over 400 years, revealing the demographic processes experienced by the population, as well as for tracking transmission and understanding the distribution of genetic variants and their impact on populational health.

Table des matières

Liste des figures

Liste des abréviations

ADN…….….. Acide désoxyribonucléique

CEU……..…Utah Residents with Northern and Western European ancestry

CHB……..…….….Han Chinese in Bejing

GBR…….….British in England and Scotland

GAC…….….Gaspé Acadians

GC……..……..…….....Genetic contribution

GCI…….….Gaspé Channel Islanders

GFC……...…….….Gaspé French Canadians

GLO…….…….….Gaspé Loyalists

GWAS……..……Étude d'association génétique à l'échelle du génome

IBD……Segment identique-par-descendance (Identical-by-descent)

INDEL……..……..…..Petites insertions et de délétions

JPT…….……..…….….Japanese in Tokyo

MAF…….…….….Minor allele frequency

MDS…….….Analyse de positionnement multidimensionnel (Multidimensional scaling)

MRCA……...……Ancêtre commun le plus récent (Most recent common ancestors)

MSL……..…….….Mende in Sierra Leone

MTL……...…….….Montreal

NSH……...…….….North Shore

PCA…..……..Analyse par composante principale (Principal component analysis)

PFE……...……Population that had undergone a founder effect

QUE……..…….….Quebec City

ROH…….……Segment d'homozygotie (Runs of homozygosity)

SAG…….…….…. Saguenay-Lac-Saint-Jean

SNP…….Polymorphismes d'un seul nucléotide

UMAP…..Analyse d'approximation et projection uniforme de variétés (Uniform Manifold Approximation and Projection)

Remerciements

Je tiens d'abord à remercier mon directeur de recherche, Simon Girard, pour son accueil au sein de du laboratoire Genopop à l'été 2020. Jamais je n'aurais pu penser que ce stage d'été en pleine pandémie m'aurait apporté là où que je suis aujourd'hui. Sa bienveillance, générosité et surtout sa confiance en moi m'ont accompagné tout au long de ma maitrise.

Je veux également remercier Claudia Moreau, la petite voix dans mon ordinateur, pour ses réponses à toutes mes questions et son support inestimable lors de mes années au sein du laboratoire. Toutes tes remises en question, interrogations et changements d'idée m'ont permis de me dépasser, mais surtout d'apprendre. Je souhaite remercier également tous mes collègues du laboratoire Genopop qui m'ont accompagné et conseillé lors de mes années au sein de ce laboratoire. Merci également à Catherine Laprise et Hélène Vézina pour leurs supports et conseils précieux.

Je tiens également à remercier les organismes suivants pour leur aide financière lors de mon parcours à la maitrise : les Instituts de recherche en santé du Canada, les fonds de recherche du Québec en santé, la corporation de recherche et d'action sur les maladies héréditaires, le centre intersectoriel en santé durable ainsi que la fondation de l'UQAC.

Un merci tout particulier à ma sœur, mon père et ma mère de m'avoir encouragé et supporté tout au long de ma maitrise et à réaliser mes rêves ; ainsi qu'à Rex et Agathe pour leur support émotionnel lors de la rédaction de ce mémoire. Merci également à belle-famille pour leur encouragement et leur appui tout au long de mes études. Merci à mon copain qui réussit toujours à me motiver, m'encourager et surtout d'être si intéressé par ce que je fais. Merci d'être mon fan #1.

Finalement, un grand merci à tous les participants des cohortes utilisées pour ce projet, sans qui rien de cela n'aurait été possible.

Avant-propos

Ce présent mémoire est divisé en cinq sections. La première est l'introduction qui a pour but de mettre en place les éléments théoriques nécessaires à la bonne compréhension de ce mémoire. Cette section commence par une description générale du domaine de la génétique et de la génétique des populations. Ensuite, une section sur les effets fondateurs et leur implication en génétique sont exposées. Cela est suivi par une présentation des analyses génétiques et généalogiques utilisées et une présentation des populations à effet fondateur étudiées dans ce mémoire. Finalement, l'introduction se termine par l'énoncé des objectifs de ce mémoire.

Les deux sections suivantes correspondent à la présentation du corps du travail. Le chapitre 1 présente un premier article scientifique soumis dans le *European Journal of Human Genetics* et qui vise à décrire correctement la structure génétique fine d'une population afin d'affiner les modèles d'analyse des associations génétiques. La structure fine d'une population permet de concentrer des variants rares en augmentant leur fréquence, et ce avec une cohorte comportant moins d'individus. Par la suite, le chapitre 2 présente le second article scientifique réalisé dans le cadre de cette maitrise et qui a été publié en avril 2023 dans le *European Journal of Human Genetics*. Maintenant que l'importance de la structure génétique fine de populations est bien comprise, ce second chapitre veut explorer plus en profondeur cette structure au sein de la population québécoise. Avec l'aide de données généalogiques, il a été possible de retracer l'évolution de la structure du Québec à travers le temps et de déterminer quels sont les phénomènes démographiques qui ont mené à la différenciation génétique de deux sous-populations du Québec.

La quatrième section correspond à une discussion approfondie de chacun de ces deux articles, mais également des objectifs communs à ces deux projets. La discussion présente aussi les énoncés novateurs, les limites ainsi que les perspectives.

Finalement, la dernière section correspond à la conclusion qui permet d'effectuer un retour global sur l'ensemble de ce mémoire.

Introduction

1.1 Génétique

La santé humaine est modulée par plusieurs facteurs. L'alimentation, l'exercice physique, le stress et les infections sont tous des déterminants qui participent à notre bien-être et à notre santé. À l'inverse, ce sont aussi des facteurs qui peuvent également impacter notre santé de façon négative. Au-delà de ces facteurs, qui sont majoritairement modulables ou qui sont des habitudes de vie, se retrouve la génétique. La génétique se décrit comme étant l'étude des gènes et de l'hérédité, donc ce qui nous est transmis par nos parents[1]. Cette transmission de caractères d'un ancêtre à son descendant se fait par l'arrangement de quatre nucléotides A, T, C et G qui composent notre acide désoxyribonucléique (ADN)[1,2]. Ces quatre lettres sont à la base de notre information génétique, et c'est cette information qui dicte les traits, comme la grandeur, la couleur des cheveux ou la couleur des yeux d'un individu. Cependant, c'est dans ce même code génétique qu'une partie du risque, ou de la susceptibilité d'avoir une maladie, est établi. Ainsi, la génétique est un domaine essentiel à étudier afin de mieux comprendre ces facteurs non modifiables qui composent chaque être humain.

1.1.1 Hérédité

L'hérédité est la transmission de caractères d'un être vivant à la prochaine génération par l'intermédiaire de l'ADN et cela se produit lors de la reproduction. Ces caractères observables sont appelés phénotype. De plus, l'ADN est transmis d'un parent à son enfant par méiose. Ainsi, chaque individu est composé de 50% du génome de sa mère et de 50% du génome de son père; et chaque position du génome (pour les autosomes) se retrouve en deux copies appelée allèles[2]. Cela apporte la notion d'homozygotie, où les 2 allèles présents sont identiques, et celle d'hétérozygotie, où les 2 allèles présents sont différents.

La transmission d'une partie du génome d'un parent a son enfant, n'est pas simplement la moitié de son génome. Il se produit un mélange du matériel génétique par un phénomène nommé recombinaison. Ce processus d'échange de matériel génétique entre deux chromosomes se produit lors de la méiose[2]. Ainsi, l'ADN contenu dans une cellule germinale est un mélange des deux copies de l'ADN parental ce qui favorise la diversité allélique au sein des populations.

1.1.2 Variations du génome

Le génome est identique à 99.9% entre tous les humains[3]. Donc, il y a seulement 0.1% de variation entre les individus au niveau de leur génome et c'est là que réside l'ensemble des types de variations génétiques qui apportent nos différences. Ces variations peuvent être de petite taille, ou excessivement grandes[4]. De plus, ces variations peuvent être provoquées par différentes causes. On y retrouve entre autres des erreurs lors de la réplication de l'ADN, des agents externes tels que les virus, des agents chimiques mutagènes ou des mutations *de novo*[4].

D'abord, les polymorphismes d'un seul nucléotide, ou SNP sont la forme la plus fréquente et simple de variation dans le génome humain (Figure 1A)[5]. Ce type de variation est la modification d'une seule base par une autre[4]. Un SNP peut être synonyme si le changement de la séquence nucléotidique ne change pas la nature de l'acide aminé et n'altère donc pas la protéine; non synonyme si le changement de la séquence nucléotidique modifie la nature de l'acide aminé ou non-sens si le changement de la séquence nucléotidique provoque l'apparition d'un codon stop et ainsi l'arrêt de la synthèse de la protéine. Ces variations peuvent influencer l'activité des promoteurs, l'expression des gènes, la conformation de l'ARN messager et plusieurs autres mécanismes de la cellule[5]. Il peut également y avoir des variations qui sont de courtes insertions et de délétions de nucléotides, nommée INDEL (Figure 1A et C)[6]. Les INDEL peuvent être de 1 à 50 paires de base[7]. Un INDEL peut causer des mutations qui

changent le cadre de lecture de l'ADN ou non tout dépendant si sa longueur est un multiple de 3 nucléotides[8,9]. Par la suite, il y a les répétitions en tandem qui sont la répétition d'un motif nucléotidique (Figure 1B)[4,10]. Les répétitions en tandem sont hautement polymorphiques entre des individus non reliés et ont un taux de mutation très élevé causé par un glissement de la polymérase lors de la réplication[10,11]. Finalement, les variants structurels sont des réarrangements de grands segments d'ADN d'au moins 50 nucléotides qui peuvent avoir des conséquences importantes par la modification de grandes sections du génome (Figure 1D)[7]. Ce type de variant peut prendre plusieurs formes, comme les translocations, les inversions, les insertions et les variations du nombre de copies (duplication et délétion). Pour toutes ces variations, une fois acquises elles sont permanentes et transmissibles à la descendance d'un individu.

Un génome typique varie de 4.1 à 5 millions de sites comparativement au génome de référence humain[12]. Ces différences sont à plus de 99.9% des SNP et des INDEL courts[12]. Parmi toutes ces variations, il existe des variants qui sont dits rares, avec une fréquence de moins de 0.5% dans le génome, et d'autres dits communs, avec une fréquence de plus de 0.5%[13]. Dans un génome typique, il y a de 40 000 à 200 000 variants rares et ils ont un plus grand impact au niveau du phénotype que les variants communs[13].

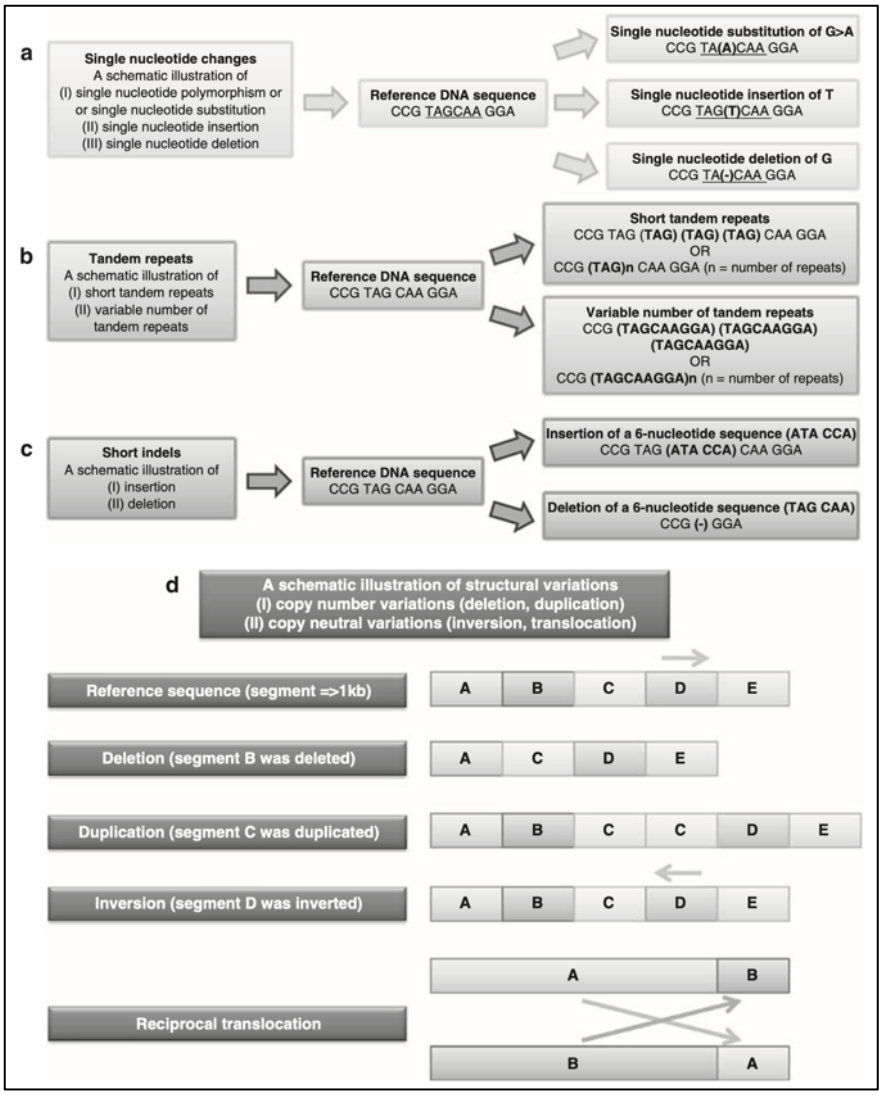


Figure 1. Schéma des différentes variations du génome.

A. Changement d'un seul nucléotide. **B.** Répétition en tandem. **C .** Petit INDEL. **D.** Variants structurels. Tiré de (avec autorisation) : Ku, C. S., Loy, E. Y., Salim, A., Pawitan, Y. & Chia, K. S. The discovery of human genetic variations and their use as disease markers: past, present and future. *J Hum Genet* **55**, 403–415 (2010).

1.1.3 Mutation et maladie

Ces variations sont l'ensemble des éléments qui entrainent des différences génétiques entre les êtres humains. La plupart de ces variations expliquent la diversité humaine et pourquoi nous sommes tous différents les uns des autres au niveau de nos caractères observables. Cependant,

ce sont également ces variations, ou mutations, qui apportent le risque génétique de développer une maladie. Une mutation est causée par une erreur durant le processus de réplication de l'ADN ou pendant la méiose. Les mutations sont étudiées depuis de nombreuses années. Par exemple, la première maladie génétique cartographiée à l'aide de polymorphisme génétique a été découverte en 1983 avec un marqueur génétique sur le chromosome 4 lié à la maladie de Huntington[14]. Depuis, le domaine de la génétique a évolué à grande vitesse et de nombreuses associations avec différents phénotypes sont faites toutes les années.

Une maladie peut être classée en deux catégories. D'un côté, il y a les maladies mendéliennes qui affectent ~7% de la population mondiale et serait associé à un seul gène défectueux[15]. De plus, ce type de maladie est transmis selon un mode de ségrégation précis. Une maladie rare peut être dominante si seulement un allèle est nécessaire pour qu'un individu soit atteint; ou à l'inverse, être récessive si les deux allèles doivent être présents pour que l'individu soit atteint de la maladie[1]. D'un autre côté, il y a les maladies complexes qui sont causées par une combinaison de plusieurs facteurs, tels que la génétique, l'environnement et les habitudes de vie[16]. L'ensemble des facteurs génétiques et non-génétiques causant ces maladies ne sont pas encore très bien connus, et ce n'est pas parce qu'un individu présente des facteurs de risque qu'il va nécessairement développer une maladie[16].

Afin d'étudier les maladies génétiques, l'ADN d'un seul individu atteint n'est pas suffisant. Il faut étudier les variations entre le génome de plusieurs individus atteints, mais également entre le génome d'individus atteints et sains. Cette comparaison permet d'évaluer quelles variations sont associées à une maladie. Toutefois, cette approche comparative est également forte utile afin d'étudier les origines et l'histoire démographique des populations. Selon l'origine d'un individu, celui-ci n'aura pas les mêmes variations qu'un autre d'origine différente. En effet, un individu d'origine africaine ou d'une population récemment métissée a hérité plus de variations

qu'un individu d'origine européenne en raison d'une plus grande diversité allélique[12]. Cette diversité est causée d'un côté par l'origine très ancienne des Africains ce qui a permis d'accumuler des mutations sur une plus grande période; et d'un autre côté les populations métissées sont la somme des variants des populations d'origines différentes. En plus des mutations, les phénomènes migratoires qui créent le mélange entre différentes populations sont aussi une cause de variation génétique[1]. Ainsi, il est important de bien comprendre et distinguer les populations puisque leur fondement génétique n'est pas le même dû à des variations dans les fréquences alléliques à travers le génome[17].

1.1.4 Génétique des populations humaines

La génétique des populations humaines est un sujet d'étude très important en raison de son implication au niveau de la compréhension de maladies génétiques. Cette discipline est la science des variations génétiques et des facteurs qui les influencent, autant les variations génétiques du passé, du présent et que du futur[1]. Elle veut étudier comment les variations génétiques sont distribuées au sein et entre les populations afin de comprendre l'influence des migrations, des phénomènes évolutifs ainsi que la distribution des mutations. Le sujet d'étude est les populations, c'est-à-dire les individus qui se reproduisent au sein d'un groupe d'individus d'une même espèce en un même lieu. La façon dont les populations se différencient est liée à des variations génétiques qui elles-mêmes peuvent résulter de plusieurs facteurs, comme la sélection naturelle. Ainsi, ce domaine est intimement lié à l'évolution puisque la sélection naturelle est un facteur important menant au changement de la génétique d'une population, et donc à son évolution[18].

1.1.4.1 Les fondements de la génétique des populations

La génétique des populations a pour but d'étudier les fréquences alléliques et les génotypes dans un ensemble populationnel et non au niveau individuel. Plus précisément, cette discipline

s'intéresse aux origines, à la distribution dans l'espace et dans le temps ainsi qu'aux forces évolutives qui dictent les variations au sein d'une population[2]. Il y a trois fondements en génétique des populations. Ces fondements sont reliés à des affirmations simples qui dictent les bases de cette discipline[2]. Ces fondements sont : l'ADN peut être répliqué, l'ADN peut muter et se recombiner ainsi que l'ADN et l'environnement interagissent afin de produire les phénotypes.

Le premier fondement est que l'ADN peut être répliqué[2]. Cette propriété permet aux fragments d'ADN et aux mutations d'être copiés, et ainsi d'être transmis aux générations suivantes. Les mutations sont donc présentes dans le temps et dans l'espace. En effet, en génétique des populations, une mutation ne peut pas être étudiée au sein d'un seul individu. La transmission de cette mutation à plusieurs individus, et ce sur plusieurs générations est à la base de cette discipline[2]. De cette manière, il est possible de définir une population comme étant un ensemble d'individus à un temps donné qui occupe un espace donné. Une population qui se reproduit est l'unité de base en génétique des populations en raison de la continuité spatiale et temporelle qui est nécessaire à l'évolution[2]. Il est possible d'étudier une population selon les fréquences alléliques et les génotypes. Pour une mutation X possédant deux états alléliques, C et T, les différents génotypes seront CC, CT et TT et il est possible de calculer la fréquence de ces allèles et de ces génotypes avec la loi de Hardy-Weinberg, $p^2 + 2pq + q^2 = 1.0$ et $q = 1 - p$, où p et q représentent la fréquence des deux allèles dans la population[1,19] (Figure 2). De plus, afin d'étudier les molécules d'ADN dans leur ensemble au sein d'une population, il est question de bassin génétique. Ce sont les fréquences alléliques qui caractérisent le bassin génétique d'un population[2].

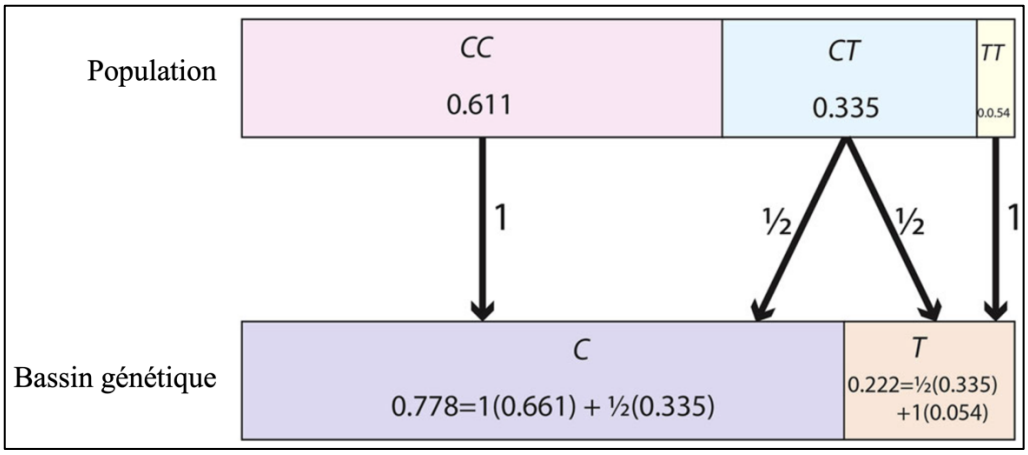


Figure 2. Relation entre fréquence des génotypes et fréquence allélique

Tiré et traduit de (avec autorisation) : Templeton, A. R. *Human population genetics and genomics*. (Academic Press, London, 2019).

Si la réplication de l'ADN fonctionnait de façon exacte, il n'y aurait pas d'évolution[2]. En effet, le mécanisme de réplication de l'ADN peut commettre des erreurs et introduire des mutations, ou variations. Ces mutations sont le second fondement en génétique des populations. Elles peuvent être sans effet, avoir un effet négatif en apportant des mutations liées à des maladies, et aussi avoir un effet positif en procurant des traits favorables à l'adaptation d'une espèce et à son évolution. La génétique des populations traite de tous les types de mutations et leur destin dans le temps et l'espace[2]. Un autre mécanisme peut également provoquer des variations génétiques et des variations des fréquences alléliques, et c'est la recombinaison. La recombinaison crée de nouvelles combinaisons avec les allèles préexistants et augmente la diversité génétique entre des individus. Ainsi, les mutations et la recombinaison créent la diversité génétique qui est à la base des changements évolutifs[2].

Le dernier fondement porte sur l'interaction entre l'environnement et l'expression d'un phénotype[2]. En effet, l'environnement peut fournir un certain contexte permettant d'exprimer certaines informations de l'ADN. Ainsi, cette interaction gène-environnement est un facteur

essentiel qui module le passage de l'ADN vers le phénotype. Il est presque impossible de séparer la génétique de l'environnement et vice-versa, car la définition même d'un phénotype est associée avec l'interaction entre l'ADN et l'environnement.

Pour finir, l'élément intégrateur de ces trois fondements est l'évolution[2]. L'évolution explore le destin des gènes à travers le temps et l'espace et comment mène à ce les populations changent et se modifie à travers les générations. L'évolution est étroitement liée au concept de *fitness*, qui est caractérisé comme un ensemble de phénotypes impliqués dans la viabilité, la réussite de l'accouplement et la fécondité (Figure 3). En génétique des populations, ce *fitness* correspond à la mesure du nombre de gamètes transmise à la prochaine génération. Dans une population, il peut y avoir des génotypes qui s'expriment de façon différente selon l'interaction avec l'environnement afin de créer un phénotype précis. La réplication de l'ADN s'effectue alors selon une reproduction différentielle entre les individus. Cela crée des individus possédant une combinaison de gènes qui ont tendance à être associés avec une meilleure adaptation et survie, et donc qui se répliquent plus[2]. De ce fait, les individus les mieux adaptés ont tendance à plus se reproduire entre eux et donc de créer une descendance mieux adaptée également. C'est ainsi que le concept de *fitness*, qui est à la base de la sélection naturelle, réunit les 3 fondements de base en génétique des populations. De cette manière, la façon dont les individus réagissent à l'environnement influence la diffusion des gènes dans l'espace et le temps.

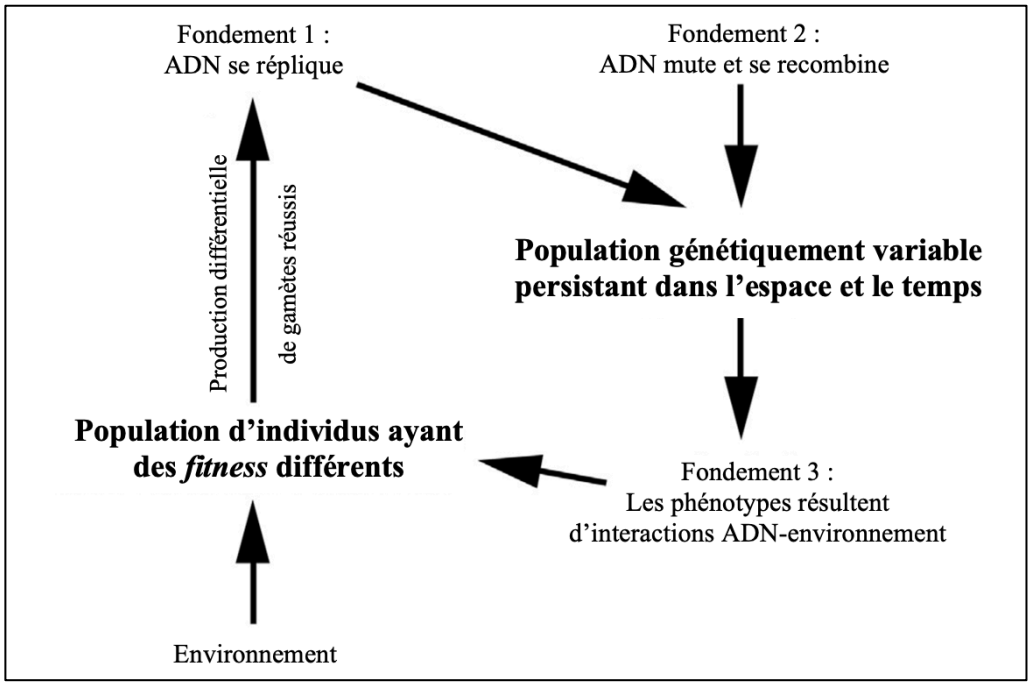


Figure 3. L'intégration des trois fondements de la génétique des populations à travers le concept de fitness

Tiré et traduit de (avec autorisation) : Templeton, A. R. *Human population genetics and genomics*. (Academic Press, London, 2019).

1.1.4.2 Concepts en génétique des populations

Il y a plusieurs concepts importants à maitriser afin de bien comprendre la génétique des populations. Tout d'abord, avec les générations, il est possible que des individus avec un apparentement éloigné se reproduisent ensemble; et ce, sans même savoir qu'ils sont liés. Le résultat lorsque deux êtres apparentés se reproduisent est la consanguinité[2]. Des individus sont dits apparentés lorsqu'ils partagent au moins un ancêtre en commun. Ces deux individus peuvent transmettre à leurs enfants un segment d'ADN identique qui provient d'un même ancêtre à une position donnée sur le génome, nommé segment identique-par-descendance (IBD pour *identical-by-descent*). La longueur de ses segments IBD est reliée à la proximité de l'ancêtre commun. Ces segments identiques peuvent mener à une augmentation des segments homozygotes (ROH pour *runs of homozygosity*) chez leurs descendants. Ces segments ROH sont la façon d'évaluer le niveau de consanguinité chez un individu[2]. Ces copies identiques au

même endroit sur les deux chromosomes d'un individu peuvent être sans conséquence. Cependant, si des allèles mutés ont été transmis par les deux parents, la descendance a plus de risque de développer des maladies récessives dû à cette augmentation d'homozygotie. Il y a donc une apparition plus fréquente de génotypes homozygotes récessifs. Cela mène à la dépression consanguine, où il y a une diminution de la condition physique et de la vitalité liée à l'augmentation du taux de consanguinité[2,20]. Chez les humains, ce concept est moins important, car les relations entre individus très apparentés sont rares. Toutefois, cela est surtout représenté par une diminution du fitness[20].

Par la suite, selon la loi de Hardy-Weinberg, la fréquence allélique ne change pas d'une génération à une autre. Cependant, cela est seulement vrai pour une population de taille infinie[1]. En réalité, les populations ont des tailles finies et les fréquences alléliques peuvent changer de façon aléatoire lorsque les gamètes sont issus du bassin génétique afin de former la génération suivante[1]. Ce changement dans les fréquences alléliques dû au hasard et donc impossible à prévoir est appelé dérive génétique aléatoire[1,2,21]. Dans de grandes populations, la dérive génétique est une force plutôt faible. Toutefois, dans des populations plus petites et isolées, les effets de la dérive génétique sont plus importants. Cela peut mener à des variations dans les fréquences alléliques, la fixation ou la perte d'un allèle. Ce phénomène est une force évolutive caractérisée par des fluctuations inévitables des fréquences alléliques et qui se déroule dans toutes les populations finies à n'importe quel locus[2]. L'avenir de la prochaine génération dépend seulement du bassin génétique actuel, et jamais du bassin génétique passé, sans avoir aucune tendance à inverser ou restaurer les fréquences alléliques à un niveau d'une génération antérieure[2]. Dans une population finie, il est possible de calculer différentes possibilités de résultats attendus de la dérive génétique, mais il est impossible de prédire lequel de ces résultats va réellement se produire[1]. Les calculs de dérive génétique prennent en compte des hypothèses comme : une population idéale, avec autant de femmes que d'hommes, de taille constante avec

accouplement aléatoire de tous les individus avec le même nombre d'enfants pour chaque individu sans contact avec d'autres populations[2,22]. Nous savons que dans la vie réelle il est impossible d'avoir des populations idéales. Ainsi, la taille effective d'une population est un concept qui veut trouver la taille d'une population idéale qui a la même force de dérive génétique que la population réelle non-idéale[2,23]. Tout ce qui augmente la variation entre les individus au niveau de leur succès reproducteur va réduire la taille effective de la population[22].

D'un autre côté, la réplication de l'ADN a une existence dans le temps, à travers les générations, mais aussi dans l'espace, à travers différentes populations[6]. Les populations sont plus connectées qu'auparavant avec la facilité de l'accès à des moyens de transport, avec des frontières de plus en plus floues et ce avec une augmentation des processus migratoires. Le flux génétique est la propagation des gènes à travers l'espace par l'accouplement d'individus provenant de différentes populations ou de bassins génétiques différents[2,21]. Autrement dit, c'est un transfert d'allèles entre populations. Ce flux génétique peut mener en sous-divisions de populations avec des bassins génétiques distincts et des fréquences alléliques différentes[2]. C'est un mécanisme évolutif essentiel puisque le flux génétique peut minimiser les effets de la consanguinité et de la dérive génétique[21]. Un autre phénomène peut se produire lorsqu'une population entière s'établit à un nouvel endroit déjà peuplé. Cette population va s'accoupler avec les individus de la population présente et va former du métissage. Le métissage est le mélange entre ces populations et est la forme majeure de flux génétique déterminant la structure de population chez les humains[2].

La dérive et le flux génétiques sont deux forces évolutives importantes et opposées qui affectent la distribution des fréquences alléliques dans une population. Tous les deux impactent la diversité allélique d'une population, mais leurs conséquences sont très différentes. La dérive génétique est une force aléatoire qui tend à réduire la diversité, car elle peut conduire à la

fixation ou à la perte d'allèles, tandis que le flux génétique augmente la diversité en introduisant de nouveaux allèles au sein d'une population.

1.2 Effet fondateur

1.2.1 Définition

L'effet fondateur se produit lorsqu'une population est créée à partir d'un petit nombre d'individus provenant d'une population mère de taille plus importante. Ce petit nombre crée un goulot d'étranglement ce qui a pour effet de diminuer le bassin des allèles qui créeront la prochaine génération. Un goulot d'étranglement est défini comme étant une réduction rapide et souvent drastique de la taille de la population sur une ou plusieurs générations consécutives[1,24]. Tous ces phénomènes mènent à des variations dans les fréquences alléliques chez la population à effet fondateur, comparativement à la population qui l'a créée. En effet, les individus de la nouvelle population portent seulement une fraction pas forcément représentative du bassin génétique de la population d'origine[25,26]. De plus, la dérive génétique peut jouer un rôle important en augmentant, diminuant ou fixant la fréquence de certains allèles de façon aléatoire[26–28]. Plus le nombre de fondateurs est petit, plus il va y avoir un effet important de la dérive génétique sur cette population[29]. La dérive génétique au cours des premières générations peut être augmentée en raison des différences de succès reproducteur entre les fondateurs[30]. La dérive génétique nuit également à la sélection naturelle et à son pouvoir d'éliminer des variants délétères[28]. Ce goulot d'étranglement peut être suivi d'un isolement culturel, géographique ou autre qui éloigne cette population pour plusieurs générations[31]. Suite à cela, il peut y avoir l'expansion de la population à effet fondateur qui mène à la présence nombreuse de liens d'apparentement éloignés entre les individus d'une population[32]. Il y a donc la possibilité d'avoir des mariages entre des individus avec un apparentement éloigné menant ainsi à une descendance avec un taux de consanguinité plus élevé[1].

Les effets fondateurs ont une implication immense dans la formation de la diversité allélique à travers le monde[28]. En effet, l'espèce humaine s'est dispersée au cours des derniers 50 000 à 100 000 ans avec plusieurs périodes de goulots d'étranglement et de mélanges entre les populations[28,33]. Ainsi, les différentes populations actuelles proviennent de nombreux évènements successifs d'effets fondateurs qui ont aux risques de développer des maladies ainsi qu'à l'évolution de l'espèce et à son adaptation.

1.2.2 Impacts et conséquences

Un effet fondateur aura des impacts et des conséquences sur la génétique de ces populations comme par l'apparition d'une structure génétique. Cette structure différencie la population à effet fondateur de la population mère. Le changement dans les fréquences alléliques apporte de l'hétérogénéité génétique entre les individus de ces deux populations et leur différenciation. Ainsi, la population est distinguable au niveau génétique, et non seulement en raison de ses traits et phénotypes[17].

Dans les populations récentes ayant vécu un effet fondateur, il y a souvent des maladies qui sont beaucoup plus fréquentes, et à l'inverse il y a des maladies qui sont presque inexistantes. L'étude des populations à effet fondateur a donc une contribution disproportionnée dans la découverte de variants liés à des maladies rares[34]. Ces maladies peuvent être introduites par seulement un ancêtre, mais qui sont devenues très importantes en raison du basin limité de fondateurs, d'un fort accroissement naturel et de la dérive génétique. Les variants rares qui ont augmenté en fréquence peuvent causer des maladies mendéliennes avec des mutations qui n'étaient pas encore connues, ou qui sont moins fréquentes dans les populations sans effet fondateur[35]. Ces maladies plus fréquentes engendrées par ces fluctuations dans les fréquences alléliques ont un impact majeur pour la santé publique. Effectivement, plusieurs populations à effet fondateur ont accès à des services de dépistages génétiques pour les maladies qui y sont

répandues[27,36,37]. Ces services sont essentiels afin de soutenir la population lors de la préconception pour diminuer le risque d'avoir une maladie génétique souvent incurable et sévère.

De plus, la notion d'endogamie, où les mariages ont lieu entre individus d'un même groupe, peut également avoir un impact important, surtout à la suite d'un effet fondateur. En effet, ce type de population est souvent isolé ce qui mène à une augmentation de l'endogamie. Ainsi, dans une société endogame, en quelques générations, la plupart des individus seront reliés les uns aux autres, et ce sur plusieurs lignées généalogiques[38]. De ce fait, cela favorise la présence des mêmes ancêtres dans la généalogie d'un individu. Ces mêmes ancêtres, qui sont présents plusieurs fois, ont plus de chance de léguer des segments génétiques identiques ce qui augmente le risque d'avoir des segments génétiques homozygotes. Ces segments ROH peuvent donc augmenter la présence de variants génétiques récessifs causant des maladies[38].

1.2.3 Phénomènes démographiques qui influencent un effet fondateur

Les études démographiques nous renseignent sur la dynamique des populations. Dans un contexte de population à effet fondateur, ces phénomènes peuvent avoir un impact important sur le futur de la population et sa structure génétique.

Tout d'abord, la dynamique migratoire d'une population est un des deux facteurs essentiels dans le processus de peuplement. Les mouvements migratoires sont un facteur qui exerce une influence importante sur la structure génétique d'un population[26]. Tous les types de mouvement, comme l'immigration, l'émigration ou des déplacements internes, vont se refléter sur cette structure[39]. Cette relation entre les migrations et la génétique est étroite et de nombreux facteurs influencent le résultat, tel que le nombre de personnes contribuant au peuplement initial, le rythme de leur arrivé, les lieux de provenance, le caractère familial, ou non, de

l'immigration ou si l'établissement est permanent ou non[26]. L'immigration d'un très grand nombre de personnes avec des origines variées n'aura pas le même impact sur la structure génétique d'une nouvelle région colonisée par un nombre d'individus limités provenant du même endroit, pourvu qu'il y ait une mixité dans la reproduction. Le second facteur essentiel dans le processus de peuplement est la reproduction. La reproduction indique à quel point une population va grandir et dicte son futur. L'accroissement naturel est la différence entre le nombre de naissances et de décès[40]. Afin qu'une population grandisse, il faut que cette valeur soit positive, donc qu'il y ait un nombre de reproduction supérieur au décès. Toutefois, c'est la fécondité utile qui est la plus importante comparativement au nombre d'enfants. La fécondité utile est le nombre d'enfants qui vont survivre jusqu'à l'âge adulte et qui vont se reproduire à leur tour[41]. De plus, le taux de l'accroissement naturel est important. Un accroissement rapide de la population peut avoir un impact important sur l'augmentation de variations rares en augmentant leur fréquence et en ne laissant pas le temps à la sélection naturelle d'agir[42,43].

Ces facteurs démographiques sont essentiels pour bien comprendre une population au niveau génétique. Dans le contexte d'une population à effet fondateur, la présence de processus migratoires et la force de l'accroissement naturel exercent une influence importante sur l'impact de l'effet fondateur sur la structure génétique de la population.

1.3 Analyses utilisées

Plusieurs analyses génétiques et généalogiques ont été effectuées dans ce mémoire. Cette section veut expliquer ces analyses utilisées fréquemment en génétique des populations.

1.3.1 Analyses génétiques

Un type d'étude important en génétique des populations se base sur des méthodes de réduction de données de grande dimension afin de visualiser la structure d'une population. L'analyse par

composante principale, ou PCA (pour *principal component analysis*) est une de ces méthodes. C'est un algorithme linéaire qui recherche de façon itérative des axes orthogonaux où des objets projetés ont la variance la plus élevée et renvoie la position de ces objets (Figure 4A)[44]. Cette méthode identifie donc le ou les éléments qui expliquent le plus la variabilité des données[45]. La visualisation en 2D ou en 3D de ces composantes principales permet de voir des motifs spécifiques à chaque ensemble de données. Un PCA est utilisé afin de visualiser les structures de population, mais également comme co-variable afin de corriger des biais potentiels liés à cette structure dans les études d'associations pangénomiques[44]. Cependant, un PCA est une méthode linéaire et identifie les directions avec une variance maximale et ignore les autres directions[45]. Une autre méthode parvient à surmonter ce problème : l'analyse d'approximation et projection uniforme de variétés ou UMAP (pour *uniform manifold approximation and projection*). Un UMAP est également une méthode de réduction de dimensionnalité qui est non-linéaire et qui a pour but de trouver une représentation des données qui préserve la structure locale au lieu de la structure globale (Figure 4B)[45].

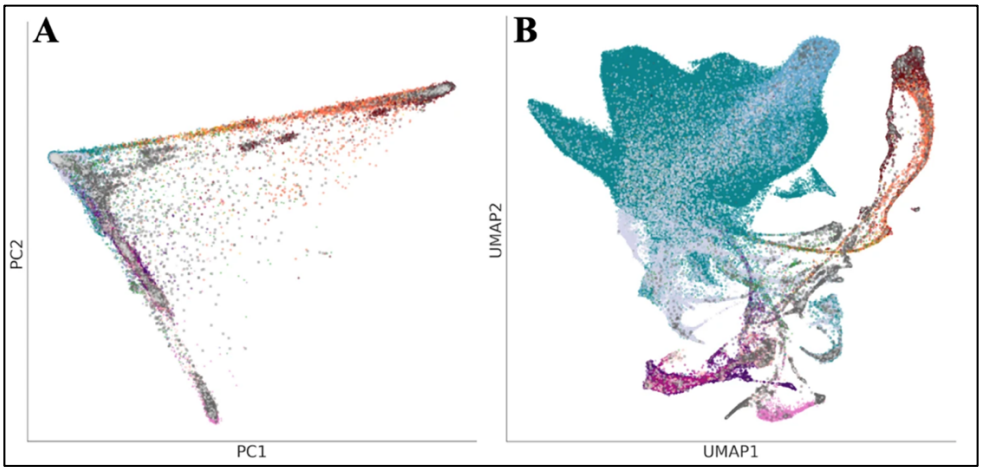


Figure 4. Représentation des données génétiques du UK Biobank.

A. Avec un PCA. **B.** Avec un UMAP. Tiré de (avec autorisation) : Diaz-Papkovich, A., Anderson-Trocmé, L. & Gravel, S. A review of UMAP in population genetics. *J Hum Genet* **66**, 85–91 (2021).

Les segments IBD renseignent sur les segments partagés entre des individus. Toutefois, la recombinaison fragmente les segments à chaque génération, donc plus un segment est ancien, plus il a de chance d'être fragmenté ou perdu[46]. Des positions avec une grande proportion de partage de segments IBD au sein d'une cohorte indiquent des positions génétiques intéressantes à analyser[47]. Par exemple, ces régions partagées entre des individus porteurs d'une même maladie indiqueraient des emplacements susceptibles d'être liés à cette maladie. Cela permet donc de prioriser l'étude de ces régions génétiques. De plus, avec ce type de segment, il est possible de considérer les analyses avec une échelle temporelle[48]. En effet, un segment IBD avec une longueur moyenne de 2cM correspondrait à un ancêtre commun ayant vécu il y a environ 25 générations[48]. Ces segments sont donc utilisés afin de mieux comprendre les ascendances communes récentes. Plus un lien entre deux individus est lointain, moins la proportion du génome partagé totale est grande et plus les segments IBD sont courts[48]. Des segments IBD chez une paire d'individus peuvent mener à la présence de segments ROH chez leur descendance. Ainsi, l'analyse de régions avec des segments ROH chez plusieurs individus peut aussi prioriser des régions à étudier et renseigne sur la consanguinité. Ces deux types de segments, ROH et IBD, peuvent être estimés par des méthodes similaires[48].

1.3.2 Données généalogiques

Les données généalogiques nous renseignent sur les phénomènes démographiques qu'une population a pu subir. Ces phénomènes démographiques sont le reflet de la société et de ses choix matrimoniaux, migratoires ou reproductifs, et c'est ce qui définit la génétique et la structure de la population[26,49,50]. Des phénomènes démographiques comme l'apparentement, la consanguinité, l'accroissement naturel et l'endogamie peuvent mener à la différenciation de populations. Toutefois, peu de populations sur terre ont accès à ce type de données[51–54]. La population québécoise est une de ces populations avec l'accès à sa généalogie sur plus de 400 ans grâce au projet BALSAC[55].

Le projet BALSAC a vu le jour à l'Université du Québec à Chicoutimi en 1972[55]. Le fichier BALSAC, contient les généalogies de l'ensemble du Québec depuis le début de la colonisation (1608) jusqu'à aujourd'hui[55]. Ces données uniques recueillies depuis maintenant plus de 50 ans permettent au Québec d'être une des rares populations au monde ayant accès à des données d'une telle qualité. Ce projet a été créé comme étant une méthodologie de jumelage de données nominatives avec l'utilisation d'informations provenant d'actes de baptême, mariage et sépulture de la région du Saguenay-Lac-Saint-Jean[26,55]. Les prêtres catholiques qui ont rédigé ces actes ont été très précis et plusieurs informations importantes y sont conservées. En effet, avec l'aide de ces actes de naissance, mariages et de décès, il est possible de retracer la vie d'un individu ; et surtout de le relier aux autres individus du Saguenay-Lac-Saint-Jean. Ainsi, la généalogie complète du Saguenay-Lac-Saint-Jean a été créée. Par la suite, ce projet a été étendu à travers toutes les régions du Québec au niveau des actes de mariage seulement. Aujourd'hui, le fichier compte plus de 5 millions d'actes d'état civil, avec plus de 6 millions d'individus répartis en 2,6 millions de famille, tout cela sur plus de 400 ans et sur l'ensemble du Québec[55].

Ce contexte a permis plusieurs recherches sur la population québécoise et son effet fondateur au niveau démographique[26,55–62]. Le jumelage de données génétiques et généalogiques a permis d'investiguer la structure de la population québécoise qui suit la géographie du Québec[50]. De ce fait, les données généalogiques sont un accès unique à l'histoire démographique d'une population, permettant d'observer directement les migrations et le taux de reproduction. Cela permet à ces populations de ne pas émettre d'hypothèses et de valider directement les relations entre les individus.

1.3.3 Analyses généalogiques et lien avec la génétique

Les généalogies sont utilisées pour calculer divers coefficients. À l'inverse de la génétique où il est seulement possible d'effectuer des hypothèses du passé à partir de données

contemporaines, les données généalogiques permettent d'avoir un portrait réel de l'évolution d'une population. Ainsi, la librairie R GENLIB a été créée afin de calculer plusieurs coefficients à partir de données généalogiques[63]. Il est possible de calculer plusieurs mesures, comme la complétude, l'apparentement, la consanguinité, la contribution génétique, et d'effectuer des simulations avec des allèles.

Le degré de complétude est une mesure calculée à chaque génération et représente la quantité d'ancêtres trouvés dans une généalogie. C'est la proportion du nombre d'ancêtres connus dans la généalogie comparée au nombre d'ancêtres attendu à chaque génération[63,64]. Le nombre d'individus attendu est calculé selon la formule 2^x, où x est le nombre de générations ; la génération des parents du proposant est la première génération[64]. Cette formule découle du fait que chaque individu, ou ancêtre, a deux parents différents. Ainsi, le nombre d'ancêtres total augmente de façon exponentielle. Une faible complétude peut mener à des biais sur plusieurs analyses généalogiques puisque cela apporte un manque de données sur les ancêtres. Ainsi, il serait possible de croire qu'il y a moins d'ancêtres, alors qu'il est seulement impossible d'avoir la vraie information. Il est donc important de faire la différence entre un vrai résultat ou à un manque de données.

Le coefficient d'apparentement est une mesure de proximité entre une paire d'individus au niveau généalogique. Selon Prost *et al.* (2022), cette valeur est mesurée pour deux personnes, A et B, comme étant « la probabilité pour qu'un gène pris au hasard chez A et un autre pris dans les mêmes conditions chez B, soit identiques-par-descendance ». Cela est aussi appliqué au niveau allélique ou au même locus chez deux individus, les deux allèles sont identiques-par-descendance[2]. Ce coefficient est calculé en recherchant tous les ancêtres communs pour une paire d'individus[65]. Plus un ancêtre commun est proche, plus il y a de probabilité que deux individus partagent un allèle identique[32]. Tandis que plus l'ancêtre est éloigné dans la

généalogie, moins les individus ont de possibilités de partager un allèle. Le coefficient d'apparentement généalogique est directement lié à la longueur totale des segments génétiques IBD chez une paire d'individus (Figure 5)[60]. Ces deux mesures, les deux calculées entre paires d'individus, nous renseignent sur les relations et liens présents au sein d'une population lorsque toutes les paires possibles sont mesurées.

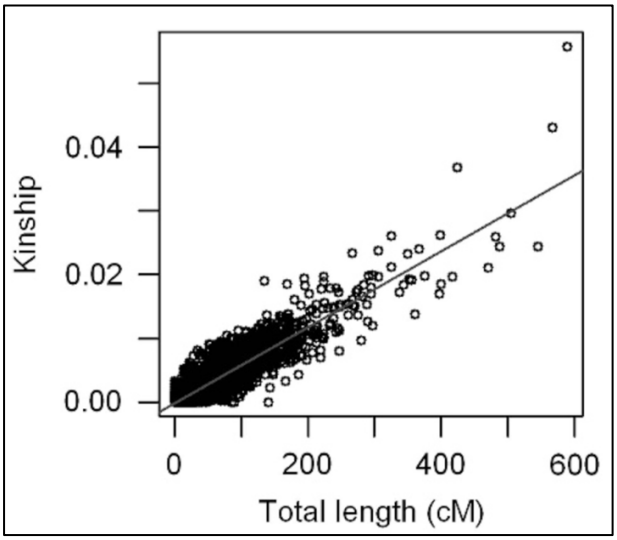


Figure 5. Corrélation entre les segments IBD et le coefficient d'apparentement

Tiré de (avec licence d'attribution non-commerciale Creative Commons) : Gauvin, H. *et al.* Genome-wide patterns of identity-by-descent sharing in the French Canadian founder population. *Eur J Hum Genet* **22**, 814–821 (2014).

L'étude des coefficients d'apparentement peut être effectuée à partir d'une analyse de positionnement multidimensionnel (MDS pour *multidimensional scaling*). Le MDS a pour but d'analyser des données de relations de grande dimension, tout en préservant les similarités[66]. Le MDS du coefficient d'apparentement d'une population reflète la structure de population qu'il est possible d'observer avec le PCA des données génétiques (Figure 6)[67,68]. Par conséquent, avec des données généalogiques, il est possible de suivre cette structure dans le temps ce qui est impossible de faire avec des données génétiques d'individus contemporains.

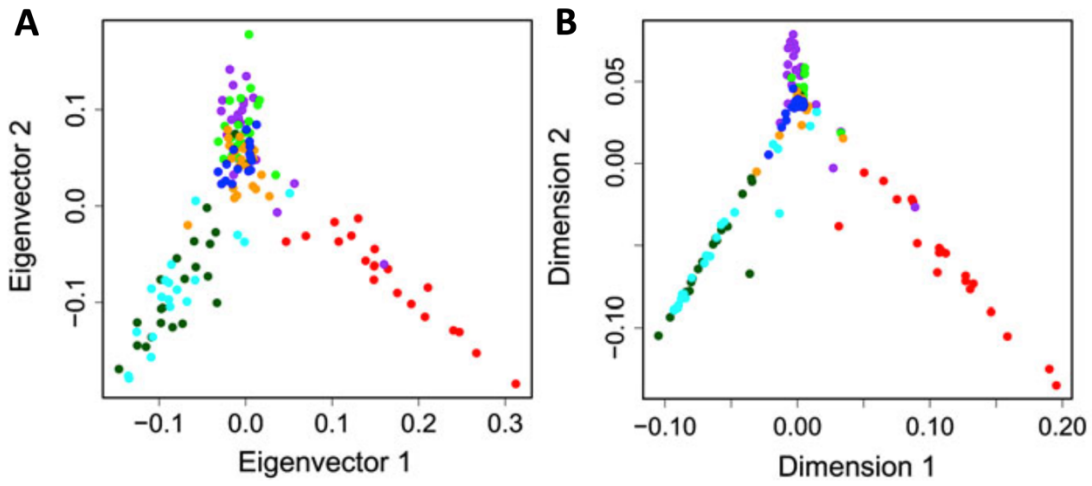


Figure 6. Structure du Québec capturée par des données génomiques et généalogiques

A. Avec un PCA des données génétiques. **B.** Avec un MDS du coefficient d'apparentement généalogique. Tiré de (avec autorisation) : Roy-Gagnon, M. H. *et al.* Genomic and genealogical investigation of the French Canadian founder population structure. *Human Genetics* **129**, 521–531 (2011).

De son côté, un individu est dit consanguin lorsque ses deux parents sont apparentés[2]. C'est donc une forme particulière d'apparentement, celle de l'apparentement dans un couple[49]. Ainsi, à un locus choisi au hasard, c'est la probabilité non nulle que deux allèles soient identiques-par-descendance, avec un allèle provenant du père et l'autre de la mère[49,69]. Le coefficient de consanguinité est relié à la notion de segment génétique ROH[2,49].

Le coefficient d'apparentement et la consanguinité sont donc fortement liés. Il est donc important, pour ces deux mesures, de faire la distinction entre un apparentement proche et éloigné. Un coefficient d'apparentement ou une consanguinité proche est défini comme étant à quatre ou cinq générations. Cette mesure renseigne sur le choix de son partenaire dans une société[49]. À l'inverse, lorsque c'est pour un nombre de générations plus élevé, il est question d'apparentement ou de consanguinité éloignée. Des niveaux du coefficient d'apparentement ou de consanguinité éloignés élevés témoignent souvent d'un effet fondateur[49]. Cette mesure aide à comprendre l'histoire démographique et son impact sur la structure d'une population.

La composition génétique d'une population dépend de ses ancêtres et de la transmission des gènes d'un ancêtre à un individu de la génération actuelle[70]. Un parent transmet 50% de son génome à son enfant. Par le fait même, des grands-parents transmettent donc 25% de leur génome à leur petit enfant. La contribution génétique est donc définie comme étant la contribution d'un ancêtre au bassin génétique actuel et la fécondité des ancêtres[70]. Le calcul de la contribution génétique à partir de la généalogie consiste à sommer les probabilités de transmission sur tous les chemins généalogiques reliant un ancêtre à un descendant. Plus un ancêtre est présent souvent, plus il va avoir une grande contribution au bassin génétique contemporain.

Une autre mesure intéressante est l'identification d'ancêtres communs les plus récents, ou aussi nommés MRCA (pour *most recent common ancestors*)[63]. Calculée pour une paire ou un groupe d'individus, l'identification de ces ancêtres est très utile afin de déterminer le coefficient d'apparentement. De plus, il est possible de mesurer la distance minimale généalogique qui relie deux individus à ce MRCA, et qui est liée à la longueur des segments génétiques IBD. Cette valeur est mesurée en méioses, où 2 méioses équivalent à une génération. Il est possible d'identifier plusieurs MRCA pour une paire ou groupe d'individus. Cela est possible lorsqu'aucun ancêtre dans l'ensemble des MRCA ne partage un descendant qui est également un ancêtre de la paire d'individus intial[60].

1.4 Populations à effet fondateur à l'étude

Cette section-ci décrit les quatre populations à effet fondateur étudiées dans ce mémoire. Ces populations ont été choisies en raison de leur effet fondateur bien défini et de la disponibilité des données en accès libre. Plusieurs autres populations à effet fondateur avaient été sectionnées pour le projet. Toutefois, nous n'avons pas pu avoir accès en raison du type de consentement.

1.4.1 Québec

Avant même l'arrivée des premiers colons européens, le territoire actuel du Québec était occupé par différents peuples autochtones. Il y a eu un léger métissage entre ces peuples et les Européens[68,71]. Toutefois, la partie la plus importante du bassin génétique des Canadiens-Français du Québec est issue des vagues d'immigrations provenant de France[32]. Les Québécois descendent d'environ 8 500 colons qui sont arrivés pour la plupart de France entre 1608 et 1759[72,73]. En suivant le fleuve Saint-Laurent, ces Européens s'installent à Québec (1608), Trois-Rivières (1634) et Montréal (1642) afin de fonder la Nouvelle-France (Figure 7)[72]. Aujourd'hui, ces villes sont de grandes régions urbaines. Avec la conquête britannique de 1759-1760, la Nouvelle-France est tombée sous le contrôle de la Grande-Bretagne[74]. À cette époque, il y avait environ 76 000 Canadiens-Français le long du fleuve Saint-Laurent[73]. Ainsi, l'immigration de France a considérablement diminué et la population francophone s'est développée principalement par accroissement naturel, fortement encouragé par le clergé. De plus, des barrières linguistique et religieuse ont limité le mélange entre les Canadiens-Français et les Canadiens-Anglais[73]. À cet effet, l'effet fondateur des Québécois d'origine canadienne-française est une conséquence du goulot d'étranglement provoqué par l'immigration initiale uniquement de France suivi par un accroissement presque exclusivement naturel jusqu'environ au début du 20^{ième} siècle[60,67,68,75]. De plus, la consanguinité proche n'a jamais été commune chez les Québécois. C'est la consanguinité éloignée qui est plus fréquente, caractérisée par la petite taille du goulot d'étranglement et de la mobilité réduite de la population et de son expansion[26,73]. Tout cela a mené à une prévalence élevée de plusieurs maladies au Québec, comme la fibrose kystique, l'hypercholestérolémie familiale et la phénylcétonurie[58,73]. Il y a au moins 28 maladies héréditaires qui seraient plus fréquentes dans au moins une région du Québec et qui sont distribuées inégalement (Figure 7)[57–59,73,76]. Certaines de ces maladies peuvent être présentes dans plus d'une région[59].

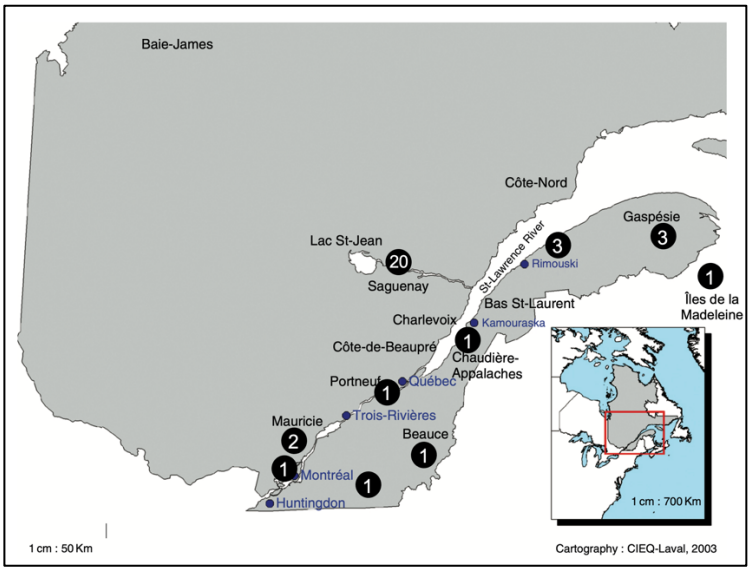


Figure 7. Carte du Québec de la distribution des maladies héréditaires plus fréquentes

Les ronds noirs indiquent la distribution géographique du nombre de maladies héréditaires plus fréquentes au Québec selon De Braekeleer et Dao (1994). Tiré de (avec modification et autorisation) : Laberge, A.-M. *et al.* Population history and its impact on medical genetics in Quebec. *Clin Genet* **68**, 287–301 (2005).

Ce fort accroissement naturel a apporté la nécessité de coloniser de nouvelles régions, incluant des régions plus éloignées et isolées, ce qui favorise la subdivision de la population et une série d'effets fondateurs régionaux[56,57,73]. C'est à ces fronts pionniers, lors de la colonisation de nouvelles régions, que le taux de natalité a été le plus élevé[73]. Par exemple, lors de la colonisation du Saguenay-Lac-Saint-Jean, la taille de la population a augmenté de 25 fois en seulement un siècle en raison d'un taux de natalité très élevé, lorsque la taille du Québec a augmenté seulement de 5 fois[60,77,78]. Chaque région que l'on connait aujourd'hui détient un bassin génétique unique, caractérisé par les premières générations de colons qui ont fondé la nouvelle région, certains ayant une contribution génétique disproportionnée[73]. Ainsi, c'est ce qui caractérise les effets fondateurs successifs du Québec. Le mariage entre individus d'une même région, l'isolement géographique et la dérive génétique sont également des facteurs ayant mené à la structure génétique du Québec. Jusqu'à maintenant, des analyses génétiques et généalogiques démontrent que deux régions se distingueraient particulièrement, soit le

Saguenay-Lac-Saint-Jean et la Gaspésie[49,60,67,68]. Plusieurs maladies génétiques rares sont plus fréquentes au Saguenay-Lac-Saint-Jean comparé au reste du monde, et même du Québec[37,79]. Il existe un test de porteur pour quatre de ces maladies pour les individus originaires du Saguenay-Lac-Saint-Jean et de la haute Côte Nord[37,80]. Ces maladies sont l'acidose lactique congénitale, l'ataxie récessive spastique de Charlevoix-Saguenay, la neuropathie sensitivomotrice héréditaire avec ou sans agénésie du corps calleux et la tyrosinémie héréditaire de type 1[37,79]. Ce test permet de savoir si les individus d'un couple sont porteurs d'une ou de plusieurs de ces maladies et offre un suivi en clinique génétique au couple porteur de la même mutation. D'un autre côté, des maladies, comme la phénylcétonurie, ne sont pas ou peu présente au Saguenay-Lac-Saint-Jean [37,61].

Ainsi, la population canadienne-française n'est pas aussi homogène qu'elle parait, avec la distinction de plusieurs groupes régionaux menant à une stratification de la population[58,68,73]. Cette stratification est caractérisée par une contribution génétique inégale et non aléatoire d'ancêtres distincts[68].

1.4.2 Juifs Ashkénazes

Le peuple juif a pris naissance au Moyen-Orient durant l'âge de bronze[81,82]. Pendant plus de 2000 ans, ils ont été un peuple migratoire et ont établi des communautés au Moyen-Orient et dans le secteur du bassin méditerranéen[81]. Au sein de ces communautés, les Juifs sont liés par la langue, la religion, les coutumes et le mariage[83]. Cela a renforcé l'identité juive jusqu'à aujourd'hui. Actuellement, trois groupes juifs sont définis : les Juifs du Moyen-Orient, les Juifs Séfarades et les Juifs Ashkénazes. Des études génétiques ont prouvé avec l'aide d'analyse de l'ADN mitochondrial et du chromosome Y qu'il y a un lien entre les peuples juifs, autant du côté matrilinéaire et patrilinéaire, et ce sur plusieurs générations[81,84]. De plus, l'endogamie est

la norme avec un taux de métissage très faible sur plus de 80 générations. Au 20e siècle, trois évènements marquants impactent la démographie des peuples juifs : l'Holocauste juif de la Seconde Guerre mondiale, l'immigration des Juifs vers Israël et le mariage de Juifs entre des Juifs et des individus non-Juifs[81].

Avec une population de plus de 10 millions d'individus établie à travers le monde aujourd'hui, et avec un nombre relativement petit de fondateurs, plusieurs goulots d'étranglement et une expansion rapide de la population, les Juifs Ashkénazes ont vécu un effet fondateur important[82,83,85–88]. Le goulot d'étranglement principal serait survenu au Moyen-Âge avec une réduction de la taille de population dans les centaines d'individus[85,88]. Toutefois, l'ensemble des goulots d'étranglement ne sont pas encore bien compris. Malgré que les Juifs Ashkénazes sont aujourd'hui présents à plusieurs endroits, ils sont très semblables au niveau génétique, sans différence marquante entre les pays de résidence actuel[85]. Cela a conduit à des caractéristiques génétiques distinctives telles que la prévalence élevée de maladies autosomiques récessives ou l'augmentation de la fréquence de certaines maladies courantes, comme la maladie de Tay-Sachs ou le cancer du sein[27,83,88,89]. Avec les années, des variants spécifiques des Juifs Ashkénazes ont été découvert puisque l'effet fondateur sévère des Juifs Ashkénazes permet l'identification de variants rares dans cette population[89–91]. Il y a même des maladies, comme la dysautonomie familiale, qui est spécifique et existe seulement chez les Juifs Ashkénazes[81]. De ce fait, des programmes de dépistage préconceptionnel existent afin de prévenir ces maladies[85].

Les goulots d'étranglement successifs, un bassin génétique limité et un fort accroissement naturel ont mené à la population juive Ashkénazes que l'on connait aujourd'hui et à son paysage génétique actuel.

1.4.3 Huttérites

Les Huttérites sont connus en tant que groupe anabaptiste créé au XVIe siècle en Europe de l'Ouest, actuellement où est situé l'Autriche[34,92–94]. Ils sont une des jeunes populations à effet fondateur les mieux caractérisées[34]. Ils se différencient des autres groupes anabaptistes par leur mode de vie en communauté de 60 à 120 individus et par la pratique de la propriété communautaire des biens[92,93]. Leur histoire est marquée par de nombreux évènements de déplacements et de persécutions religieuses, qui ont entrainé une réduction de la population; mais également par des périodes de prospérité et d'augmentation du nombre d'individus[34,92]. Entre 1874 et 1879, un peu plus de 400 individus de la population Huttérite ont migré d'Europe vers les États-Unis en trois vagues[92,95]. Ils se sont établis en trois colonies, nommées Leuts : Schmiedeleut, Dariusleut, et Lehrerleut, au Dakota du Sud[34,92,94]. Ces individus constituent la population ancestrale Huttérite d'Amérique du Nord[96]. La population Huttérite détient des archives généalogiques détaillées montrant que la grande majorité des Huttérites d'Amérique du Nord contemporains descendent de 89 ancêtres[34,92,96]. Lors de la Première Guerre mondiale, les Huttérites refusent de participer à l'effort de guerre et à la conscription[92]. Ils se sont déplacés vers le Canada puisqu'on leur avait offert l'exemption militaire, la liberté religieuse et le droit à leurs propres écoles[92]. Ils se sont établis en 15 colonies dans les provinces des prairies canadiennes. La population a connu une croissance spectaculaire, atteignant aujourd'hui plus de 40 000 membres dans plus de 400 colonies communautaires dans le centre nord des États-Unis et les Prairies canadiennes (Figure 8)[92,95,96]. Chacune de ces colonies contemporaines descend d'un des trois Leuts originaux[95]. En plus d'être séparé en 3 groupes de façon géographique, chaque Leut a développé des différences culturelles et une identité propre. De plus, ils pratiquent l'endogamie au sein du même groupe[93,96]. Alors, les mariages entre les Leuts n'arrivent que rarement et les mariages avec des individus de confession non Huttérite sont proscrits[92,93,95]. Également, la croissance depuis leur arrivée aux États-Unis s'est presqu'uniquement faite par accroissement naturel, avec la présence de grandes familles, la

quasi absence de célibat, l'interdiction d'utiliser des moyens de contraception et l'interdiction de divorcer[93–95]. Le taux de natalité serait considéré comme le plus élevé au monde avec un taux de mortalité infantile faible[93]. En plus de toutes ces caractéristiques, les Huttérites ont un mode de vie austère et agricole, créant un environnement uniforme[96]. Tout cela a favorisé une augmentation du niveau de consanguinité dans la population. Une paire d'individus pris au hasard dans la population Huttérite serait reliée en moyenne à un niveau plus élevé que celui de petit cousin[93,95]. Ceci fait en sorte que les Huttérites sont un exemple extrême d'effet fondateur[32].

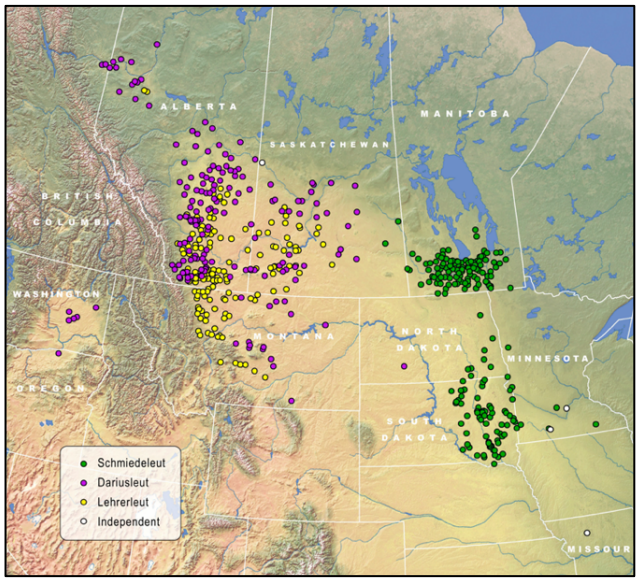


Figure 8. Carte des différentes colonies des Huttérites en Amérique du Nord

Tiré de (avec licence d'attribution Creative Commons) : Katz, Y. & Lehr, J. The digital revolution and the Hutterite community: The rules and reality. *Prairie Perspectives: Geographical Essays* **21**, 9–15 (2019).

Cette population a permis de faire des avancées au niveau de la recherche sur des maladies génétiques. En effet, leur petit nombre de fondateurs jumelé à des caractéristiques sociales a apporté une isolation génétique du reste des populations d'Amérique[92]. Cela a mené à une propagation d'allèles causant des maladies génétiques au sein de la population. Plus de 28

maladies ont été cartographiées et cela a permis d'identifier le gène et la mutation spécifiques concernés chez plus de la moitié de celle-ci[34,92]. Certaines maladies à traits complexes sont également plus fréquentes dans cette population. Les Huttérites sont un excellent exemple où l'effet fondateur a apporté une réduction de la variation génétique, et où la dérive génétique a mené à la perte de la majorité des variants rares[34].

1.4.4 Himba

En ce qui concerne les Himbas, il s'agit d'un groupe pastoral semi-nomade vivant en Namibie, dans la région du Kaokoland[97–100]. Les Himbas sont un sous-groupe indépendant du peuple Héréro, un groupe bantou qui s'est installé en Namibie au milieu du 16^{e} siècle[97,98,100]. Leur langue et leur culture sont similaires et ils sont proches au niveau génétique, mais les Himbas sont restés plus isolés de la technologie que les Héréros[97,98]. Leur mode de vie est basé sur une économie pastorale de leur bétail et de l'horticulture, avec l'absence d'eau et d'électricité, et avec un accès limité à l'économie de marché qu'ils utilisent seulement pour vendre leur bétail[99,101]. Les ménages sont composés de familles élargies avec une taille de 8 à 25 indvidus[101]. Il est courant pour les hommes et les femmes, mariés ou non, d'avoir des partenaires simultanés[99,101]. Il y a un taux élevé d'enfants issus de relations extra-conjugales où les femmes déclarent que 17% de leur descendance sont issue d'une relation hors couple[97].

Les facteurs culturels et le mode de vie des Himbas ont mené à une population assez isolée. Il y a de hauts niveaux d'homozygotie et de parenté chez les Himbas révélés avec des études sur les segments ROH et IBD[102]. Depuis environ 170 ans, les Himba ont vécu plusieurs évènements qui ont contribué au déclin de leur population[102]. Aux 19^{e} et 20^{e} siècles, des évènements tels que la peste bovine et la pleuropneumonie contagieuse bovine ont causé des pertes dévastatrices dans leur bétail[102,103]. Ils ont également connu d'autres évènements tels que plusieurs sécheresses sévères qui ont apporté des pertes lourdes de leur bétail, des pressions

génocidaires, des taxes élevées, une mobilité réduite et l'isolement[100,102,103]. Selon les données génétiques, un goulot d'étranglement aurait eu lieu il y a 12 générations, démontrant l'impact notable de ces facteurs environnementaux sur la population[102]. Tous ces facteurs ont mené à une diminution de la population et à la présence de long ROH chez les Himbas. Environ 2.6% de leur génome serait composé de longs segments d'homozygoties de plus de 1 500kb[102]. Ces longs ROH auraient un impact négatif sur la fertilité chez les Himbas en raison d'une charge de mutation récessive[102].

1.5 Objectifs

Les populations à effet fondateur ont été particulièrement étudiées en raison de leur structure génétique distincte due à leur histoire démographique unique et de leur implication au niveau des maladies génétiques. Toutefois, ce type de population n'a pas révélé tout son potentiel et il est encore essentiel de les étudier. En effet, l'étude des populations à effet fondateur peut être un outil puissant dans l'identification des variations encore inconnues liées à des maladies et dans la meilleure compréhension de la structure fine des populations.

L'objectif principal de mon projet de maitrise est d'étudier les effets fondateurs à deux niveaux. D'abord, afin de mieux comprendre leur structure génétique. Ensuite, afin de mieux comprendre l'importance des phénomènes démographiques dans l'apparition de ces effets fondateurs. Mon projet de maitrise est divisé en deux chapitres. L'objectif du premier chapitre est de décrire correctement la structure fine d'une population pour affiner les modèles d'analyse pour la découverte de nouvelles associations génétiques. Pour ce faire, plusieurs populations à effet fondateur sont étudiées ensemble afin de comparer, quantifier et mieux comprendre leur structure fine. Ainsi, l'impact de cette structure fine sur la fréquence de variants rares déjà connus pour être associés à un effet fondateur spécifique est étudié. Cela a pour objectif de comprendre l'implication des structures de populations dans l'identification de variantes rares.

Par la suite, au niveau du second chapitre, l'objectif est d'utiliser les données génétiques et généalogiques de la population du Québec afin de suivre l'évolution de la structure démographique régionale depuis sa colonisation jusqu'à nos jours. Ensuite, l'investigation en profondeur de la structure fine avec l'aide de données généalogiques et génétiques a pour objectif de mieux comprendre l'effet des processus démographiques qui a pu mener à ces effets fondateurs régionaux.

Chapitre 1 : Fine-scale genetic structure and rare variant frequencies

Avant-propos

L'article *Fine-scale genetic structure and rare variant frequencies* présente les résultats de travaux de recherches effectués sur quatre populations à effet fondateur afin d'explorer leur structure fine et son impact sur les études d'association génétique. Cet article est publié sur bioR*x*iv, un serveur de prépublication, le 13 février 2024 sous le DOI suivant : https://doi.org/10.1101/2024.02.02.578687. Il a également été soumis au *European Journal of Human Genetics* le 3 avril 2024. L'article a été écrit et soumis avec un format de publication court. Les figures et tableaux supplémentaires se retrouvent à l'Annexe 1 de ce mémoire.

Au niveau des différentes cohortes, je me suis occupé des recherches afin de trouver les cohortes de trois populations à effet fondateur disponible en libre accès sur dbGAP, soit les Juifs Ashkénazes, les Huttérites et les Himbas, que nous voulions utiliser. Les processus administratifs afin d'avoir accès aux données ont été effectués par mon directeur Simon Girard. Je me suis également occupée des données contrôle du *1000 Genomes Project*. Pour ce qui est des données de la dernière population à effet fondateur, le Québec, la cohorte était accessible par le laboratoire. J'ai joué un rôle important dans la conception du projet. J'ai effectué toutes les analyses et j'ai fait toutes les figures et les tableaux. J'ai écrit la première version de l'article que j'ai amélioré avec les commentaires et suggestions des co-auteurs. J'ai également effectué la publication sur bioR*x*iv et la soumission au *European Journal of Human Genetic*.

Fine-scale genetic structure and rare variant frequencies

Laurence Gagnon,[1,2] Claudia Moreau,[1,2] Catherine Laprise,[1,2,3] Simon L. Girard[1,2,4].

1. Département des sciences fondamentales, Université du Québec à Chicoutimi, Saguenay, Québec, G7H 2B1, Canada.

2. Centre Intersectoriel en Santé Durable (CISD), Université du Québec à Chicoutimi, Saguenay, Québec, G7H 2B1, Canada.

3. Centre Intégré Universitaire en Santé et Services Sociaux du Saguenay–Lac-Saint-Jean, Saguenay, Québec, QC G7H 7K9, Canada

4. Centre de recherche CERVO, Université Laval, Québec, Québec, G1V 0A6, Canada.

Contact info

Simon Girard, PhD

Université du Québec à Chicoutimi

555, boulevard de l'Université

Chicoutimi (Québec) G7H 2B1

simon2_girard@uqac.ca

Résumé

En réponse au défi actuel des études génétiques visant à établir de nouvelles associations génétiques, nous préconisons un virage vers l'utilisation de la structure fine des populations. En utilisant les données génétiques de quatre populations ayant subi un effet fondateur, notre projet met en lumière des structures de populations fines avec la présence de regroupement, remettant en question la perception antérieure d'homogénéité. Ensuite, nous avons évalué l'impact de ses regroupements sur la fréquence d'allèles associés à des maladies fréquentes dans certaines populations à effets fondateurs spécifiques. Ainsi, cela souligne que des cohortes plus petites, mais bien définies, présentent une importante augmentation des fréquences de variants rares, offrant une avenue prometteuse pour la découverte de nouveaux variants génétiques.

Abstract

In response to the current challenge in genetic studies to make new associations, we advocate for a shift toward leveraging population fine-scale structure. Our exploration brings to light distinct fine-structure within populations having undergone a founder effect, challenging the prior perception of homogeneity. This underscores that smaller, but well-defined cohorts, demonstrate an important increase in rare variant frequencies, offering a promising avenue for new genetic variants' discovery.

Keywords

Population with a founder effect; Population genetics; Rare variants; Clusters

Introduction

Common variants are the primary source of variation identified by genetic association studies. However, despite numerous association analyses conducted in the last years, a significant proportion of the genetic predisposition for many diseases still remains unknown. To address this issue, it is crucial to study rare variants, which often have more significant phenotypic effects, to better understand this "missing heritability". Nevertheless, associations with rare variants present a reduction in statistical power due to the scarcity of individuals carrying these alleles (1). Thus, given their unique structure, populations that had undergone a founder effect (PFE) have the potential to more readily reveal new genetic associations of rare variants that have potential implications for human health (2–7), and may help to reduce the "missing heritability" (8). Therefore, our study aims to properly describe the fine-scale genetic structure of a population to refine analysis models for genetic associations.

Subjects and methods

This study was approved by the University of Quebec in Chicoutimi (UQAC) ethics board.

Cohorts

The data consist of 4 different cohorts of populations that had undergone a founder effect. We gained access to data from Quebec, Ashkenazi Jews, Himba and Hutterites (Table S1). We also used the data from the 1000 Genomes Project phase 3 as reference groups from Africa (Mende (MSL)), Europe (British, Northern and Western Europe (GBR and CEU)) and East Asia (Han Chinese and Japanese (CHB and JPT)). Thus, this led to a final sample size of 3,683 individuals. Please refer to the Supplementary information for comprehensive details on the cleaning process and identical-by-descent (IBD) segments computation.

Statistical analysis

A Uniform Manifold Approximation and Projection (UMAP) was performed on the first 8 principal components of the Principal Component Analysis (PCA) to capture as much variance of the fine structure as possible. Additional UMAPs were completed on each individual dataset, using the first 5, 4, 5, 8 principal components for the Ashkenazi Jews, Quebec, Himba and Hutterites, respectively (Figure S1). The UMAPs were realized with the R package "umap" v0.9.2.0 (9). The n neighbors variable was set to the maximal value (number of individuals in the dataset) (Table S1). The min distance value was set to 0.9 for the UMAP on the merged dataset to promote dots splitting and ensure good visualization; while it was set to 0.01 for the UMAPs on individual datasets of populations to better capture the structure and promote clustering (10).

The DBScan method was employed to generate clusters from UMAP of each population. This method was chosen for its ability in density-based clustering, allowing the capture of clusters with various shapes, including non-convex shapes (11). The "dbscan" R library v1.1-11 was used for clustering with the minPts parameter set to 4, and the epsilon value adjusted according to each individual population. The Himba and Hutterites only displayed a single cluster each, while Ashkenazi Jews and Quebec exhibited 4 and 5 clusters, respectively (Figure S1).

Minor allele frequency (MAF) of known disease-causing variants were computed using PLINK software v1.9 on imputed data, see Supplementary information for details. It's important to note that using imputed data may result in a loss of rare variants or in the underestimation of the real frequency of these variants in a PFE. The founder variants (listed in Table 1 and Table S2) were selected because of their association with specific populations. However, limited literature is available concerning specific variants in the Acadians of Gaspe (Quebec-3) which could explain why we found only one variant associated of interest in this cluster.

Results

Fine-scale population structure

We characterized the genetic structure of four PFE alongside three reference groups from the 1000 Genomes Project. These populations are the Himba of Namibia, the Hutterites of South Dakota in the US, the population of the Quebec' province in Canada and the Ashkenazi Jews of Europe and across the World (Table S1). A UMAP analysis was conducted to analyze their genetic structure (Figure 1A). The Himba and Hutterites form two tightly packed clusters, whereas the Ashkenazi Jews and Quebec exhibit a more dispersed pattern and are even interconnected. For Hutterites and Himba, they can be studied as a whole as they do not subdivide into clusters. Indeed, Hutterites are known to practice endogamy and live in community and Himba individuals were documented to practice polygyny and live in a pastoralist way (12,13). This way of living promotes very close links between individuals and could explain the absence of fine-structure. This is also evident in the proportion of pairs sharing an IBD segment across the genome, which reaches the highest level among the Himba and Hutterites compared to the other PFE. In contrast, a fine-structure is observed within the Ashkenazi Jews and the Quebec population so that they can be subdivided into clusters created by DBScan method (Figure 1B). The fine-structure in these populations is also reflected in the patterns of sharing of IBD segments; indeed, some of the clusters exhibit increased IBD sharing across the genome (Table S3). The proportion of pairs sharing an IBD segment through the genome within these clusters reaches an average threshold comparable to the one observed among Himba and Hutterites (Table S3). The Ashkenazi Jews and Quebec went through unique histories of migration, isolation and population expansion. The Ashkenazi Jews-1 cluster represents the Ashkenazi Jewish ancestry, i.e. individuals for whom all four grand-parents were of Ashkenazi Jewish ancestry (Figure S2A). In contrast, the Ashkenazi Jews-2 cluster appears to represent a more admixed ancestry, due to its connection to the European reference group and Quebec (Figure 1B). This cluster likely contains individuals with only 1 to 3 of their grand-

parents with Ashkenazi Jewish ancestry (6). Moreover, the analysis of the genetic relatedness among and between clusters reveals that the Ashkenazi Jews clusters are distinct and exhibit greater relatedness within than between clusters (Figure S3A). As for Quebec, clusters can be associated with specific ethnocultural groups. Specifically, the Quebec-2 and Quebec-3 represent the Saguenay-Lac-Saint-Jean and the Acadians of Gaspe, respectively (Figure S2B). These two groups are known for having a genetic structure which distinguishes them from the broader Quebec population that can be associated with Quebec-1 cluster (14–16). This cluster may be related to the initial founder effect in Quebec. This is evident in the lower genetic relatedness, for example more diversity amongst founders or larger fraction of recent immigrants amongst their ancestors, observed within the Quebec-1 cluster compared to all the others (Figure S3B). Undoubtedly, populations like the Ashkenazi Jews and Quebec cannot be treated as single entities due to the presence of fine-structure, even if they were initially perceived as “homogenous populations”.

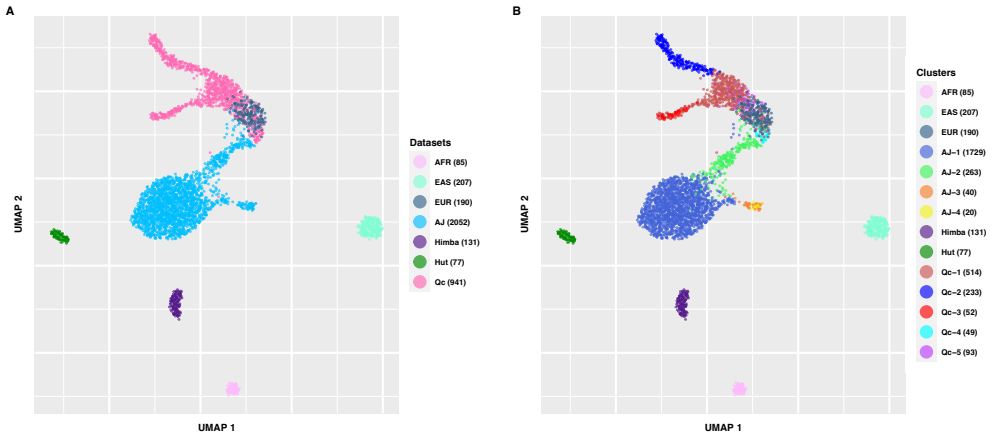


Figure 1. UMAP of the first 8 principal components of the merged dataset. Colored according to the origin of the population (A) and the clusters (on Supplementary Figure 1) (B). AFR African from the 1000 Genomes Project, EAS East Asian from the 1000 Genomes Project, EUR European from the 1000 Genomes Project, AJ Ashkenazi Jews, Hut Hutterites, Qc Quebec.

Impact of clustering on genetic association

To assess the impact of clustering on genetic association, we conducted an analysis of known founder disease variants. These variants were selected for their genetic association among a specific group. The analysis of the frequency of these variants in the associated group (Quebec-2 for Saguenay-Lac-St-Jean, Quebec-3 for Acadians of Gaspe, or Ashkenazi Jews-1 for Ashkenazi Jews) (Figure S2), revealed higher allele frequencies of the variant in the specific cluster compared with the other clusters (Table 1). They were also nearly absent from the European reference group. Remarkably, within the Quebec population, the variants associated with spastic ataxia of Charlevoix-Saguenay and Usher syndrome type I exhibit a 4 and 8 fold increase, respectively, when comparing the specific cluster and the whole population. Notably, this was calculated with a much smaller sample size of 4 and 18 fold, respectively. This trend is also observable while investigating the variant of Familial dysautonomia in the Ashkenazi Jews and other diseases associated with a specific population (Table 1 and Table S2). So, it is crucial to consider these clusters not only in PFE(17), but also in outbred populations since the presence of clusters in more diverse or admixed populations could even have a more striking effect on genetic associations (18,19). This approach could result in the need for much smaller cohorts consisting of individuals with well-known fine-scale genetic structure, offering cost-effectiveness and increased statistical power.

Table 1. Frequency of disease-causing variants known to be associated with a specific population. Only the main population and the associated cluster are shown on the table.

Population	MAF of the European reference group of 1000GP	MAF of the PFE	Number of individuals
Familial dysautonomia (chr9:108899816:A:G) (ClinVar: 6085)			
Ashkenazi Jews	0.0000	0.0178	2052
Ashkenazi Jews-1	0.0000	0.0211	1729
Spastic ataxia of Charlevoix-Saguenay (chr13:23335031:TA:T) (ClinVar: 5512)			
Quebec	0.0000	0.0080	941
Quebec-2	0.0000	0.0323	233
Usher syndrome type I (chr11:17531431:C:T) (ClinVar: 5143)			
Quebec	0.0000	0.0048	941
Quebec-3	0.0000	0.0385	52

Discussion

In conclusion, we suggest a novel approach, parallel to the already existing strategies, that uses the fine-scale genetic structure of a population to refine analysis models for genetic associations. This cost-effectiveness method would help to enhance the value of existing large cohorts and to develop new analytical methods. We believe that cohorts composed of fewer individuals with a common genetic background would help in discovering new rare genetic associations, as they would be easier to find given their increased frequency. Investigating more prevalent diseases within targeted populations has the potential to generate positive impacts on public health at the community level and on the discovery of new genes that could be new therapeutic targets.

Data Availability

The Quebec cohort genotypes are available upon request to BALSAC at https://balsac.uqac.ca/ (20). The Ashkenazi Jews cohort data may be available via dbGaP study accession number phs000448.v1.p1, the Himba cohort data may be available via dbGaP study accession number phs001995.v1.p1 and the Hutterites cohort data may be available via dbGaP study accession number phs001033.v1.p1.

Code Availability

The code used for this study can be found in the following GitHub repository: https://github.com/laugag17/world_pop_with_founder_effect.

References

1. Goswami, C., Chattopadhyay, A. & Chuang, E. Y. Rare variants: data types and analysis strategies. *Ann Transl Med* **9**, 961 (2021).
2. Casals, F. *et al.* Whole-exome sequencing reveals a rapid change in the frequency of rare functional variants in a founding population of humans. *PLoS Genet* **9**, e1003815 (2013).
3. Lencz, T. *et al.* Novel ultra-rare exonic variants identified in a founder population implicate cadherins in schizophrenia. *Neuron* **109**, 1465–1478 (2021).
4. Southam, L. *et al.* Whole genome sequencing and imputation in isolated populations identify genetic associations with medically-relevant complex traits. *Nat Commun* **8**, 15606 (2017).
5. Scriver, C. R. Human Genetics : Lessons from Quebec Populations. *Annual Review of Genomics and Human Genetics* **2**, 69–101 (2001).
6. Guha, S. *et al.* Implications for health and disease in the genetic signature of the Ashkenazi Jewish population. *Genome Biology* **13**, R2 (2012).
7. Carmi, S. *et al.* Sequencing an Ashkenazi reference panel supports population-targeted

personal genomics and illuminates Jewish and European origins. *Nat Commun* **5**, 4835 (2014).

8. Uricchio, L. H. Evolutionary perspectives on polygenic selection, missing heritability, and GWAS. *Hum Genet* **139**, 5–21 (2020).
9. McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:180203426. 2020;
10. Allaoui M, Kherfi ML, Cheriet A. Considerably Improving Clustering Algorithms Using UMAP Dimensionality Reduction Technique: A Comparative Study. In: El Moataz A, Mammass D, Mansouri A, Nouboud F, editors. Image and Signal Processing. 2020. p. 317–25.
11. Hahsler M, Piekenbrock M, Doran D. dbscan: Fast Density-Based Clustering with R. Journal of Statistical Software. 2019 Oct 31;91:1–30.
12. Nimgaonkar VL, Fujiwara TM, Dutta M, Wood J, Gentry K, Maendel S, et al. Low prevalence of psychoses among the Hutterites, an isolated religious community. Am J Psychiatry. 2000 Jul;157(7):1065–70.
13. Scelza BA. Female mobility and postmarital kin access in a patrilocal society. Hum Nat. 2011 Dec;22(4):377–93.
14. Gagnon L, Moreau C, Laprise C, Vézina H, Girard SL. Deciphering the genetic structure of the Quebec founder population using genealogies. European Journal of Human Genetics. 2024 Jan 1;32(1):91–7.
15. 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. Human Genetics. 2011 May;129(5):521–31.
16. Bherer C, Labuda D, Roy-Gagnon MH, Houde L, Tremblay M, Vézina H. Admixed ancestry and stratification of Quebec regional populations. American Journal of Physical Anthropology. 2011;144(3):432–41.

17. Xue Y, Mezzavilla M, Haber M, McCarthy S, Chen Y, Narasimhan V, et al. Enrichment of low-frequency functional variants revealed by whole-genome sequencing of multiple isolated European populations. Nat Commun. 2017 Jun 23;8(1):15927.
18. Gouveia MH, Bentley AR, Leal TP, Tarazona-Santos E, Bustamante CD, Adeyemo AA, et al. Unappreciated subcontinental admixture in Europeans and European Americans and implications for genetic epidemiology studies. Nat Commun. 2023 Nov 7;14(1):6802.
19. Koyama S, Wang Y, Paruchuri K, Uddin MM, Cho SMJ, Urbut SM, et al. Decoding Genetics, Ancestry, and Geospatial Context for Precision Health. medRxiv. 2023 Oct;2023.10.24.23297096.
20. BALSAC. BALSAC. [cited 2022 May 27]. BALSAC. Available from: https://balsac.uqac.ca/

Acknowledgements

This work was supported by funding from the Canada Research Chair in Genetics and Genealogy hold by SLG. It was also made possible by the Digital Research Alliance of Canada which provided access to storage and computing resources. We are extremely grateful to all participants of this research. We would like to thank Hélène Vézina, Damian Labuda and their team for the Quebec Regional Reference Sample cohort constitution. LG received funding from the Fonds de recherche du Québec - Santé and the Canadian Institutes of Health Research. CL is the director of the Centre intersectoriel en santé durable (http://www.uqac.ca/santedurable), the chairholder of the Canada Research Chair in the Genomics of asthma and allergic diseases (http://www.chairs.gc.ca) and co-holder of the Chaire en santé durable du Québec (http://www.chairesantedurable.ca).

Author contributions

All authors acquired data and approved the final version of the manuscript. LG and CM played an important role in interpreting the results. LG, CM and SLG conceived and designed the study and drafted the manuscript. CL, designed, built and manages the Saguenay-Lac-Saint-Jean cohort, obtaining funding for genealogical constructs and genomic data acquisition. CL also revised the manuscript.

Ethical approval

This study was approved by the University of Quebec in Chicoutimi (UQAC) ethics board.

Competing interests

The authors declare no competing interests.

Chapitre 2 : Deciphering the genetic structure of the Quebec founder population using genealogies

Avant-propos

L'article *Deciphering the genetic structure of the Quebec founder population using genealogies* présente les résultats de nos travaux de recherche sur la structure de la population du Québec. Cet article a été soumis au European Journal of Human Genetics le 30 septembre 2022. Suivant le processus de révision par les pairs il a été resoumis avec modification le 7 mars 2023 et il a été accepté pour publication le 22 mars 2023. L'article a été publié le 4 avril 2023 en accès libre sous une licence Creative Commons Attribution 4.0 International. Le DOI est le suivant : https://doi.org/10.1038/s41431-023-01356-2. Les figures et tableaux supplémentaires se retrouvent à l'Annexe 2 de ce mémoire. Le tableau supplémentaire 1 se retrouve sur le site de la version web de l'article en raison de sa taille (https://www.nature.com/articles/s41431-023-01356-2).

La création de la cohorte a été effectuée avant le commencement de ce projet, par conséquent, je n'ai pas participé à ces étapes de cette étude. J'ai joué un rôle important dans la conception de ce projet. J'ai effectué l'ensemble des analyses mentionnées et j'ai effectué toutes les figures et les tableaux. J'ai écrit la première version de l'article, et je l'ai amélioré avec les commentaires des co-auteurs. Je me suis également occupée de tout le processus de soumission au journal et j'ai effectué les modifications demandées à la suite de la révision par les pairs.

Deciphering the genetic structure of the Quebec founder population using genealogies

Laurence Gagnon,[1,2] Claudia Moreau,[1,2] Catherine Laprise,[1,2,3] Hélène Vézina,[2,4,5] Simon L. Girard[1,2,6].

1. Département des sciences fondamentales, Université du Québec à Chicoutimi, Saguenay, Québec, G7H 2B1, Canada.

2. Centre Intersectoriel en Santé Durable (CISD), Université du Québec à Chicoutimi, Saguenay, Québec, G7H 2B1, Canada.

3. Centre Intégré Universitaire en Santé et Services Sociaux du Saguenay–Lac-Saint-Jean, Saguenay, Québec, QC G7H 7K9, Canada

4. Département des sciences humaines et sociales, Université du Québec à Chicoutimi, Saguenay, Québec, G7H 2B1, Canada.

5. Projet BALSAC, Université du Québec à Chicoutimi, Saguenay, Québec, G7H 2B1, Canada.

6. Centre de recherche CERVO, Université Laval, Québec, Québec, G1V 0A6, Canada.

Contact Info

Simon Girard, PhD

Université du Québec à Chicoutimi

555, boulevard de l'Université

Chicoutimi (Québec) G7H 2B1

simon2_girard@uqac.ca

Résumé

L'utilisation de la généalogie pour étudier l'histoire démographique d'une population permet d'éliminer les modèles et les hypothèses souvent utilisés en génétique des populations. La population européenne et à effet fondateur du Québec est l'une des rares populations au monde ayant accès à la généalogie complète des 400 dernières années. L'objectif de cette étude est de suivre l'évolution de la structure de la population québécoise au fil du temps, depuis le début de la colonisation européenne jusqu'à nos jours. Pour ce faire, nous avons calculé les coefficients d'apparentement de toutes les paires d'ancêtres dans la généalogie ascendante de 665 sujets issus de huit groupes régionaux et ethnoculturels par période de 25 ans. Nous montrons que la structure de la population québécoise est apparue progressivement dans la vallée du Saint-Laurent dès 1750, avec la distinction des régions du Saguenay et de Gaspésie. À cette époque, les ancêtres de deux groupes, les Saguenéens et les Acadiens de la péninsule de la Gaspésie, ont connu une augmentation marquée des niveaux d'apparentement et de consanguinité, ce qui a façonné la structure génétique contemporaine. Il est intéressant de noter que cette structure est apparue avant la colonisation de la région du Saguenay et dès le tout début du peuplement de la péninsule de la Gaspésie. Ces effets fondateurs régionaux ont entrainé des différences dans le partage de segments identique-par-descendance, les groupes de la Gaspésie et de la Côte-Nord partagent davantage de segments plus longs, tandis que les Saguenéens partagent davantage de segments plus courts. Cela se reflète également dans la distribution du nombre d'ancêtres communs les plus récents à différentes générations et de leur contribution génétique aux sujets étudiés.

Abstract

Using genealogy to study the demographic history of a population makes it possible to overcome the models and assumptions often used in population genetics. The Quebec founder population is one of the few populations in the world having access to the complete genealogy of the last 400 years. The goal of this study is to follow the evolution of the Quebec population structure over time from the beginning of European colonization until the present day. To do so, we calculated the kinship coefficients of all ancestors' pairs in the ascending genealogy of 665 subjects from eight regional and ethnocultural groups per 25-year period. We show that the Quebec population structure appeared progressively in the St. Lawrence valley as early as 1750 with the distinction of the Saguenay and Gaspesian groups. At that time, the ancestors of two groups, the Sagueneans and the Acadians from the Gaspé Peninsula, experienced a marked increase in kinship and inbreeding levels which have shaped the structure and led to the contemporary population structure. Interestingly, this structure arose before the colonization of the Saguenay region and at the very beginning of the Gaspé Peninsula settlement. The resulting regional founder effects in these groups led to differences in the present-day identity-by-descent sharing, the Gaspé and North Shore groups sharing more large segments and the Sagueneans more short segments. This is also reflected by the distribution of the number of most recent common ancestors at different generations and their genetic contribution to the studied subjects.

Introduction

Founder populations have been particularly helpful in demonstrating how past demographic events have shaped present-day genetic structure and its consequences on human health [1,2,3]. Studying the past demographic history of a population often relies on current genetic data and models of ascending genealogical trees [4, 5]. However, developing efficient methods for inferring the underlying genealogy has proved challenging [6, 7] or requires lots of contemporary and ancient genomes data [8]. To avoid using such assumptions, one would need the complete genealogy of the population. Few populations in the world have access to such genealogical data [9,10,11]. The Quebec province of Canada relies on the BALSAC population register, a large collection of linked data from parish records, to reconstruct the genealogy of the vast majority of Quebecers, mostly of French Canadian descent, but also of other origins, since the foundation of the colony in the 17th century until recent times [12]. This invaluable resource allows the detailed mapping of the population structure over time. Indeed, it has been shown that the genealogical lines covering the last 400 years explain most of the present-day genetic structure of the Quebec population [13, 14].

This study will focus on Quebecers genealogically anchored into five regions (from west to east): the Montreal and Quebec City areas, the Saguenay-Lac-St-Jean (Saguenay) and North Shore regions, and the Gaspé Peninsula (Gaspé) where four subgroups were sampled (Acadians, French Canadians, Loyalists, and Channel Islanders). Most Quebecers of French Canadian ancestry are descendants of around 8500 settlers who came predominantly from France between 1608 and 1760 [15]. These European newcomers first settled in Quebec City (1608) and Montreal (1642) which are now two major urban regions (Supplementary Fig. S1A) and along the shores of the St.Lawrence river. Following the British Conquest of 1760, French immigration decreased dramatically, and the French-speaking population expanded mostly

through natural increase. Population growth led to the colonization of new regions, including more remote and isolated regions, favoring population subdivision [13].

Permanent European settlement in Gaspé began during the second half of the 18th century with the arrival of Acadians, who escaped deportation by the British [16]. They were soon joined by English-speaking United Empire Loyalists who chose to remain under British rule after the American Declaration of Independence in 1776. From 1830–1840, many Quebecers of French Canadian ancestry from the lower part of the St. Lawrence valley also settled in the Gaspé peninsula [16]. At the same time, a fourth group, inhabitants of the Channel Islands, came to Gaspé for the fishing industry.

The settlement of Saguenay started in 1838 with founders mostly coming from the neighboring region of Charlevoix which was colonized earlier by the end of the 17th century. The Saguenay population size underwent a 25-fold increase between 1861 and 1961, mostly due to a high birth rate [17, 18], while the whole Quebec population increased only 5-fold. The western part of the North Shore was colonized by ancestors who came from the Charlevoix and Bas-St-Laurent regions [19] while the eastern part pioneers were mostly fishermen from Iles-de-la-Madeleine and Gaspé.

Genetic data is often used to study the contemporary population structure [20] and we have previously shown that the genetic structure of the Quebec population is well correlated with the one inferred using genealogical measures [13, 21, 22]. However, genealogies are an invaluable tool to study how the population structure was shaped in the past generations. The goal of this study was to follow the evolution of the Quebec regional population structure from its colonization until the present day. To do so, we looked at the kinship and inbreeding levels

of all ancestors in the genealogies. We also deciphered the population fine structure inferred with present-day genetic identity-by-descent (IBD) sharing using genealogical measures.

Subjects and Methods

This study was approved by the University of Quebec in Chicoutimi (UQAC) ethics board. Written informed consent was obtained from all adult participants.

Cohort

The data consist of 579 subjects from the Quebec Regional Reference Sample and 86 unaffected subjects from the Saguenay-Lac-St-Jean asthma familial cohort (Supplementary Fig. S1A and Table 1) [13, 21, 23]. The subjects are distributed in five regions and eight groups (based on geographical and ethnocultural criteria) of the province of Quebec: the Montreal and Quebec City areas, the Saguenay and North Shore regions, and the Gaspé Peninsula. For the latter, four subgroups were sampled, namely Acadians, French Canadians, Loyalists, and Channel Islanders. Subjects were sampled regardless of their proportion in the population. To ensure regional connection, subjects needed to have their four grandparents born in the Quebec province and one or two parents born in the particular region except for the Montreal area where only first criterion (the four grandparents) was applied. The four ethnocultural subgroups of Gaspé were self-reported. A strong correspondence between ancestral origins traced in genealogies and self-reported origins was found in a previous study [24]. Genotyping and ascending genealogical data are available for the 665 subjects.

Genotyping data and genomic analyses

Genotyping of the 665 subjects was conducted on Illumina Omni Express (~740,000 SNPs) and Illumina Omni 2.5 chips (~2.5 M SNPs) chips. Both chips have been merged to keep only common SNPs (702,216). Quality control filters were applied at the individual and SNP levels

using PLINK software v1.9 [25]. We retained subjects with at least 98% genotypes among all SNPs. At the SNP level, we retained SNPs with at least 98% genotypes among all subjects, located on the autosomes and in Hardy–Weinberg equilibrium $p > 0.001$ (calculated on the whole cohort), yielding 659,219 SNPs. Closely related subjects (first cousins, kinship coefficient ≥ 0.0625) were eliminated to avoid bias in the population structure analysis, yielding a final sample size of 665 subjects (Table 1). A principal component analysis (PCA) was performed on SNPs with a minor allele frequency of at least 5% and after pruning to remove SNPs in LD (96,915 SNPs left) using PLINK software to confirm that our final dataset reflects the previously described Quebec population structure [13, 21] (Supplementary Fig. S2).

The assessment of pairwise IBD segments was performed using refinedIBD software v17Jan20 [26] on phased genotypes (done using Beagle software version 18May20.d20). This software was selected for its power and accuracy in detecting IBD segments [20, 26]. Only segments of 2 cM or more and with a LOD score greater than 3 were retained.

Table 1. Regional and ethnocultural contemporary groups' description.

Group	Abbreviation	Sample Size	Average year of Parents' Marriage
Montreal Area	MTL	138	1950
Quebec City Area	QUE	70	1945
Saguenay-Lac-Saint-Jean	SAG	86	1977
North Shore	NSH	47	1949
Gaspé French Canadians	GFC	97	1945
Gaspé Loyalists	GLO	71	1939
Gaspé Channel Islanders	GCI	67	1941
Gaspé Acadians	GAC	89	1942

Genealogical data and analyses

Genealogical data were obtained through the BALSAC project [12]. Ascending genealogies were reconstructed for the 665 (contemporary) subjects for whom we have genotype data with average completeness of at least 60% up to the tenth generation for all groups, except for the Loyalists and Channel Islanders of Gaspé (explained in part by their later time of arrival in Quebec), consistent with previous results [13] (Supplementary Fig. S3). The completeness is the proportion of ancestors found at each generation in the genealogy compared to the maximum possible number of ancestors. Information on the parents' year (±5 years for confidentiality concerns) and region of marriage or if outside Quebec, country of origin was obtained for 94,076 distinct individuals (ancestors and subjects) throughout the genealogy. There are 20 regions of marriage in our data (Supplementary Fig. S1B). We inferred the unknown parents' marriage years as being the children or grandchildren parents' marriage year minus 30 or 60 years, the average time between parents' marriage and their children's marriages in our data being 32 years. They were grouped into 25-year periods to minimize the parent–child overlap within the same period (2.5% overlap in the 1676–1925 period),

The kinship coefficient at the maximum generational depth was computed using the R GENLIB library v1.1.6 [22] for each pair of ancestors within 25-year periods. Multidimensional scaling (MDS) was performed on the pairwise kinship distance matrix (1-kinship coefficient). Ancestors who had no kinship ties with anybody were removed from this analysis (either founders who have no known parents in the genealogy or ancestors close to founders such as their children or grandchildren who could not be linked to anybody else in the genealogy). For this analysis, the parents' marriage region was assigned directly to each ancestor and colored according to the 20 regions in Fig. S1B.

The average pairwise kinship and inbreeding coefficients for the ancestors of each group were calculated at the maximal depth for each period. In this analysis we did not assign the region according to the parents' marriage place of each ancestor, rather, we assigned the ethnocultural or regional group to all ancestors of the contemporary subjects of each group. Consequently, ancestors could be assigned to many groups if they happened to be present in the genealogies of subjects from different groups. In the genealogies, kinship is measured on each pair of individuals (ancestors whose parents were married in each period in our case) and inbreeding is measured within one ancestor and is equal to the kinship coefficient of their parents. Consequently, within the same period of time, the mean kinship for all pairs of ancestors will not necessarily reflect the mean inbreeding for these ancestors since two persons have to mate to produce a child with a certain level of inbreeding. Of course, mating will not happen between all pairs of ancestors in the population. The inbreeding reflects the mating pattern and not necessarily the mean kinship of the population, especially if mating is not random.

Most recent common ancestors (MRCAs) were counted using GENLIB for the subjects' pairs within groups for each distance in meioses (for example, there are four meioses between two cousins). The minimal distance (the shortest genealogical path) relating both subjects through the MRCA was calculated using GENLIB. A pair can have more than one MRCA as long as no ancestor in the set of MRCAs shares a descendant who is also an ancestor of the pair of subjects. The expected genetic contribution (GC), consisting in summing the transmission probabilities over all genealogical paths connecting an ancestor to a descendant given that parents transmit half of their genome to each child, was also calculated for each MRCA to both descendants using GENLIB. The GC product to both subjects was summed over all MRCAs. Groups were resampled down to 47 subjects 1000 times to avoid size bias. For each bootstrap, we randomly selected 47 subjects in each group and reconstructed the genealogy for this new

subjects' subset. This results in less ancestors in the genealogy and is essential to consider when counting the absolute number of ancestors which depends on the number of subjects.

Results

How population structure was shaped

The evolution of the Quebec population structure was assessed using pairwise kinship coefficients of all ancestors at the maximum depth for each period (Fig. 1) in the ascending genealogies of the 665 subjects all together. In this figure, we used the parents' marriage region (Supplementary Fig. S1B) to color each dot (representing an ancestor whose parents married in that period). Before 1750, the ancestors from three regions (Côte-de-Beaupré, Côte-du-Sud, and Charlevoix), progressively differentiated from each other along the x axis and from another group of immigrants along the y axis, whose parents did not marry in Quebec but whose country of origin is known (see interactive Fig. 1 for countries of origin). These immigrants are mainly coming from Acadia (83%) and a lower proportion from France (5%) and other origins. The present-day population structure (Supplementary Fig. S2) [13] progressively appears as early as 1751–1775 distinguishing Charlevoix and the ancestors of the Gaspé groups (Fig. 1). Note that Gaspé ethnocultural groups can't be distinguished from the data used in Fig. 1. Prior to 1750, ancestors of each Gaspé group were found almost everywhere on the MDS (Supplementary Fig. S4). Following this period, we can see the effect of the Acadians and Loyalists immigration by the multiplication of ancestors specific to only one Gaspé group and their progressive differentiation along the y axis (Supplementary Fig. S4). By 1826–1850, the first ancestors married in Saguenay appeared and the population structure at that time was very similar to the one depicted on the PCA of the present-day subjects (Supplementary Fig. S2) [13].

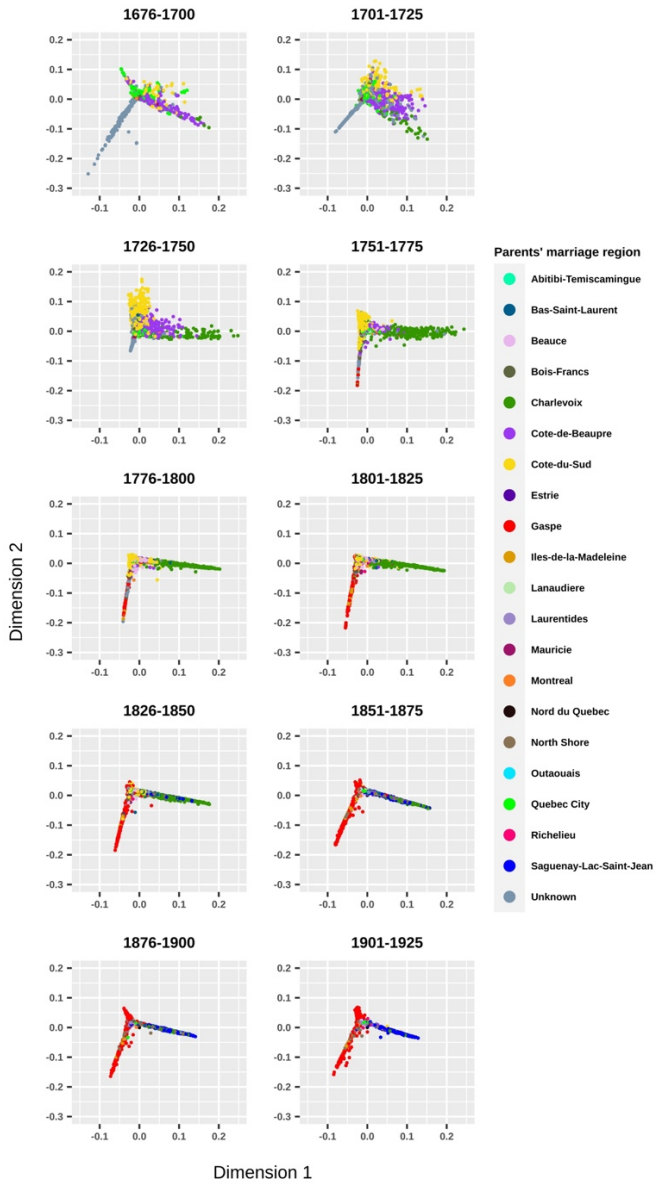


Fig. 1. Multidimensional scaling (MDS) of the pairwise kinship coefficients of ancestors of all groups per 25-year period.

MDS was performed on the pairwise kinship distance matrix, (i.e., 1-kinship coefficient) of ancestors whose parents were married at each period. The pairwise kinship coefficient was computed using the R GENLIB library at the maximal depth. The external interactive version of this figure is available at https://laugag17.github.io/quebec_founder_pop_interactive_figure/figure2_interactive_graph.html.

Mean kinship and inbreeding over time

We averaged for ancestors of each contemporary group the genealogical kinship and inbreeding coefficients at the maximum depth for each period (based on the parents' marriage date) (Fig. 2A and Supplementary Table S1 for counts). From 1750, the GAC and Saguenay ancestors went through a marked increase in averaged kinship compared to the ancestors of the other groups. By 1825, the GAC ancestors' mean kinship had continued to increase while the Saguenay ancestors had reached a plateau. The ancestors' mean kinship increase in GAC and Saguenay groups was accompanied by an increase in inbreeding (Fig. 2B). Until 1850, the average inbreeding coefficient at the maximum depth was higher for the Saguenay ancestors. After 1850, the GAC ancestors' mean inbreeding exceeded the one of the Saguenay ancestors, whereas the latter reached a plateau. The three other Gaspé groups' average inbreeding coefficient increase was slower at the beginning, but almost reached the Saguenay ancestors' inbreeding value by 1925. Interestingly, for all groups the increase in inbreeding was more important than the increase in kinship levels.

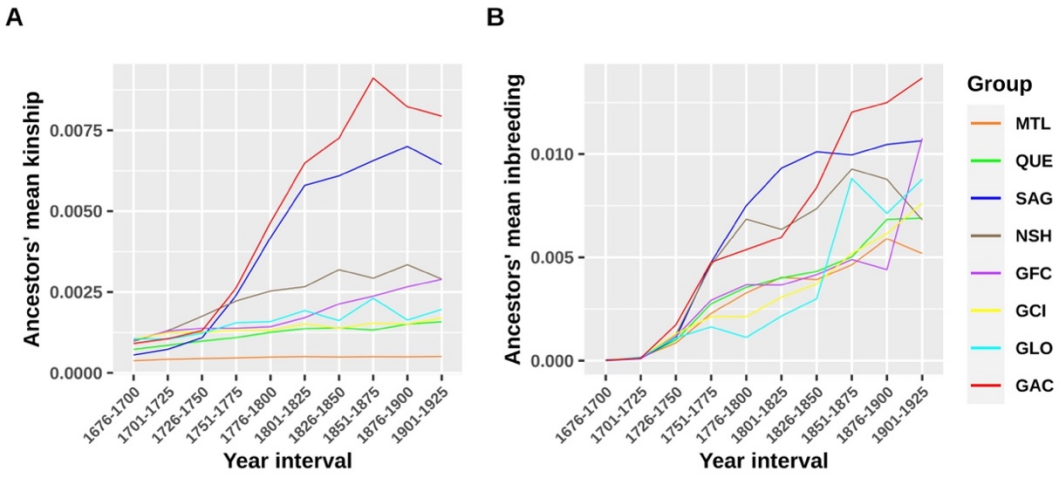


Fig. 2. Average kinship (A) and inbreeding (B) coefficients of ancestors of each group per 25-year period.

The average pairwise kinship and inbreeding coefficients for the ancestors of each group were calculated at the maximal depth for each period. The sample sizes are reported in Supplementary Table S1.

GAC=Gaspé Acadians ; GCI=Gaspé Channel Islanders ; GFC=Gaspé French Canadians ; GLO=Gaspé Loyalists ; MTL=Montreal ; NSH=North Shore ; QUE=Quebec City ; SAG=Saguenay.

IBD sharing and most recent common ancestors

For each group, we plotted the mean number of IBD segments shared among subjects' pairs per segment length (bins of 5 cM) and compared this with the cumulative MRCA counts per meiosis (Fig. 3A, B). Note that MRCAs are not unique so the same MRCAs can appear for many subjects' pairs and they will be counted each time they appear. These two metrics, one using genetic data and the other using genealogical data, show very similar patterns. Interestingly, the Sagueneans' pairs shared more IBD segments of short lengths (<22 cM), but less of long lengths (>37 cM) compared to the North Shore and the four Gaspé groups leaving only the urban and older groups (Montreal and Quebec City areas) behind for longer segments.

This is also reflected by the cumulative MRCA count until ten meioses (Supplementary Table S2). Inversely, the Gaspé Loyalists (GLO) shared less short segments and more long segments. Note that the MRCA count for the GLO is biased towards the right of the graph (Fig. 3B) since their completeness decreases faster than the other groups (Supplementary Fig. S3). This was also observed for the Gaspé Channel Islanders to a lesser extent.

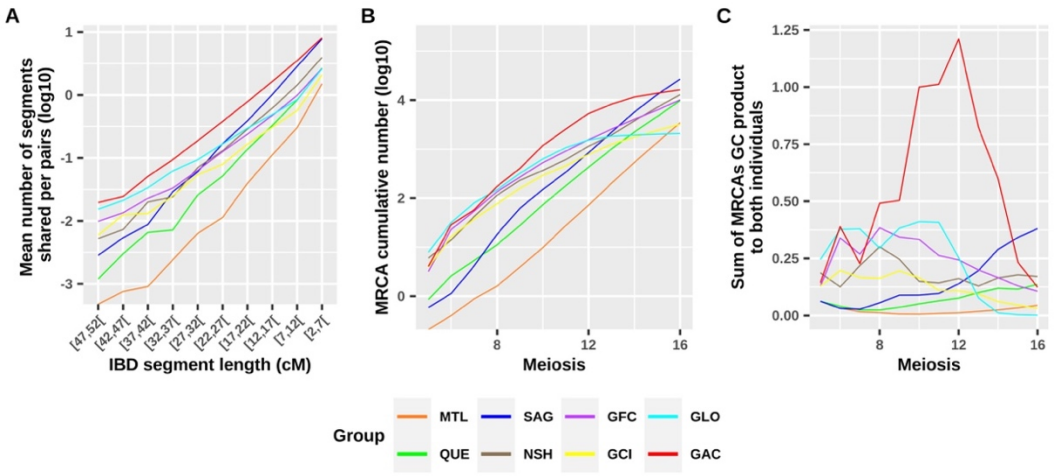


Fig. 3. Within groups IBD segment sharing by length (A) as well as MRCA cumulative count (B) and genetic contribution (C) per meiosis.

IBD segment lengths were binned into 5cM intervals. The sample sizes for (A) are reported in Table 1. The MRCA cumulative count (B) and the sum (not cumulative) of their genetic contribution to the contemporary subjects (C) are the averages of 1,000 bootstraps of 47 subjects. Fig. S5 presents the cumulative MRCA counts (B) and genetic contribution (C) until 30 meioses and with the bootstrap intervals.

GAC=Gaspé Acadians ; GCI=Gaspé Channel Islanders ; GFC=Gaspé French Canadians ; GLO=Gaspé Loyalists ; MTL=Montreal ; NSH=North Shore ; QUE=Quebec City ; SAG=Saguenay.

We also calculated in the genealogies the product of the genetic contributions of each MRCA to both subjects and we summed these products for all MRCAs. Figure 3C presents this GC sum averaged for 1000 bootstraps of 47 subjects (see also Supplementary Fig. S5 for intervals). Note the very high genetic contribution of GAC close MRCAs. For the three other Gaspé groups and the North Shore, closer MRCAs also had a higher GC than those of Montreal, Quebec City, and Saguenay. However, for the latter, the GC sum is higher for more distant MRCAs.

Discussion

In this study, we show how the Quebec founder's population structure was shaped over time. We found that the previously described structure differentiating the Gaspé and the Saguenay groups [13] emerged early in the colonization process (1750 or before), almost a hundred years before the colonization of the Saguenay region (1840) [17, 18, 27,28,29] (Fig. 1). At this time, Saguenay ancestors were mostly located in the Charlevoix region where they had established only two or three generations before. The small number of founding families in Charlevoix followed by a rapid expansion in Saguenay in the 19th century led to changes in the frequencies of alleles and diseases [30]. Similarly, GAC subjects descend from a small number of founding families, but they did not go through a rapid expansion like the Saguenay region. Instead, they mostly married inside their community due to linguistic and cultural barriers present with the other Gaspé groups [31]. Additionally, they were the only ones in the area until 1780 [16]. Both groups' colonization started with a limited number of founding families implicating that a smaller number of founders explains a higher proportion of the present-day gene pool compared to the other groups [22]. Despite their different subsequent colonization processes, Saguenay and GAC ancestors have a very similar mean kinship increase (Fig. 2A) starting in 1750 when the contemporary population structure appeared. For both Saguenay and GAC ancestors, spouses had higher chances of being related than those of the other groups. However,

the average inbreeding was higher among Saguenay ancestors until 1850 when GAC ancestors went through a marked increase which lasted until recent times (Fig. 2B). This is consistent with previous findings showing that close inbreeding of contemporary subjects is the lowest in the province for Sagueneans and the highest for Gaspé [27] (Supplementary Fig. S6). In fact, for both kinship and inbreeding, the Saguenay ancestors reached a plateau at the time the region was colonized (1838) and the expansion started (around 1860) while GAC never went through such a rapid expansion. This increase in inbreeding followed by stabilization among Saguenay ancestors was previously explained by the evolution of nonrandom mating as well as by the evolution of inbreeding resulting from drift [32]. This would need further investigation to understand its implications on the contemporary population. Nevertheless, the GAC and the other Gaspé ethnocultural groups did not reach such a plateau.

The regional fine structure could be observed within groups by comparing IBD sharing patterns (Fig. 3A). To ensure that this fine structure is explained by the recent population history (after the European colonization of Quebec), we focused on large IBD segments which are expected to come from more recent ancestors [33, 34]. The GAC and Saguenay groups share more IBD segments <22 cM than the other groups, in line with previous results on shorter segments [21] and consistent with the higher ancestors' mean kinship and inbreeding compared to other groups by 1750. However, for long segments >37 cM, the IBD sharing of Sagueneans' pairs decreases more rapidly than the one of Gaspé and North Shore groups. Note that the North Shore sampling in the present study was extended to both eastern and western parts compared to a previous analysis which focused more on the western part [13]. This gives us a higher resolution and reveals differences that were not seen before since both parts have had a different colonization process. The observed IBD pattern is explained by the recent MRCA counts from five to ten meioses (Fig. 3B and Supplementary Table S2) which are more numerous in the Gaspé and North Shore groups than in the Sagueneans. In other words, the Sagueneans have

less recent, but more distant common ancestors than other eastern groups. This is consistent with the close inbreeding being less important in Saguenay than among the Gaspé and the North Shore contemporary subjects (Supplementary Fig. S6) [27]. As MRCA can appear many times in this analysis and even though the Saguenean subjects descend from fewer founders than the other groups except GAC, they have more numerous MRCAs (after 13 meioses), which means that some of them appear very often, consistent with an expanding population due to a high birth rate and also observed in previous studies [29]. Another interesting ethnocultural group is the Gaspé Loyalists (GLO) which is genetically different from other Quebec groups on the second and the third PC (Supplementary Fig. S2). The GLO is among the groups with the lowest mean number of short IBD segments shared per pair, but they reach the second highest mean number of pairs sharing IBD segments (after the GAC group) above 27 cM. Indeed, GLO ancestors did not undergo a marked kinship increase around 1750 like Saguenay and GAC ancestors, they show an inbreeding increase after 1850 (Fig. 2B) which is consistent with the highest inbreeding values found at the 6–7 generations for GLO subjects (Supplementary Fig. S6) and corresponds to the appearance of the Gaspé groups differentiation (Supplementary Fig. S4). The GLO group comes from more numerous and diversified founders compared with GAC which could have affected the sharing of short IBD segments [35]. After their settlement, the GLO ancestors have remained quite isolated for more than 150 years as shown by their rapid inbreeding increase after 1850 (Fig. 2B), which could have exacerbated their sharing of long IBD segments. This is again consistent with previous findings on paternal and maternal lineages [31] and explains the particular IBD sharing among GLO subjects.

In this study, we show a similar pattern for the shared IBD segment lengths' distribution and the cumulative number of genealogical MRCAs per subject pair (Fig. 3AB). Shared IBD segment length distribution depends on the number of common ancestors and the distance connecting both subjects to their common ancestor as it has been shown before using

simulations on two individuals' pairs [22]. The chance of transmitting a segment also depends on the GC of the common ancestor to both descendants [22]. Thus, common ancestors who contributed a lot to the present-day gene pool would be more susceptible to transmitting an IBD segment than those who contributed less. Usually, the closer the ancestors are to their descendants, the bigger their GC is (Fig. 3C). But in Saguenay, there are unusually great contributors among distant ancestors [29]. The number of MRCAs above 20 meioses is in the same order of magnitude for Saguenay as for Montreal and Quebec City subjects. However, the Saguenay ancestors' GC above 10 meioses was higher and the resulting IBD sharing of shorter segments (less than 22 cM) is also higher. Sagueneans share more segments of less than 22 cM than any other group except GAC. In turn, GAC close common ancestors have a larger GC and they have the highest IBD sharing for all length bins even if their close MRCA counts (until 8 meioses) are similar to the other Gaspé groups and the North Shore subjects, suggesting that close MRCAs might have transmitted not only long, but also short IBD segments to their descendants.

Some limitations are present in this work. The genealogical completeness is not consistent across all groups (Supplementary Fig. S3) but was left uncorrected to retain two groups of Gaspé that would have been filtered out otherwise (Gaspé Loyalists and Channel Islanders) [13, 21]. This explains the aberrant curves in Fig. 3B, especially for Loyalists' MRCA counts above 12 meioses. Also, note that if a structure was already present in the Quebec founders (for whom we reach the limit of the genealogy and we don't know the parents), we do not have this information and we are unable to interpret its impact on the present-day population structure. We also reported an unequal number of participants across groups (Table 1). To overcome this, a bootstrap method was performed for specific genealogical analyses (Fig. 3BC). Finally, a generation gap was present between the Saguenay and the other groups (Table 1) that was not accounted for in Fig. 3 since the IBD sharing to be compared with genealogical

MRCAs also includes this generation gap. For the other genealogical analyses presented in this study, this was not relevant since ancestors of each period were grouped regardless of the subjects' generation.

In conclusion, genealogies are an invaluable tool to study the evolution of the population structure over time and to understand how the present-day genetic structure was shaped. We have shown that the Quebec population structure subdividing the Saguenay and Gaspé (especially Acadians and Loyalists) groups appeared early in the history of the province, even before the colonization of the Saguenay region. At that time, the Saguenay and GAC groups both experienced a marked average kinship and inbreeding increase until more recent times, when Saguenay reached a plateau and was almost joined by the other Gaspé groups. The resulting strong founder effect that occurred led to differences in the present-day IBD sharing and is linked to less numerous recent, but more numerous distant MRCAs for the Sagueneans compared to the GAC. Another understudied group, the GLO, was shown to have numerous recent MRCAs resulting in higher sharing of long IBD segments compared to all other groups except GAC. However, as their founders were more numerous and diversified and also due to the lower genetic contribution of their close MRCA, they did not go through a kinship increase as the Saguenay and GAC groups around 1750, but they did later and their resulting founder effect is less striking.

Data Availability

The 665 subjects' genealogies and genotypes are available upon request to BALSAC at https://balsac.uqac.ca (12).

Code Availability

The code used for this study can be found in the following GitHub repository: https://github.com/laugag17/quebec_founder_pop.

References

1. Kere J. Human population genetics: lessons from Finland. Annu Rev Genomics Hum Genet. 2001;2:103–28.
2. Scriver CR. Human Genetics : Lessons from Quebec Populations. Annual Review of Genomics and Human Genetics. 2001;2:69–101.
3. Locke AE, Steinberg KM, Chiang CWK, Service SK, Havulinna AS, Stell L, et al. Exome sequencing of Finnish isolates enhances rare-variant association power. Nature. 2019;572(7769):323–8.
4. Marchi N, Schlichta F, Excoffier L. Demographic inference. Current Biology. 2021;31(6):R276–9.
5. Tournebize R, Chu G, Moorjani P. Reconstructing the history of founder events using genome-wide patterns of allele sharing across individuals. PLOS Genetics. 2022 juin;18(6):e1010243.
6. McVean GAT, Cardin NJ. Approximating the coalescent with recombination. Philosophical Transactions of the Royal Society B: Biological Sciences. 2005;360(1459):1387–93.
7. Rasmussen MD, Hubisz MJ, Gronau I, Siepel A. Genome-Wide Inference of Ancestral Recombination Graphs. PLoS Genetics. 2014;10(5):e1004342.
8. Wohns AW, Wong Y, Jeffery B, Akbari A, Mallick S, Pinhasi R, et al. A unified genealogy of modern and ancient genomes. Science. 2022 Feb 25;375(6583):eabi8264.
9. Helgason A, Hrafnkelsson B, Gulcher JR, Ward R, Stefánsson K. A Populationwide Coalescent Analysis of Icelandic Matrilineal and Patrilineal Genealogies: Evidence for a Faster Evolutionary Rate of mtDNA Lineages than Y Chromosomes. The American Journal of Human Genetics. 2003;72:1370–88.

10. Pettay JE, Kruuk LEB, Jokela J, Lummaa V. Heritability and genetic constraints of life-history trait evolution in preindustrial humans. Proceedings of the National Academy of Sciences of the United States of America. 2005;102(8):2838–43.
11. Pluzhnikov A, Nolan DK, Tan Z, McPeek MS, Ober C. Correlation of intergenerational family sizes suggests a genetic component of reproductive fitness. American Journal of Human Genetics. 2007;81(1):165–9.
12. BALSAC. BALSAC. Available from: https://balsac.uqac.ca/
13. 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. Human Genetics. 2011 May;129(5):521–31.
14. Anderson-Trocmé L, Nelson D, Zabad S, Diaz-Papkovich A, Baya N, Touvier M, et al. On the Genes, Genealogies, and Geographies of Quebec. bioRxiv. 2022 Jan 1;2022.07.20.500680.
15. Charbonneau H, Desjardins B, Légaré J, Denis H. The population of the St-Lawrence Valley, 1608-1760. In: A Population History of North America. 2000. p. 99–142.
16. Desjardins M, Frenette Y, Bélanger J. Histoire de la Gaspésie. Éditions de l'IQRC. Revue d'histoire de l'Amérique française. Sainte-Foy, Québec; 1999. 797 p.
17. Pouyez C, Lavoie Y. Les Saguenayens. Introduction à l'histoire des populations du Saguenay. Presse de l'UQ. 1983. 386 p.
18. Jette R, Gauvreau D, Guérin M. Aux origines d'une région: le peuplement fondateur de Charlevoix avant 1850. In: Histoire d'un génome Population et génétique dans l'est du Québec Québec: Presses de l'Université du Québec. 1991. p. 75–106.
19. Frenette P. Histoire de la Côte-Nord. Sainte-Foy, Québec: Institut québécois de recherche sur la culture; 1996. 667 p. (Collection Les régions du Québec).

20. Nait Saada J, Kalantzis G, Shyr D, Cooper F, Robinson M, Gusev A, et al. Identity-by-descent detection across 487,409 British samples reveals fine scale population structure and ultra-rare variant associations. Nat Commun. 2020 Nov 30;11(1):6130.

21. Gauvin H, Moreau C, Lefebvre JF, Laprise C, Vézina H, Labuda D, et al. Genome-wide patterns of identity-by-descent sharing in the French Canadian founder population. Eur J Hum Genet. 2014 Jun;22(6):814–21.

22. Gauvin H, Lefebvre JF, Moreau C, Lavoie EM, Labuda D, Vézina H, et al. GENLIB: an R package for the analysis of genealogical data. BMC Bioinformatics. 2015 May 15;16(1):160.

23. Laprise C. The Saguenay-Lac-Saint-Jean asthma familial collection: the genetics of asthma in a young founder population. Genes Immun. 2014 Apr;15(4):247–55.

24. Vézina H, Tremblay M, Lavoie ÈM, Labuda D, Stringer L. Concordance Between Reported Ethnic Origins and Ancestral Origins of Gaspé Peninsula Residents. Population (English Edition, 2002-). 2014;69(1):7–27.

25. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet. 2007 Sep;81(3):559–75.

26. Browning BL, Browning SR. Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics. 2013 Jun;194(2):459–71.

27. Vézina H, Tremblay M, Houde L. Mesures de l'apparentement biologique au Saguenay-Lac-St-Jean (Québec, Canada) à partir de reconstitutions généalogiques. Annales de démographie historique. 2004;108:67–83.

28. Lavoie EM, Tremblay M, Houde L, Vézina H. Demogenetic study of three populations within a region with strong founder effects. Community Genet. 2005;8(3):152–60.

29. Bherer C, Labuda D, Roy-Gagnon MH, Houde L, Tremblay M, Vézina H. Admixed ancestry and stratification of Quebec regional populations. American Journal of Physical Anthropology. 2011;144:432–41.

30. Bouchard G, De Braekeleer M. Histoire d'un genome: Population et genetique dans l'est du Quebec. Sillery, Québec: Presses de l'Université du Québec; 1991. 607 p.

31. Moreau C, Vézina H, Yotova V, Hamon R, Knijff PD, Sinnett D, et al. Genetic heterogeneity in regional populations of Quebec - Parental lineages in the Gaspe Peninsula. American Journal of Physical Anthropology. 2009 Aug;139(4):512–22.

32. Mourali-Chebil S, Heyer E. Evolution of inbreeding coefficients and effective size in the population of Saguenay Lac-St.-Jean (Quebec). Hum Biol. 2006 Aug;78(4):495–508.

33. Browning SR, Browning BL. Identity by descent between distant relatives: detection and applications. Annu Rev Genet. 2012;46:617–33.

34. Ralph P, Coop G. The Geography of Recent Genetic Ancestry across Europe. PLOS Biology. 2013 mai;11(5):e1001555.

35. Matthews GJ, Gentilcore RL. Historical Atlas of Canada [Internet]. University of Toronto Press; 1993. 184 p. Available from: http://www.jstor.org/stable/10.3138/9781442675759.

Acknowledgments

This work was supported by funding from the Canadian Institutes of Health Research (#420021) in the SLG lab. It was also made possible by the Digital Research Alliance of Canada which provided access to storage and computing resources. We are extremely grateful to all participants in this research. We would like to thank Damian Labuda and his team for the Quebec Regional Reference Sample cohort constitution. LG received funding from the Fonds de recherche du Québec - Santé, the Fonds de recherche du Québec - Nature et technologies, and the Canadian Institutes of Health Research. CL is the director of the Centre intersectoriel

en santé durable (http://www.uqac.ca/santedurable) and the chairholder of the Canada Research Chair in the Environment and Genetics of Respiratory Diseases and Allergy (http://www.chairs.gc.ca).

Author contributions

All authors acquired data and approved the final version of the manuscript. LG and CM played an important role in interpreting the results. LG, CM and SLG conceived and designed the study and drafted the manuscript. CL and HV revised the manuscript.

Funding

This work was supported by funding from the Canadian Institutes of Health Research (#420021).

Competing interest

The authors declare no competing interests.

Ethical approval

This study was approved by the University of Quebec in Chicoutimi (UQAC) ethics board.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41431-023-01356-2.

Correspondence and requests for materials should be addressed to Simon L. Girard.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Chapitre 3 : Discussion

Les populations à effet fondateur ont été très étudiées afin d'analyser les variants associés à des maladies, mais ces dernières permettent également de mieux comprendre les impacts des phénomènes démographiques sur la structure génétique d'une population. Cette discussion est un retour sur les objectifs spécifiques et communs des deux chapitres précédents. Les limitations et perspectives de ce mémoire seront également discutées.

3.1 Retour sur les objectifs

3.1.1 Chapitre 1

L'article *Fine-scale genetic structure and rare variant frequencies,* qui se retrouve au Chapitre 1, démontre la présence de structure fine au sein de certaines populations à effet fondateur. L'objectif de ce volet était de décrire correctement la structure fine d'une population et de mieux comprendre l'implication des structures de populations dans l'identification de variants rares. L'investigation de ces structures démontre qu'elles sont essentielles à prendre en compte afin de concentrer des variants rares et ainsi augmenter leur fréquence, facilitant leur identification.

L'exploration des effets fondateurs régionaux a déjà été réalisée au Québec[49,57,60,67,68], avec la différenciation du Saguenay-Lac-Saint-Jean et de la Gaspésie comparativement à l'ensemble de la population. Au niveau des Juifs Ashkénazes, certaines structures génétiques ont également été identifiées[91,104,105]. Toutefois, toutes ces structures ne sont pas nécessairement prises en compte dans les larges études d'association, ce qui peut résulter en biais et en faux positifs. Les études d'associations de variants communs utilisent des méthodes, comme des PCA ou des modèles linaires mixtes, pour corriger l'ascendance et les effets de structure de population[106–108]. Néanmoins, ces méthodes ne sont pas suffisantes. En effet, une récente étude de Gouveia *et al.* 2023 a révélé la présence de métissage chez les Européens et son impact sur

les associations génétiques. En ajustant pour l'ascendance à l'échelle du génome et pour l'ascendance spécifique au locus, ils ont déterminé que des associations déjà identifiées et publiées étaient finalement fausses. Pour ce qui est des études d'associations de variants rares, l'impact de la structure de population est encore plus important, surtout lorsque cette structure est causée par des changements démographiques récent[107]. Des méthodes de PCA de variants rares, de IBD ou d'ascendance spécifique au locus sont appropriées pour corriger l'ascendance et les structures fines de population[106,109]. Au-delà de corriger pour ces structures, ce projet a pour but de déterminer une façon d'effectuer des associations génétiques en utilisant la structure fine de population. Nous croyons que cela réside dans la connaissance approfondie de nos populations, autant à effet fondateur ou non.

Ainsi, l'objectif de ce projet est atteint. La prise en compte de la structure génétique fine, qui malheureusement n'est pas toujours investiguée, a permis de mieux comprendre les structures présentes au sein d'une population. En effet, au Québec, il serait plus adéquat de séparer la province en différents groupes afin d'effectuer des études d'associations génétiques puisque la structure fine présente au sein des groupes du Saguenay-Lac-Saint-Jean et de la Gaspésie pourrait mener à des biais. À cet effet, nous proposons donc une approche inverse aux études génétiques actuelles en voulant mettre en valeur les effets fondateurs ou structures populationnelles présentes au sein de grandes cohortes. L'utilisation d'une sous-partie de ces cohortes avec une structure bien définie, comportant moins d'individus, faciliterait l'identification de variants rares en raison de leur augmentation en fréquence. De ce fait, cela met en valeur des cohortes existantes en augmentant leur rentabilité et en diminuant le besoin de générer de nouveaux ensembles de données. Il serait intéressant de vérifier cette approche au Québec en utilisant une grande cohorte génétique et populationnelle, comme celle de CARTaGENE[110].

3.1.2 Chapitre 2

Ce chapitre a pour but d'analyser en profondeur la structure d'une population à effet fondateur présenté au Chapitre 1. Avec l'utilisation de données généalogiques composées d'un total de 94 076 ancêtres, l'article *Deciphering the genetic structure of the Quebec founder population using genealogies* dévoile et suit dans le temps l'évolution la structure génétique fine de la population québécoise. De plus, cela avait pour objectif de mieux comprendre l'impact de certains processus démographiques ayant mené à l'apparition d'effets fondateurs régionaux au Québec, avec l'aide de données génétiques et généalogiques.

La structure génétique du Québec a déjà été rapportée par d'autres études[26,49,60,67,68,76]. Également, l'impact de l'effet fondateur au Saguenay-Lac-Saint-Jean a bien été étudié en raison du grand nombre de maladies plus fréquentes dans cette région[26,58]. En revanche, la nouveauté de ce projet est au-delà de l'étude de la structure du Québec. À cet effet, notre objectif est atteint où des données généalogiques ont été utilisées afin de suivre l'évolution de cette structure dans le temps, depuis la période contemporaine jusque dans les années 1600. Cela a permis de constater que la structure de la population québécoise qui distingue le Saguenay-Lac-Saint-Jean et la Gaspésie est apparue dès 1751 et persiste jusqu'à aujourd'hui. Ainsi, les conséquences de cette structure ont encore une implication importante actuellement, surtout au niveau des maladies génétiques plus fréquentes. Néanmoins, il serait intéressant de voir comment cette structure va évoluer dans le futur et si elle va être conservée ou diminuée dans le temps. Un autre facteur de nouveauté réside dans le calcul du coefficient d'apparentement et de consanguinité à partir d'ancêtres par tranche de temps de 25 ans. Les analyses généalogiques sur la population québécoise effectuées par d'autres groupes de recherche mesurent toujours ces coefficients à partir de la première génération, soit les individus contemporains, et effectuent le calcul pour chaque génération dans le passé[49,67]. Ainsi, l'approche inverse en recalculant ces coefficients toujours à partir de sous-groupes

d'ancêtres, et non les individus de la première génération, a permis de voir l'évolution de cette structure et à quel moment elle est apparue.

Une fois que cette structure finie a été établie, un autre objectif a été atteint où il a été possible de comprendre comment cette structure a apporté la différenciation de ces deux populations. Un lien a été observé entre les intervalles de la longueur des segments IBD et le nombre cumulatif de MRCA et avec la contribution génétique de ceux-ci. Ces données, génétiques et généalogiques, révèlent l'impact de certains phénomènes démographiques au Québec. Ainsi, il a été possible de comprendre quel type de phénomènes démographiques est derrière chacun des sous-effets fondateurs du Saguenay-Lac-Saint-Jean et de la Gaspésie. Au niveau du Saguenay-Lac-Saint-Jean, l'effet fondateur est surtout causé par une forte expansion qui a mené à la structure de la population actuelle. Il est possible de remarquer un plateau du niveau de consanguinité au Saguenay-Lac-Saint-Jean, associé à de plus courts segments IBD, qui témoigne de ce fort accroissement naturel. Cette structure qui distingue le Saguenay-Lac-Saint-Jean a émergée à partir de Charlevoix dès 1751, presque 100 ans avant même la colonisation de la région. Pour les Acadiens de Gaspésie, les résultats montrent les conséquences de l'endogamie et le peu de mélange avec les autres populations, comme démontré dans d'autres études[57]. Pour les Loyalistes de Gaspésie, un groupe peu étudié, il a été démontré qu'ils ont des segments génétiques IBD plus longs par rapport aux autres groupes, sauf les Acadiens de Gaspésie, liés à un nombre plus élevé de MRCA récent. Les longs segments IBD présents chez les groupes de Gaspésie sont liés à une augmentation continue du taux de consanguinité.

De plus, une croyance populaire veut croire que les individus du Saguenay-Lac-Saint-Jean soient plus consanguins qu'ailleurs au Québec. Cependant, ce projet confirme que cette croyance est fausse, comme démontré dans d'autres études[49,111]. En effet, le Saguenay-Lac-Saint-Jean présente une forte consanguinité éloignée qui est liée au fort taux d'accroissement

naturel et au petit nombre de fondateurs prolifiques, où le grand nombre d'enfants a fait augmenter la fréquence de mêmes ancêtres dans les généalogies. Pour ce qui est de la consanguinité proche, celle qui importe au niveau du choix de partenaire, le Saguenay-Lac-Saint-Jean est un groupe présentant le niveau le plus bas de consanguinités dans tout le Québec[49]. Ainsi, cette croyance populaire est totalement fausse.

3.1.3 Retour sur les deux projets

L'objectif global visant à étudier la structure de populations à effet fondateur et son importance dans l'apparition de phénomènes démographiques est atteint. L'utilisation de populations à effet fondateur a permis d'étudier la présence et l'importance de la structure génétique d'une population. Des phénomènes démographiques, comme l'apparentement, la consanguinité, l'accroissement naturel ou l'endogamie, menant à la différenciation de populations sont clarifiés. L'apparition et l'évolution de la structure de la population au Québec sont également expliquées. De plus, une meilleure compréhension de nos populations mène à l'ajout de meilleures stratégies pour des études sur l'identification de nouveaux variants ou des études liées à des maladies. Ainsi, mieux comprendre les structures de population présentes aujourd'hui et comment elles sont apparues offre un apport incroyable pour les études génétiques et un éclaircissement sur comment elles pourraient évoluer dans le futur.

Au niveau du Québec, ces approches impactent la recherche actuelle en démontrant que l'effet fondateur québécois est très hétérogène et que la structure est apparue avant même la formation d'une des régions impliquées. Le Saguenay-Lac-Saint-Jean a très bien été étudié en raison des maladies qui y sont fréquentes. De plus, un test de porteur disponible pour sa population aide les parents au moment de la préconception[37,79]. Toutefois, les populations de la Gaspésie présentent un profil similaire d'effet fondateur, même que les résultats du coefficient de consanguinité et d'apparentement révèle des niveaux plus élevés qu'au Saguenay-Lac-Saint-

Jean dans les dernières générations. Cependant, aucun dépistage génétique ou même des tests de porteur n'y sont offerts. Au niveau des Acadiens du Nouveau-Brunswick, il y a une nouvelle attention portée à des variants plus fréquents[112]. De ce fait, il serait important d'investiguer la présence de variants rares plus communs chez les populations de Gaspésie, qui semblent être une région indiquée pour l'implémentation d'un test de porteur.

Ces résultats sont également valables pour des populations n'ayant pas vécu d'effet fondateur. L'ascendance continentale n'est pas simplement une composante homogène de la diversité génétique, mais est beaucoup plus complexe. Il y a des populations métissées, comme des individus d'ascendance africaine et latino-américaine qui ont des structures de population importante ce qui impacte les analyses génétiques[113]. De plus, il a été démontré que la population européenne est dotée d'une structure et est métissé, ce qui n'est pas toujours pris en compte dans les études d'association génétique, ce qui mène à des biais et des faux-positifs[114]. Également, les scores polygéniques sont sensibles à la structure de population[115]. Ainsi, la tendance génétique actuelle veut construire les plus grandes cohortes possibles avec des individus d'origine diverse. Les chercheurs devront être très attentifs à ces structures fines de population pour ne pas causer de biais. De ce fait, il est important de prendre en compte ces structures, au sein de toutes les populations, afin de bien effectuer des associations génétiques, au lieu de seulement les corriger.

3.2 Limitations

Ces deux projets comprennent tous les deux des limitations. Tout d'abord, la cohorte de référence du Québec utilisée pour les deux projets à une très bonne représentation des groupes ethnoculturels sélectionnés. Toutefois, il y a certaines régions du Québec où il y avait très peu d'individus et d'autres régions qui n'ont tout simplement pas été étudiées. À cet effet, si je pouvais créer ma propre cohorte populationnelle d'individus québécois, j'essayerais d'être plus

inclusif de l'ensemble du Québec. En effet, il y a peut-être d'autres régions qui se différencient et qui seraient importantes de connaitre ; mais qui sont encore inconnues en raison d'un manque de données génétiques d'individus vivant en région, et non seulement dans les grands centres.

Au niveau du Chapitre 1, nous avons réussi à avoir accès à quatre cohortes de populations à effet fondateur. Trois de ces cohortes sont d'origine européenne et la dernière d'origine africaine. Malgré un effort afin de sélectionner des populations provenant de plusieurs origines, il n'était pas toujours facile d'obtenir l'accès à ces données. En effet, les populations devaient avoir vécu un effet fondateur, ensuite les données devaient être en accès libre et finalement le consentement des données devait être ouvert à des études en génétique des populations. C'est souvent ce dernier critère qui éliminait des cohortes que j'ai pu trouver, où le consentement autorisait seulement des recherches sur une maladie précise. Ainsi, l'ajout d'autres cohortes d'origines diverses aurait été souhaitable. De plus, l'intégration de plusieurs cohortes impliquait un nettoyage de données important. Seulement les SNP en commun ont été considérés pour les analyses subséquentes afin de ne pas causer de biais. Ainsi, seulement un nombre de SNP limité répondait à ces critères et un total de 199 238 SNP ont été utilisés pour les analyses. Ce nombre est suffisant, mais si cela avait été possible, avoir un nombre supérieur de SNP aurait été l'idéal afin de mieux discriminer les structures de population. L'identification des segments ROH dépend du nombre de SNP et d'autres analyses ont finalement été éliminées, car ils nécessitaient un plus grand nombre de SNP et d'individus, comme l'analyse de la taille effective de population[116,117]. Finalement, puisque les données auxquelles nous avons eu accès sont des données de génotypage, il a été nécessaire d'imputer les données afin de vérifier la présence des variants rares associés à des populations spécifiques. L'utilisation de données imputées a pu entrainer une perte de variantes rares ou une sous-estimation de la fréquence réelle de ces variantes dans les populations à effet fondateur. En effet, l'imputation de données génétiques dépend des SNP présents qui proviennent de populations européennes

n'ayant pas vécu d'effet fondateur, et un nombre faible peu mal représenter les haplotypes réels.

Pour ce qui est du Chapitre 2, l'utilisation de données rares telle que les généalogies est un outil incroyable. Toutefois, ces données ont une limite. Les données généalogiques de BALSAC recensent le peuplement des territoires du Québec depuis sa colonisation en 1608 jusqu'à aujourd'hui. Ainsi, il est impossible de remonter plus loin dans le passé et d'avoir plus d'information. De plus, si une lignée s'arrête tôt en raison d'un manque de données, il est impossible d'avoir les informations manquantes. Ces données sont donc un outil puissant, mais il faut les analyser avec précaution lorsqu'on arrive vers la fin des lignées généalogiques en raison du biais de complétude. De plus, une méthode *bootstrap* a été appliquée dans l'article présenté au chapitre 2 afin de limiter le biais de complétude entre les différentes régions. Finalement, il est impossible d'avoir l'information sur les fondateurs, c'est-à-dire les fins de lignées généalogiques, donc le lien entre ces individus est inconnu.

3.3 Perspectives

Ces deux projets utilisent la structure fine des populations afin de mieux comprendre leur évolution et l'implication de ces structures comme méthode pour faciliter l'identification de variants rares. Nous croyons que l'utilisation de cette structure est la prochaine avenue que devraient emprunter les études d'association de variants rares, et même de variants communs. La mise en valeur de cohortes existantes et la recherche de structure fine à l'intérieur de celles-ci sont un outil qui aide à concentrer ces variants rares et à en identifier des nouveaux. En utilisant seulement des individus d'un même regroupement, il y a plus de chance que les individus restants possèdent les mêmes variants rares et que donc ceux-ci soient concentrés au sein du regroupement. Une fois identifiés, ces variants peuvent nous aider à mieux comprendre le mécanisme des maladies associées. De plus, l'utilisation d'une structure fine permettrait

aussi de réévaluer des associations déjà établies afin de vérifier si celles-ci sont véridiques. Également, dans des populations où des maladies sont plus fréquentes, tel que dans certaines populations à effet fondateur, cette méthode pourrait servir à identifier les variants impliqués dans des maladies propres à ces régions. Des procédures pourraient être implantées, telles que les tests de porteur, afin d'aider directement ces populations.

Ensuite, ce projet vise également à promouvoir l'utilisation de données en accès ouvert. À l'époque où de nombreux ensembles de données génétiques existent, il est important d'encourager la réutilisation de ces données dans d'autres projets pour leur donner une nouvelle vie. Notre approche va dans le même sens où l'identification de ces structures fines au sein de grandes cohortes pourrait augmenter leur valeur au lieu de toujours vouloir créer les plus grandes cohortes possibles ce qui réduit les coûts associés.

Par ailleurs, l'investigation de ces sous-structures au sein du Québec avec les données généalogiques est une avenue importante afin de comprendre la transmission de traits génétiques. Les données généalogiques ont permis d'obtenir une nouvelle résolution où deux populations présentant un sous-effet fondateur puissant se différenciaient de l'effet fondateur du Québec. Les analyses généalogiques ont dévoilé que les processus démographiques, tels qu'un fort accroissement naturel et l'endogamie, ayant façonné ces populations sont complètement différents et peuvent impacter la transmission de variants liés à des maladies. Ainsi, en comprenant mieux d'où viennent nos populations, il est possible de mieux étudier leur futur.

Ce mémoire de maitrise démontre que la génétique des populations est une discipline essentielle. Passant de meilleures approches afin d'étudier des variants rares à mieux comprendre les phénomènes démographiques vécus par une population, l'étude et la

connaissance de la génétique des populations sont essentielles afin d'effectuer des recherches sur des maladies. De plus, l'étude des populations à effet fondateur aide à comprendre leur évolution. De nos jours, la facilité de se déplacer favorise une connexion plus aisée entre différents lieux, ce qui engendre une augmentation du métissage. Par conséquent, l'impact de la structure génétique de populations à effet fondateur sera de plus en plus moindre. Cela apportera certains résultats, tels qu'une diminution de la fréquence des maladies liées à ce type de population. Toutefois, les projets de colonisation spatiale mèneront sans aucun doute à l'apparition de forts effets fondateurs. Ainsi, les études actuelles sur les populations à effet fondateur permettront d'avoir de meilleures connaissances afin de bien préparer ce phénomène.

Conclusion

Ce projet avait comme objectif de mettre en évidence la structure génétique fine de populations à effet fondateur et d'étudier l'importance des processus démographiques dans l'apparition de cette structure. Pour ce faire, une analyse de la structure génétique fine de quatre populations à effet fondateur a été réalisée. Cette structure fine a ensuite été utilisée afin de calculer l'augmentation en fréquence de variants rares dans les regroupements créés, comparativement à la population globale. Ainsi, il a été déterminé que l'utilisation de plus petites cohortes, avec une structure génétique semblable, est un outil puissant pour l'identification de variants rares. D'un autre côté, l'investigation en profondeur de la structure fine présente au Québec a révélé le moment de son apparition en 1750 grâce à des analyses généalogiques, distinguant le Saguenay-Lac-Saint-Jean et la Gaspésie. De plus, ces mêmes données ont révélé comment les processus démographiques du passé, comme un fort accroissement naturel et l'endogamie, ont pu mener à la structure génétique contemporaine. Ce projet permet de saisir l'importance d'étudier la génétique de la population avant d'effectuer des analyses liées à des maladies. En effet, ces approches démontrent l'impact d'une structure d'une population sur l'étude de variants rares. Ces résultats veulent être une nouvelle façon d'effectuer des études génétiques liées à des maladies, en tenant compte de la génétique des populations.

Bibliographie

1. Griffiths, A. J. F., Wessler, S. R., Carroll, S. B. & Doebley, J. *Introduction to genetic analysis*. (W.H. Freeman ;, New York, 2012).

2. Templeton, A. R. *Human population genetics and genomics*. (Academic Press, 2019).

3. National Human Genome Research Institute. Genetics vs. Genomics Fact Sheet. *Genome.gov* https://www.genome.gov/about-genomics/fact-sheets/Genetics-vs-Genomics (2018).

4. Ku, C. S., Loy, E. Y., Salim, A., Pawitan, Y. & Chia, K. S. The discovery of human genetic variations and their use as disease markers: past, present and future. *J Hum Genet* **55**, 403–415 (2010).

5. Shastry, B. S. SNPs: impact on gene function and phenotype. *Methods Mol Biol* **578**, 3–22 (2009).

6. Mullaney, J. M., Mills, R. E., Pittard, W. S. & Devine, S. E. Small insertions and deletions (INDELs) in human genomes. *Hum Mol Genet* **19**, R131–R136 (2010).

7. Collins, R. L. *et al.* A structural variation reference for medical and population genetics. *Nature* **581**, 444–451 (2020).

8. Lin, M. *et al.* Effects of short indels on protein structure and function in human genomes. *Sci Rep* **7**, 9313 (2017).

9. Rodriguez-Murillo, L. & Salem, R. M. Insertion/Deletion Polymorphism. in *Encyclopedia of Behavioral Medicine* (eds. Gellman, M. D. & Turner, J. R.) 1076–1076 (Springer, New York, NY, 2013). doi:10.1007/978-1-4419-1005-9_706.

10. Fan, H. & Chu, J.-Y. A Brief Review of Short Tandem Repeat Mutation. *Genomics Proteomics Bioinformatics* **5**, 7–14 (2007).

11. Marshall, J. N. *et al.* Variable number tandem repeats – Their emerging role in sickness and health. *Exp Biol Med (Maywood)* **246**, 1368–1376 (2021).

12. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

13. Goswami, C., Chattopadhyay, A. & Chuang, E. Y. Rare variants: data types and analysis strategies. *Ann Transl Med* **9**, 961 (2021).

14. Gusella, J. F. *et al.* A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* **306**, 234–238 (1983).

15. Chung, B. H. Y., Chau, J. F. T. & Wong, G. K.-S. Rare versus common diseases: a false dichotomy in precision medicine. *npj Genom. Med.* **6**, 1–5 (2021).

16. Craig, J. Complex Diseases: Research and Applications. *Nature Education* (2008).

17. Hellwege, J. *et al.* Population Stratification in Genetic Association Studies. *Curr Protoc Hum Genet* **95**, 1.22.1-1.22.23 (2017).

18. Okasha, S. Population Genetics. in *The Stanford Encyclopedia of Philosophy* (eds. Zalta, E. N. & Nodelman, U.) (Metaphysics Research Lab, Stanford University, 2023).

19. Mayo, O. A Century of Hardy–Weinberg Equilibrium. *Twin Research and Human Genetics* **11**, 249–256 (2008).

20. Charlesworth, D. & Willis, J. H. The genetics of inbreeding depression. *Nat Rev Genet* **10**, 783–796 (2009).

21. Maia, R. T. & Araújo, M. C. de. *Population genetics*. (IntechOpen, London, 2022).

22. Kliman, R., Sheehy, B. & Schultz, J. Genetic Drift and Effective Population Size | Learn Science at Scitable. *Nature Education* (2008).

23. Charlesworth, B. Effective population size and patterns of molecular evolution and variation. *Nat Rev Genet* **10**, 195–205 (2009).

24. Wein, T. & Dagan, T. The Effect of Population Bottleneck Size and Selective Regime on Genetic Diversity and Evolvability in Bacteria. *Genome Biol Evol* **11**, 3283–3290 (2019).

25. Kivisild, T. Founder Effect. in *Brenner's Encyclopedia of Genetics (Second Edition)* (eds. Maloy, S. & Hughes, K.) 100–101 (Academic Press, San Diego, 2013). doi:10.1016/B978-0-12-374984-0.00552-0.

26. Bouchard, G. & De Braekeleer, M. *Histoire d'un Genome: Population et Genetique Dans l'est Du Quebec*. (Sillery, Québec: Presses de l'Université du Québec, 1991).

27. Jain, A., Sharma, D., Bajaj, A., Gupta, V. & Scaria, V. Founder variants and population genomes-Toward precision medicine. *Adv Genet* **107**, 121–152 (2021).

28. Tournebize, R., Chu, G. & Moorjani, P. Reconstructing the history of founder events using genome-wide patterns of allele sharing across individuals. *PLOS Genetics* **18**, e1010243 (2022).

29. Cavalli-Sforza, L. L. (Luigi L., 1922-2018. & Bodmer, W. F. (Walter F., 1936-. *The genetics of human populations*. (W.h. Freeman, San Francisco, 1971).

30. Santos, J. *et al.* From nature to the laboratory: the impact of founder effects on adaptation. *Journal of Evolutionary Biology* **25**, 2607–2622 (2012).

31. Arcos-Burgos, M. & Muenke, M. Genetics of population isolates. *Clinical Genetics* **61**, 233–247 (2002).

32. Rouleau, G. Portrait démogénétique de la RMR de Saguenay. (Université Laval, Québec, 2017).

33. Henry, J.-P. Génétique et origine d'Homo sapiens. *Med Sci (Paris)* **35**, 39–45 (2019).

34. Chong, J. X., Ouwenga, R., Anderson, R. L., Waggoner, D. J. & Ober, C. A Population-Based Study of Autosomal-Recessive Disease-Causing Mutations in a Founder Population. *Am J Hum Genet* **91**, 608–620 (2012).

35. Xue, Y. *et al.* Enrichment of low-frequency functional variants revealed by whole-genome sequencing of multiple isolated European populations. *Nat Commun* **8**, 15927 (2017).

36. Scott, S. A. *et al.* Experience with carrier screening and prenatal diagnosis for 16 Ashkenazi Jewish genetic diseases. *Human Mutation* **31**, 1240–1250 (2010).

37. Bchetnia, M. *et al.* Genetic burden linked to founder effects in Saguenay-Lac-Saint-Jean illustrates the importance of genetic screening test availability. *J Med Genet* **58**, 653–665 (2021).

38. Bonner, J. D. *et al.* Pedigree structure and kinship measurements of a mid-Michigan community: A new North American population isolate identified. *Human biology* **86**, 59–68 (2014).

39. Sergerie, François. Le peuplement fondateur de la région de Lotbinière et ses conséquences démogénétiques. (Université de Montréal (Faculté des arts et des sciences), Montréal, 2010).

40. Gouvernement du Canada, S. C. Croissance démographique: l'accroissement migratoire l'emporte sur l'accroissement naturel. https://www150.statcan.gc.ca/n1/pub/11-630-x/11-630-x2014001-fra.htm#a1 (2014).

41. Bouchard, G., Charbonneau, H., Desjardins, B., Heyer, É. & Tremblay, M. Mobilité géographique et stratification du pool génique canadien-français sous le Régime français. in *Les chemins de la migration en Belgique et au Québec, du XVIIe au XXe siècle.* 51–59 (1995).

42. Keinan, A. & Clark, A. G. Recent Explosive Human Population Growth Has Resulted in an Excess of Rare Genetic Variants. *Science* **336**, 740–743 (2012).

43. Casals, F. *et al.* Whole-exome sequencing reveals a rapid change in the frequency of rare functional variants in a founding population of humans. *PLoS Genet* **9**, e1003815 (2013).

44. François, O. *et al.* Principal component analysis under population genetic models of range expansion and admixture. *Mol Biol Evol* **27**, 1257–1268 (2010).

45. Diaz-Papkovich, A., Anderson-Trocmé, L. & Gravel, S. A review of UMAP in population genetics. *J Hum Genet* **66**, 85–91 (2021).

46. Severson, A. L., Carmi, S. & Rosenberg, N. A. The Effect of Consanguinity on Between-Individual Identity-by-Descent Sharing. *Genetics* **212**, 305–316 (2019).

47. Browning, S. R. & Thompson, E. A. Detecting Rare Variant Associations by Identity-by-Descent Mapping in Case-Control Studies. *Genetics* **190**, 1521–1531 (2012).

48. Browning, S. R. & Browning, B. L. Identity by descent between distant relatives: detection and applications. *Annu Rev Genet* **46**, 617–633 (2012).

49. Vézina, H., Tremblay, M. & Houde, L. Mesures de l'apparentement biologique au Saguenay-Lac-St-Jean (Québec, Canada) à partir de reconstitutions généalogiques. *Annales de démographie historique* **108**, 67–83 (2004).

50. Anderson-Trocmé, L. *et al.* On the genes, genealogies, and geographies of Quebec.

Science **380**, 849–855 (2023).

51. Pettay, J. E., Kruuk, L. E. B., Jokela, J. & Lummaa, V. Heritability and genetic constraints of life-history trait evolution in preindustrial humans. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 2838–2843 (2005).

52. Helgason, A., Hrafnkelsson, B., Gulcher, J. R., Ward, R. & Stefánsson, K. A Populationwide Coalescent Analysis of Icelandic Matrilineal and Patrilineal Genealogies: Evidence for a Faster Evolutionary Rate of mtDNA Lineages than Y Chromosomes. *The American Journal of Human Genetics* **72**, 1370–1388 (2003).

53. Pluzhnikov, A., Nolan, D. K., Tan, Z., McPeek, M. S. & Ober, C. Correlation of intergenerational family sizes suggests a genetic component of reproductive fitness. *American Journal of Human Genetics* **81**, 165–169 (2007).

54. O'Brien, E., Zenger, R. & Jorde, L. B. Genetic structure of the Utah Mormons: A comparison of kinship estimates from DNA blood groups, genealogies, and ancestral arrays. *American Journal of Human Biology* **8**, 609–614 (1996).

55. BALSAC. BALSAC. *BALSAC* https://balsac.uqac.ca/.

56. Roy-Gagnon, M. H. *et al.* Genomic and genealogical investigation of the French Canadian founder population structure. *Human Genetics* **129**, 521–531 (2011).

57. Moreau, C. *et al.* Genetic heterogeneity in regional populations of Quebec - Parental lineages in the Gaspe Peninsula. *American Journal of Physical Anthropology* **139**, 512–522 (2009).

58. Scriver, C. R. Human Genetics : Lessons from Quebec Populations. *Annual Review of Genomics and Human Genetics* **2**, 69–101 (2001).

59. De Braekeleer, M. & Dao, T. N. Hereditary disorders in the French Canadian population of Quebec. I. In search of founders. *Human biology* **66**, 205–23 (1994).

60. Gauvin, H. *et al.* Genome-wide patterns of identity-by-descent sharing in the French Canadian founder population. *Eur J Hum Genet* **22**, 814–821 (2014).

61. Bouchard, G. & De Braekeleer, M. *Pourquoi Des Maladies Héréditaires?: Population et Génétique Au Saguenay-Lac-Saint-Jean*. (Les éditions du Septentrion, 1992).

62. Moreau, C. *et al.* Deep human genealogies reveal a selective advantage to be on an expanding wave front. *Science* **334**, 1148–50 (2011).

63. Gauvin, H. *et al.* GENLIB: an R package for the analysis of genealogical data. *BMC Bioinformatics* **16**, 160 (2015).

64. Cazes, P. & Cazes, M.-H. Comment mesurer la profondeur généalogique d'une ascendance ? *Population* **51**, 117–140 (1996).

65. Karigl, G. A recursive algorithm for the calculation of identity coefficients. *Annals of Human Genetics* **45**, 299–305 (1981).

66. Tzeng, J., Lu, H. H.-S. & Li, W.-H. Multidimensional scaling for large genomic data sets. *BMC Bioinformatics* **9**, 179 (2008).

67. Roy-Gagnon, M.-H. *et al.* Genomic and genealogical investigation of the French Canadian founder population structure. *Hum Genet* **129**, 521–531 (2011).

68. Bherer, C. *et al.* Admixed ancestry and stratification of Quebec regional populations. *American Journal of Physical Anthropology* **144**, 432–441 (2011).

69. Asselin, G. Facteurs contribuant à l'homogénéisation du pool génique de la population humaine de l'Île-aux-Coudres à partir de l'étude des contributions génétiques de ses fondateurs. (Université du Québec à Montréal, 2003).

70. Roberts, D. F. Genetic Effects of Population Size Reduction. *Nature* **220**, 1084–1088 (1968).

71. Moreau, C. *et al.* Native American admixture in the Quebec founder population. *PLoS One* **8**, e65507 (2013).

72. Charbonneau, H., Desjardins, B., Légaré, J. & Denis, H. The population of the St-Lawrence Valley, 1608-1760. in *A Population History of North America* 99–142 (2000).

73. Laberge, A.-M. *et al.* Population history and its impact on medical genetics in Quebec. *Clin Genet* **68**, 287–301 (2005).

74. Fyson, D. Between the Ancien Régime and Liberal Modernity: Law, Justice and State Formation in colonial Quebec, 1760–1867. *History Compass* **12**, 412–432 (2014).

75. Vézina, H., Tremblay, M., Desjardins, B. & Houde, L. Origines et contributions génétiques des fondatrices et des fondateurs de la population québécoise. *cqd* **34**, 235–258 (2005).

76. Gagnon, A. & Heyer, E. Fragmentation of the Québec population genetic pool (Canada): Evidence from the genetic contribution of founders per region in the 17th and 18th centuries. *American Journal of Physical Anthropology* **114**, 30–41 (2001).

77. Jette, R., Gauvreau, D. & Guérin, M. Aux origines d'une région: le peuplement fondateur de Charlevoix avant 1850. in *Histoire d'un génome. Population et génétique dans l'est du Québec. Québec: Presses de l'Université du Québec* 75–106 (1991).

78. Pouyez, C. & Lavoie, Y. *Les Saguenayens. Introduction à l'histoire Des Populations Du Saguenay*. (1983).

79. Cruz Marino, T. *et al.* Portrait of autosomal recessive diseases in the French-Canadian founder population of Saguenay-Lac-Saint-Jean. *Am J Med Genet A* **191**, 1145–1163 (2023).

80. Gouvernement du Québec. Offre de tests de porteur pour quatre maladies héréditaires récessives chez les personnes originaires des régions du Saguenay–Lac-Saint-Jean, de Charlevoix et de la Haute-Côte-Nord. *Gouvernement du Québec* https://www.quebec.ca/sante/conseils-et-prevention/depistage-et-offre-de-tests-de-porteur/tests-de-porteur-maladies-hereditaires-recessive/description-tests-porteur (2024).

81. Ostrer, H. A genetic profile of contemporary Jewish populations. *Nat Rev Genet* **2**, 891–898 (2001).

82. Ostrer, H. & Skorecki, K. The population genetics of the Jewish people. *Hum Genet* **132**, 119–127 (2013).

83. Bray, S. M. *et al.* Signatures of founder effects, admixture, and selection in the Ashkenazi Jewish population. *Proceedings of the National Academy of Sciences* **107**, 16222–16227 (2010).

84. Behar, D. M. *et al.* Counting the Founders: The Matrilineal Genetic Ancestry of the Jewish Diaspora. *PLoS One* **3**, e2062 (2008).

85. Waldman, S. *et al.* Genome-wide data from medieval German Jews show that the Ashkenazi founder event pre-dated the 14th century. *Cell* **185**, 4703-4716.e16 (2022).

86. Lencz, T. *et al.* Genome-wide association study implicates NDST3 in schizophrenia and bipolar disorder. *Nat Commun* **4**, 2739 (2013).

87. Goes, F. S. *et al.* Genome-wide association study of schizophrenia in Ashkenazi Jews. *Am J Med Genet B Neuropsychiatr Genet* **168**, 649–659 (2015).

88. Carmi, S. *et al.* Sequencing an Ashkenazi reference panel supports population-targeted personal genomics and illuminates Jewish and European origins. *Nat Commun* **5**, 4835 (2014).

89. Risch, N., Tang, H., Katzenstein, H. & Ekstein, J. Geographic Distribution of Disease Mutations in the Ashkenazi Jewish Population Supports Genetic Drift over Selection. *Am J Hum Genet* **72**, 812–822 (2003).

90. Palamara, P. F., Lencz, T., Darvasi, A. & Pe'er, I. Length distributions of identity by descent reveal fine-scale demographic history. *Am J Hum Genet* **91**, 809–822 (2012).

91. Guha, S. *et al.* Implications for health and disease in the genetic signature of the Ashkenazi Jewish population. *Genome Biology* **13**, R2 (2012).

92. Boycott, K. M. *et al.* Clinical genetics and the Hutterite population: a review of Mendelian disorders. *Am J Med Genet A* **146A**, 1088–1098 (2008).

93. Hostetler, J. A., Opitz, J. M. & Reynolds, J. F. History and relevance of the Hutterite population for genetic studies. *American Journal of Medical Genetics* **22**, 453–462 (1985).

94. Ober, C. *et al.* HLA and mate choice in humans. *Am J Hum Genet* **61**, 497–504 (1997).

95. O'Brien, E., Kerber, R. A., Jorde, L. B. & Rogers, A. R. Founder Effect: Assessment of Variation in Genetic Contributions among Founders. *Human Biology* **66**, 185–204 (1994).

96. Nimgaonkar, V. L. *et al.* Low prevalence of psychoses among the Hutterites, an isolated religious community. *Am J Psychiatry* **157**, 1065–1070 (2000).

97. Scelza, B. A. Female mobility and postmarital kin access in a patrilocal society. *Hum Nat* **22**, 377–393 (2011).

98. Oliveira, S. *et al.* Matriclans shape populations: Insights from the Angolan Namib Desert into the maternal genetic history of southern Africa. *Am J Phys Anthropol* **165**, 518–535 (2018).

99. Scelza, B. A. *et al.* High rate of extrapair paternity in a human population demonstrates diversity in human reproductive strategies. *Science Advances* **6**, eaay6195 (2020).

100. Bollig, Michael. *Risk management in a hazardous environment : a comparative study of two pastoral societies*. (Springer, New York, 2006). doi:10.1007/978-0-387-27582-6.

101. Scelza, B. & Prall, S. Partner preferences in the context of concurrency: What Himba want in formal and informal partners. *Evolution and Human Behavior* **39**, 212–219 (2017).

102. Swinford, N. A. *et al.* Increased homozygosity due to endogamy results in fitness consequences in a human population. 2022.07.25.501261 Preprint at https://doi.org/10.1101/2022.07.25.501261 (2022).

103. Bollig, M. Risk and Risk Minimisation among Himba Pastoralists in Northwestern Namibia. *Nomadic Peoples* **1**, 66–89 (1997).

104. Gladstein, A. L. & Hammer, M. F. Substructured Population Growth in the Ashkenazi Jews Inferred with Approximate Bayesian Computation. *Molecular Biology and Evolution* **36**, 1162–1171 (2019).

105. Feder, J., Ovadia, O., Glaser, B. & Mishmar, D. Ashkenazi Jewish mtDNA haplogroup distribution varies among distinct subpopulations: lessons of population substructure in a closed group. *Eur J Hum Genet* **15**, 498–500 (2007).

106. Uffelmann, E. *et al.* Genome-wide association studies. *Nat Rev Methods Primers* **1**, 1–21 (2021).

107. Bomba, L., Walter, K. & Soranzo, N. The impact of rare and low-frequency genetic variants in common disease. *Genome Biology* **18**, 77 (2017).

108. Auer, P. L. & Lettre, G. Rare variant association studies: considerations, challenges and opportunities. *Genome Medicine* **7**, 16 (2015).

109. Shriner, D. *et al.* Universal genome-wide association studies: Powerful joint ancestry and association testing. *HGG Adv* **4**, 100235 (2023).

110. Awadalla, P. *et al.* Cohort profile of the CARTaGENE study: Quebec's population-based biobank for public health and personalized genomics. *Int J Epidemiol* **42**, 1285–1299 (2013).

111. Mourali-Chebil, S. & Heyer, E. Evolution of inbreeding coefficients and effective size in the population of Saguenay Lac-St.-Jean (Quebec). *Hum Biol* **78**, 495–508 (2006).

112. Robichaud, P. P. *et al.* Pathogenic variants carrier screening in New Brunswick: Acadians reveal high carrier frequency for multiple genetic disorders. *BMC Medical Genomics* **15**, 98 (2022).

113. Brugger, S. W. & Davis, M. F. Influence of Admixture on Phenotypes. *Curr Protoc* **3**, e953 (2023).

114. Gouveia, M. H. *et al.* Unappreciated subcontinental admixture in Europeans and European Americans and implications for genetic epidemiology studies. *Nat Commun* **14**, 6802 (2023).

115. Sohail, M. *et al.* Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *Elife* **8**, e39702 (2019).

116. Ceballos, F. C., Joshi, P. K., Clark, D. W., Ramsay, M. & Wilson, J. F. Runs of homozygosity: windows into population history and trait architecture. *Nat Rev Genet* **19**, 220–234 (2018).

117. Browning, S. R. & Browning, B. L. Accurate Non-parametric Estimation of Recent Effective Population Size from Segments of Identity by Descent. *Am J Hum Genet* **97**, 404–418 (2015).

Certification éthique

Ce mémoire a fait l'objet d'une certification éthique au CER de l'UQAC. Le numéro du certificat est : 2021-560 - Études des populations fondatrices.

Annexe 1

L'Annexe 1 comporte les informations supplémentaires de l'article *Fine-scale genetic structure and rare variant frequencies* présenté au Chapitre 1. Le tableau supplémentaire 2 se retrouve au lien suivant : https://www.biorxiv.org/content/10.1101/2024.02.02.578687v2, en raison de sa taille.

Supplementary information

Fine-scale genetic structure and rare variant frequencies
Laurence Gagnon, Claudia Moreau, Catherine Laprise, Simon L. Girard.

Supplementary subjects and methods

Genotyping data and cleaning

Each dataset underwent cleaning using PLINK software v1.9, ensuring individuals with at least 95% genotypes among all SNPs were retained (1). At the SNP level, we retained SNPs with at least 95% genotypes among all individuals, located on the autosomes and in Hardy–Weinberg equilibrium $p > 0.001$ (calculated on each whole cohort). A Principal Component Analysis (PCA) was done on each individual dataset using SNPs with a minor allele frequency (MAF) of at least 5%, and after pruning to remove SNPs in linkage disequilibrium.

Subsequently, all datasets were merged (lifting over to hg19 for the Ashkenazi Jews) to retain only common bi-allelic SNPs that are in intersection between all datasets. After the merge, individuals with less than 95% genotypes among all SNPs and SNPs with less than 95% genotypes across all individuals were once again filtered out. The final dataset comprises 199,238 SNPs and 4,259 individuals. Related individuals (PLINK pihat >= 0.25) were filtered out, resulting in a final sample size of 3,683 subjects. This unusually high threshold was applied to retain two populations with high relatedness (Table S4). Hutterites are indeed recognized for practicing endogamy and communal living, while Himba individuals have a pastoralist lifestyle and practice polygyny (2,3). The Figure S4 demonstrates the low impact of different genetic relatedness thresholds on the pairwise sum of identity-by-descent (IBD) segments length and number. The merged dataset was imputed on TOPMed imputation server, using the reference panel topmed-r2 after lifting over to hg38 (4). Postimputation quality control filters were applied to remove SNPs within imputed data with an imputation quality score of <0.3 and only biallelic SNPs were kept for further analyses.

Finally, a PCA was performed on SNPs with a MAF of at least 5% and after pruning (threshold 50 5 0.2) to remove SNPs in linkage disequilibrium (73,624 SNPs left).

Analyses of IBD segments

The assessment of pairwise identity-by-descent (IBD) segments was performed on the common variants that are in intersection between all datasets using refinedIBD software v17Jan20 on phased genotypes, which was done using Beagle software version 18May20.d20 (5). The identified segments were then merged with merge-ibd-segments.17Jan20.102.jar. This software was selected for its robustness and precision in detecting IBD segments (5). Only segments of 2 cM or more and with a LOD score greater than 3 were retained for further analysis on the level of IBD sharing across the genome.

Supplementary figures

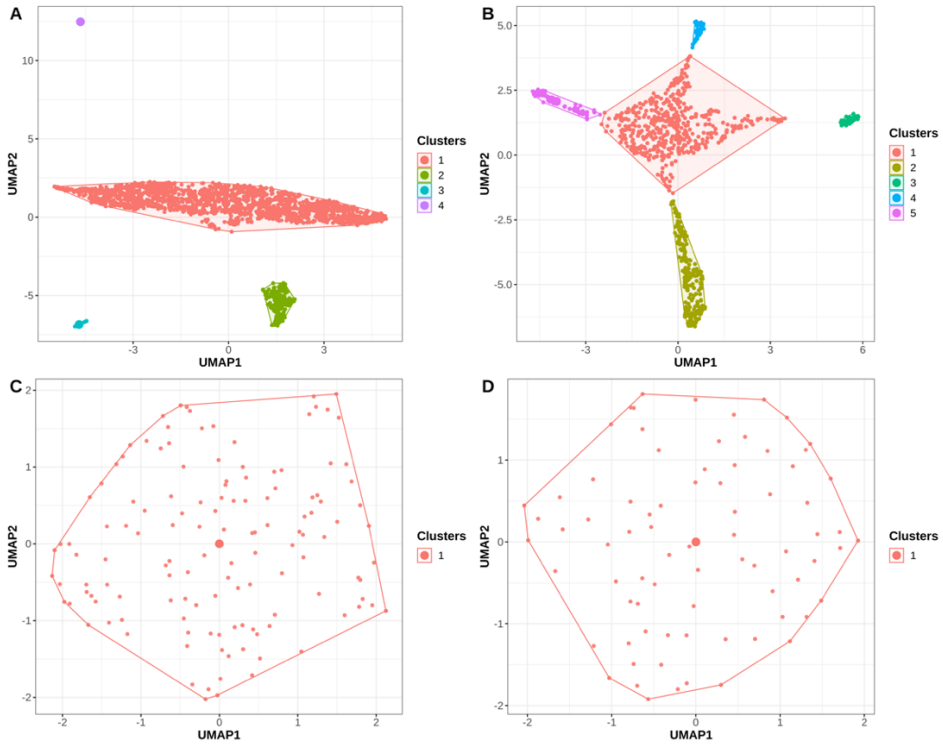


Figure S1. UMAP clustering with DBScan for each dataset of PFE.
A. Ashkenazi Jews. B. Quebec. C. Himba. D. Hutterites

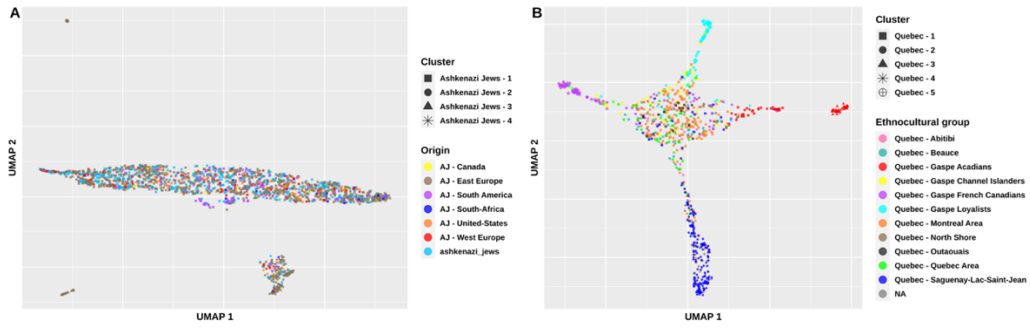


Figure S2. UMAP colored according to the origin or ethnocultural group and shaped according to the clustering.
A. Ashkenazi Jews. B. Quebec. For panel A, Ashkenazi Jews have reached these locations already long after the nearly complete mixing of the population in Eastern and Central Europe. Therefore, no genetic differences were expected across these locations (6).

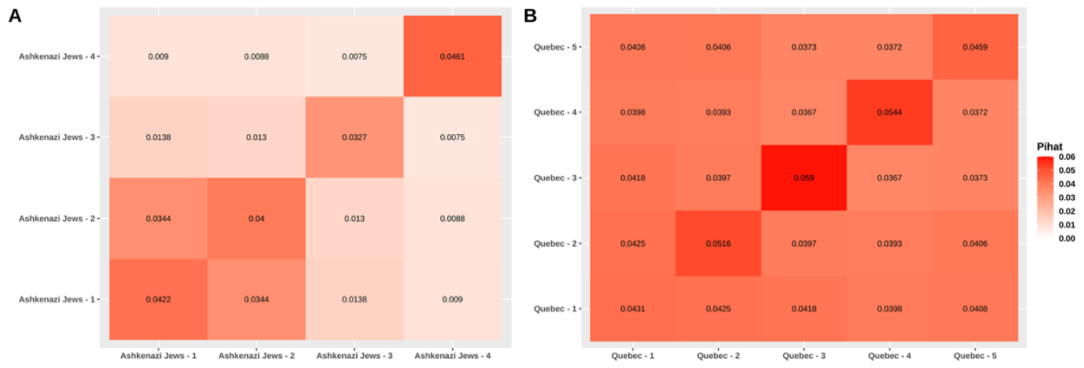


Figure S3. Heatmap of the averaged genetic relatedness (PLINK pihat) between and within clusters.

A. Ashkenazi Jews. B. Quebec.

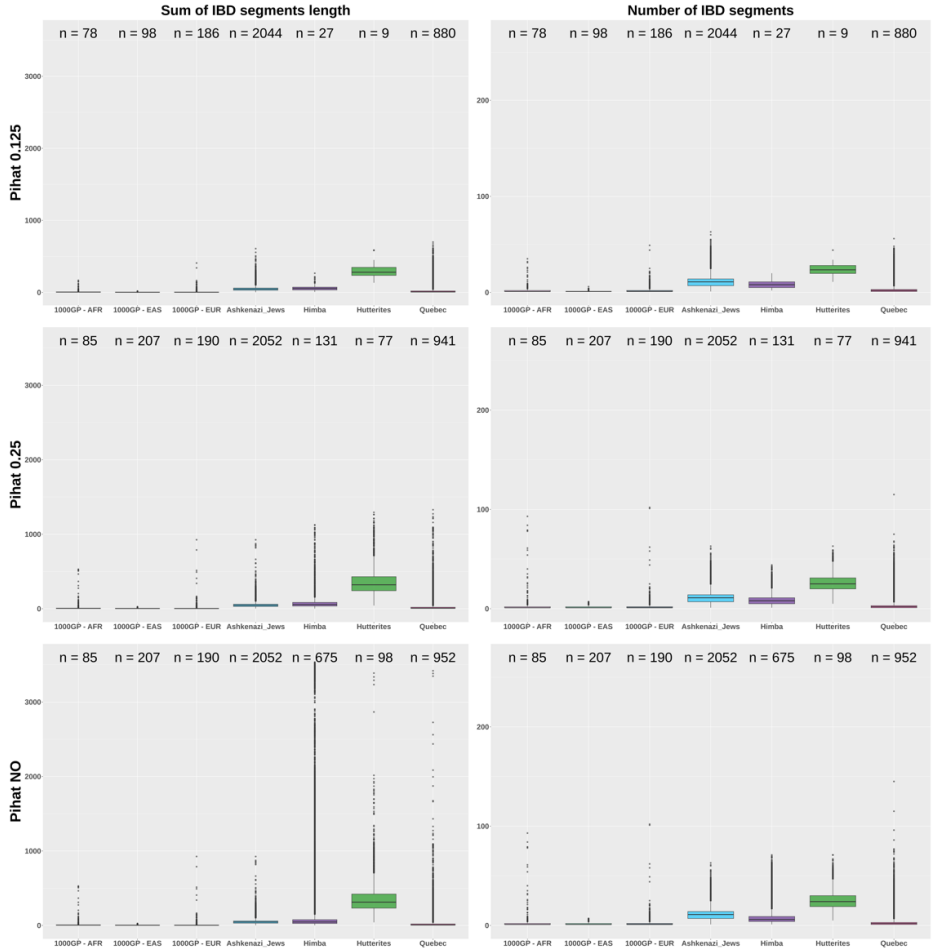


Figure S4. Pairwise sum of IBD segments length and number of IBD segments according to different genetic relatedness filters (pihat 0.125, pihat 0.25 and no cleaning).

Supplementary tables

Table S1. Populations and datasets

Cohort	Sample size	Populations included	Genetic data	Sources
Quebec	941	French Canadian; English-speaking United Empire Loyalists and Acadians of the Gaspe Peninsula.	Illumina Omni Express or Illumina Omni 2.5 chips.	Quebec Regional Reference Sample (6) and unaffected individuals from the Saguenay-Lac-St-Jean asthma familial cohort (7).
Ashkenazi Jews	2,052	Ashkenazi Jews.	Illumina HumanOmni1-Quad arrays.	dbGaP study accession number phs000448.v1.p1 (8).
Himba	131	Himba.	Illumina MEGAex and H3Africa arrays.	dbGaP study accession number phs001995.v1.p1 (9).
Hutterites	77	Hutterites.	SNP genotypes derived from sequence data.	dbGaP study accession number phs000185.v8.p1(10).
African reference group	85	Mende of Sierra Leone (MSL).	SNP genotypes derived from sequence data.	1000 Genomes Project (11).
European reference group	190	British (GBR), Northern Europe and Western Europe (CEU).	SNP genotypes derived from sequence data.	1000 Genomes Project (11).
East Asian reference group	207	Han Chinese (CHB) and Japanese (JPT).	SNP genotypes derived from sequence data.	1000 Genomes Project (11).

Table S3. Mean, maximum and minimum proportion of pairs sharing an IBD segment through the genome for each PFE, clusters and reference population.

Population	Mean (%)	Maximum (%)	Minimum (%)
Himba	2.182	3.464	0.611
Hutterites	10.309	14.730	2.461
Quebec	0.396	0.819	0.073
Quebec - 1	0.251	0.606	0.032
Quebec - 2	2.130	3.630	0.359
Quebec - 3	4.897	10.106	1.207
Quebec - 4	3.013	8.078	0.595
Quebec - 5	1.653	3.296	0.234
Ashkenazi Jews	1.211	1.770	0.081
Ashkenazi Jews - 1	1.524	2.196	0.105
Ashkenazi Jews - 2	0.261	0.633	0.012
Ashkenazi Jews - 3	2.520	10.256	0.128
Ashkenazi Jews - 4	5.986	27.895	0.526
1000GP - EUR	0.094	1.470	0.006
1000GP - EAS	0.048	1.351	0.005
1000GP - AFR	0.171	1.261	0.028

Table S4. Number of individuals remaining according to different genetic relatedness filters.

1000GP - AFR	1000GP - EAS	1000GP - EUR	Ashkenazi Jews	Himba	Hutterites	Quebec	Genetic relatedness
85	207	190	2052	675	98	952	No filter
85	207	190	2052	131	77	941	0.25
78	98	186	2044	27	9	880	0.125

Supplementary references

1. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet. 2007 Sep;81(3):559–75.

2. Nimgaonkar VL, Fujiwara TM, Dutta M, Wood J, Gentry K, Maendel S, et al. Low prevalence of psychoses among the Hutterites, an isolated religious community. Am J Psychiatry. 2000 Jul;157(7):1065–70.

3. Scelza BA. Female mobility and postmarital kin access in a patrilocal society. Hum Nat. 2011 Dec;22(4):377–93.

4. 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 Feb;590(7845):290–9.

5. Browning BL, Browning SR. Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics. 2013 Jun;194(2):459–71.

6. Gagnon L, Moreau C, Laprise C, Vézina H, Girard SL. Deciphering the genetic structure of the Quebec founder population using genealogies. European Journal of Human Genetics. 2024 Jan 1;32(1):91–7.

7. Laprise C. The Saguenay-Lac-Saint-Jean asthma familial collection: the genetics of asthma in a young founder population. Genes Immun. 2014 Apr;15(4):247–55.

8. Lencz T, Guha S, Liu C, Rosenfeld J, Mukherjee S, DeRosse P, et al. Genome-wide association study implicates NDST3 in schizophrenia and bipolar disorder. Nat Commun. 2013 Nov 19;4(1):2739.

9. Scelza BA, Prall SP, Swinford N, Gopalan S, Atkinson EG, McElreath R, et al. High rate of extrapair paternity in a human population demonstrates diversity in human reproductive strategies. Science Advances. 2020 Feb 19;6(8):eaay6195.

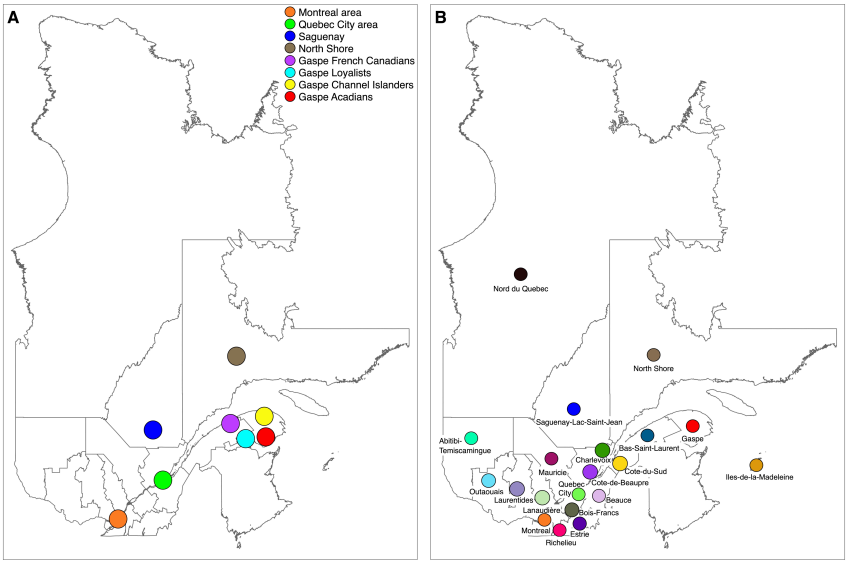
10. Ober C, Nord AS, Thompson EE, Pan L, Tan Z, Cusanovich D, et al. Genome-wide association study of plasma lipoprotein(a) levels identifies multiple genes on chromosome 6q. J Lipid Res. 2009 May;50(5):798–806.

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

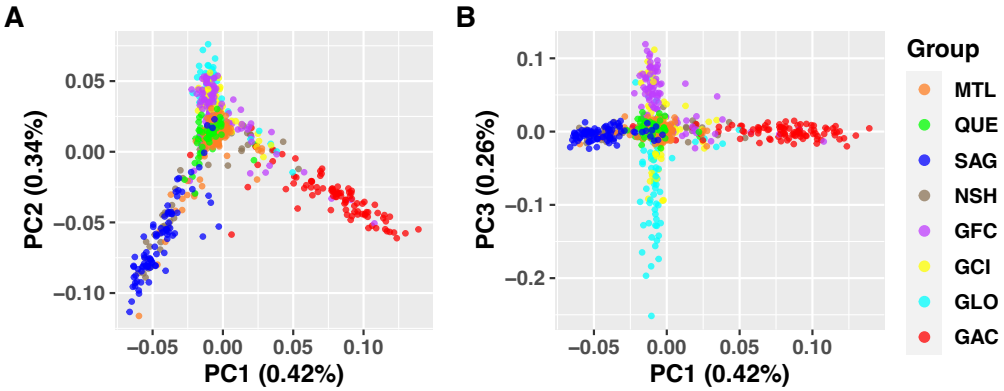
Annexe 2

L'Annexe 2 comporte les informations supplémentaires de l'article *Deciphering the genetic structure of the Quebec founder population using genealogies* présentées au Chapitre 2. Toutefois, le tableau supplémentaire 1 est disponible sur le site internet de la version web de l'article en raison de sa taille (https://www.nature.com/articles/s41431-023-01356-2).

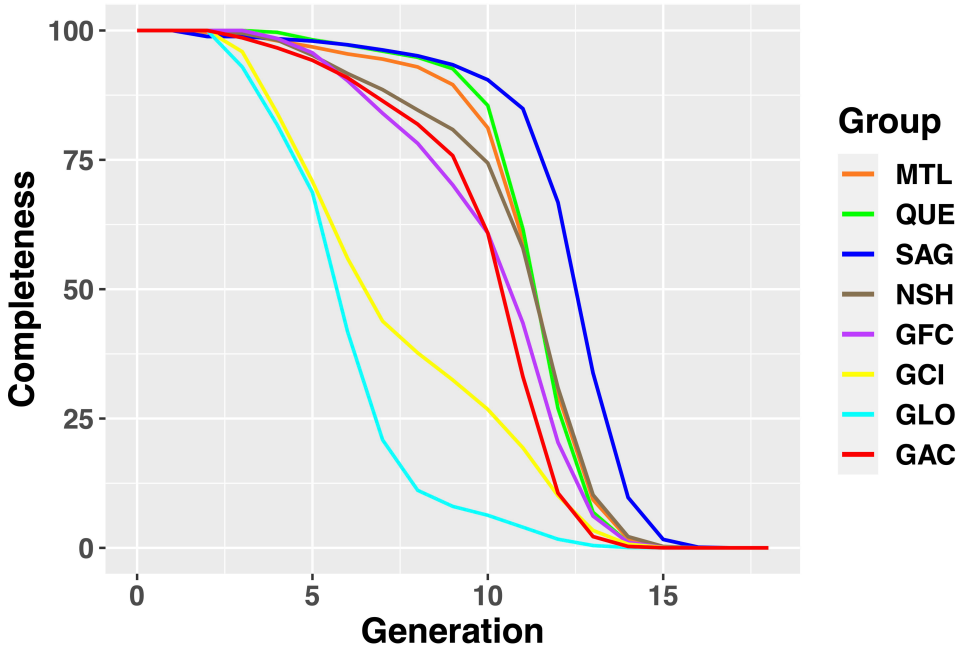
Supplemental Information



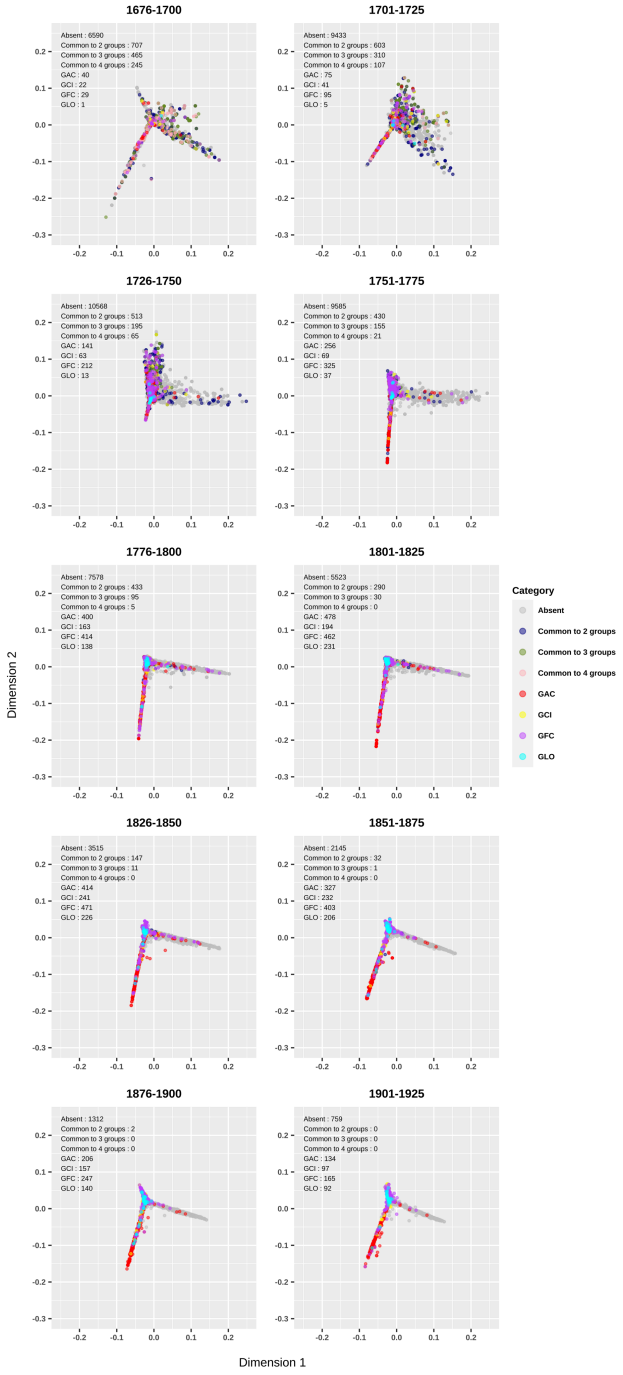
Supplementary Fig. S1. Geographical location of groups investigated in this study (A) and parents' regions of marriage used in figure 1 (B)



Supplementary Fig. S2. Principal component analysis of genotype data
Dots are colored according to the ethnocultural or regional group. Sample sizes are reported in Table 1 of the main text.
GAC=Gaspé Acadians ; GCI=Gaspé Channel Islanders ; GFC=Gaspé French Canadians ; GLO=Gaspé Loyalists ; MTL=Montreal ; NSH=North Shore ; QUE=Quebec City ; SAG=Saguenay.



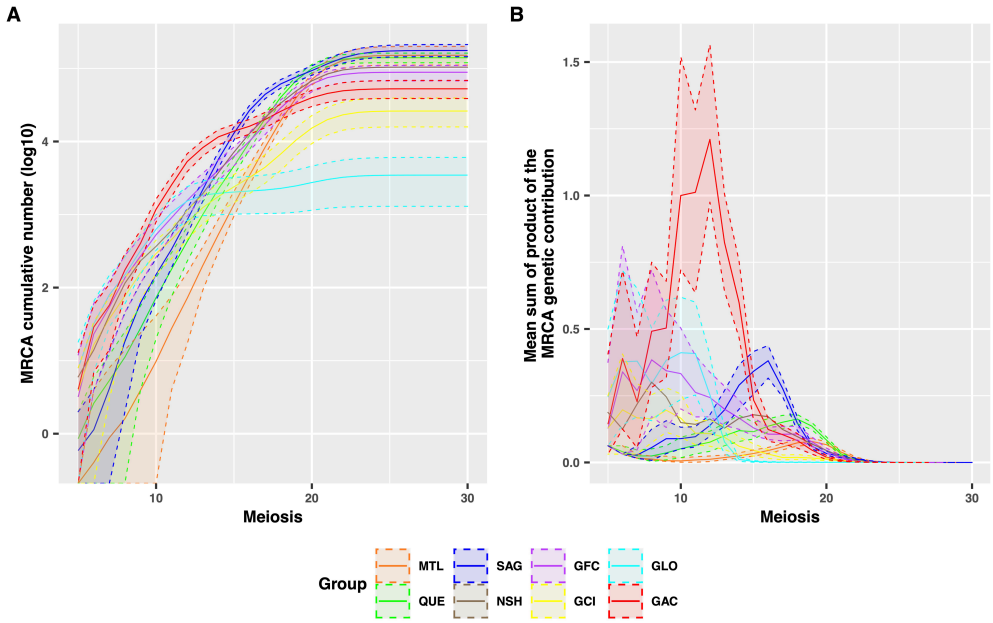
Supplementary Fig. S3. Contemporary groups' mean completeness per generation
The completeness is the proportion of ancestors present in the genealogy at each generation compared to the maximum possible number of ancestors. Sample sizes are reported in Table 1 of the main text.
GAC=Gaspé Acadians ; GCI=Gaspé Channel Islanders ; GFC=Gaspé French Canadians ; GLO=Gaspé Loyalists ; MTL=Montreal ; NSH=North Shore ; QUE=Quebec City ; SAG=Saguenay.



Supplementary Fig. S4. Multidimensional scaling (MDS) of the pairwise kinship coefficients of ancestors of Gaspé groups per 25-year period

MDS was performed on the pairwise kinship distance matrix, (i.e., 1-kinship coefficient) of ancestors whose parents were married at each period. The pairwise kinship coefficient was computed using the R GENLIB library at the maximal depth. Dots were colored according to the contemporary Gaspé group. If the ancestors happened to be the ancestor of more than one group, they were colored accordingly.

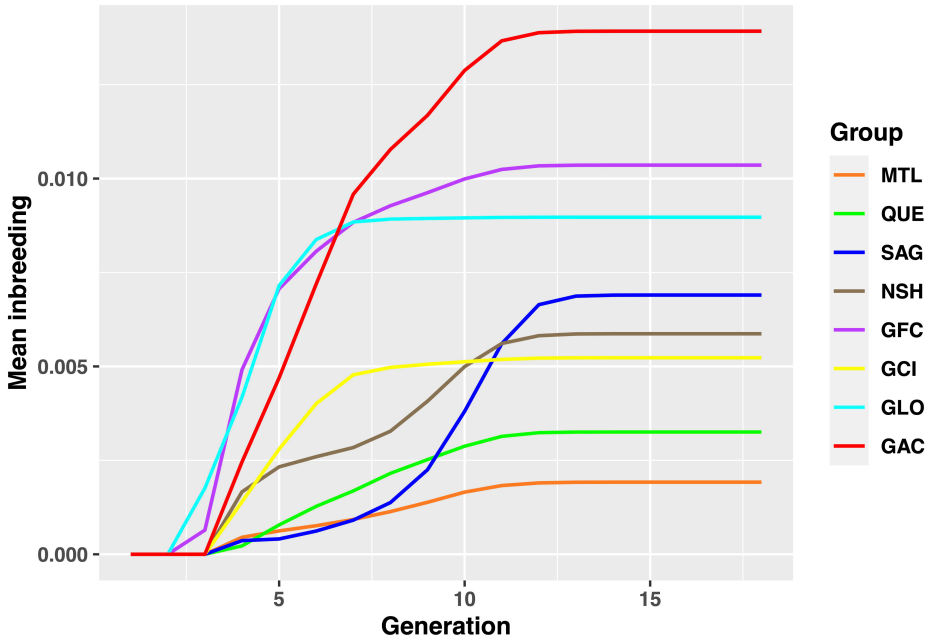
GAC=Gaspé Acadians ; GCI=Gaspé Channel Islanders ; GFC=Gaspé French Canadians ; GLO=Gaspé Loyalists.



Supplementary Fig. S5. MRCA cumulative count (A) and genetic contribution (B) per meiosis within groups

The solid line represents the mean and the dashed lines are the maximum and minimum values of 1,000 bootstraps of 47 individuals.

GAC=Gaspé Acadians ; GCI=Gaspé Channel Islanders ; GFC=Gaspé French Canadians ; GLO=Gaspé Loyalists ; MTL=Montreal ; NSH=North Shore ; QUE=Quebec City ; SAG=Saguenay.



Supplementary Fig. S6. Mean inbreeding coefficient of contemporary subjects per group per generation

Mean inbreeding was calculated on the contemporary subjects of each group at each generation depth (x axis) using GENLIB. The inbreeding coefficient depends on the number of common ancestors present in both parents' genealogy. Close inbreeding (until ~4 generations) provides information on the choice of a spouse while distant inbreeding rather reflects the demographic history of the population. Sample sizes are reported in Table 1 of the main text.

GAC=Gaspé Acadians ; GCI=Gaspé Channel Islanders ; GFC=Gaspé French Canadians ; GLO=Gaspé Loyalists ; MTL=Montreal ; NSH=North Shore ; QUE=Quebec City ; SAG=Saguenay.

Supplementary Table S2. Cumulative MRCA counts
Mean of 1,000 bootstraps of 47 individuals.
GAC=Gaspé Acadians ; GCI=Gaspé Channel Islanders ; GFC=Gaspé French Canadians ; GLO=Gaspé Loyalists ; MTL=Montreal ; NSH=North Shore ; QUE=Quebec City ; SAG=Saguenay.

	6	**8**	**10**	**12**	**14**	**16**	**18**	**20**	**22**	**24**	**26**	**28**	**30**
MTL	0	2	10	72	534	3441	19239	70845	129160	148650	151048	151164	151165
QUE	3	11	73	428	2238	9779	40217	100058	134125	139772	140152	140154	140154
SAG	1	18	151	845	5701	26733	61315	93452	139673	168269	175136	175891	175902
NSH	14	115	367	1142	3756	12913	33438	67604	95445	103129	104070	104097	104097
GFC	23	136	529	1559	4048	10120	28829	61358	82731	87996	88586	88612	88612
GCI	17	77	292	783	1767	3214	6734	15264	23194	25801	26118	26119	26119
GLO	32	150	630	1535	1954	2100	2281	2740	3239	3430	3462	3463	3463
GAC	29	175	1181	5322	11577	16182	25855	39748	49457	52188	52614	52621	52621