UQAC
Université du Québec
à Chicoutimi

Recherche de maladies à effet fondateur par l'étude de variants pathogènes enrichis dans la population du Saguenay

par Elisa Michel

Mémoire présenté à l'Université du Québec à Chicoutimi en vue de l'obtention du grade de Maîtrise ès sciences (M. Sc.) en santé durable

Québec, Canada

Résumé

Le Saguenay–Lac-Saint-Jean (SLSJ), une région localisée dans la province de Québec, Canada, permet l'étude d'une population unique en lien avec l'histoire de son peuplement. Le travail effectué pendant des décennies par les cliniciens avec notamment des diagnostics de patients permet aujourd'hui d'affirmer sans nul doute que cette région présente une prévalence élevée de certaines maladies génétiques connues pour être rares ailleurs dans le monde. Ce fait doit être mis en lien avec le peuplement de la région afin d'en comprendre l'origine. En effet, cette population est issue de groupements d'individus ayant migré d'abord de France pour arriver au Québec puis dans la région de Charlevoix et enfin du SLSJ. Ces migrations de groupes d'individus limités en nombre dans une autre région et ayant eu de nombreux descendants sont à l'origine de l'effet fondateur. On peut donc constater une augmentation en fréquence de certains allèles en comparaison à d'autres populations sans que la consanguinité n'en soit la cause. Ainsi, en s'appuyant sur la base de données CARTaGENE qui regroupe les données génétiques de nombreux individus du Québec, l'objectif de ce mémoire est de mettre en évidence des variants pathogènes plus fréquents, et cela en lien avec l'effet fondateur, dans cette population. Lors de ce projet, des variants pathogènes plus présents au Québec ont été mis en évidence puis analysés génétiquement par l'étude des segments identiques-par-descendance (IBD) partagés entre les individus porteurs. Il a ainsi été possible de mettre en évidence le caractère fondateur de variants par une forte proportion de segments IBD partagés au niveau de ces derniers parmi des individus non-apparentés. Cela signifie que ce segment d'ADN provient d'un seul ou de très peu d'ancêtres qui l'ont transmis à leur descendance il y a plusieurs générations. Parmi les variants déterminés comme fondateurs lors de mon étude, certains étaient déjà répertoriés et connus pour être plus fréquents dans la région. Cependant, mes analyses ont aussi permis la détermination de variants fondateurs qui n'ont jamais été répertoriés jusqu'à ce jour au SLSJ. L'approche de mon travail est originale car au lieu de regarder au niveau des symptômes déclarés du patient comme le font depuis de nombreuses années les cliniciens, je me base sur l'étude du génome d'individus sélectionnés au hasard dans la population afin de déterminer avec précision les variants pour lesquels la fréquence est élevée dans la population. Ainsi, mon projet met en lumière de nouveaux variants à évaluer lors de considérations portant sur l'étude de la population ou le domaine de la santé publique. Cela pourrait permettre un meilleur diagnostic des maladies rares au SLSJ et donc une meilleure prise en charge des symptômes. Ainsi, ce travail pourrait contribuer à la mise en place d'outils de prévention comme il en existe déjà avec les tests de porteurs proposés pour les quatre maladies considérées fondatrices au SLSJ.

Abstract

The Saguenay–Lac-Saint-Jean (SLSJ) region in Quebec, Canada, provides a unique opportunity to study a population shaped by its distinct historical settlement. Decades of work by clinicians in identifying and diagnosing conditions have revealed that the region now exhibits higher frequencies of certain genetic diseases known to be rare worldwide. To fully grasp this phenomenon, we must examine the settlement history of the region. This population traces its origins to a limited number of French migrants who initially settled in Quebec, then moved to Charlevoix, and eventually to the SLSJ region. These migrations, involving a limited number of individuals moving to a new region followed by rapid population expansion with large families, form the basis of the founder effect. As a result, we observe an increase in the frequency of certain alleles compared to other populations, despite the fact that consanguinity does not play a role. Utilizing the CARTaGENE database, which contains genetic data from numerous individuals across Quebec, this study aims to demonstrate the presence of pathogenic variants whose frequencies have increased due to the founder effect. In this project, pathogenic variants

with elevated frequencies in Quebec were identified and subsequently analyzed genetically by examining the identical-by-descent (IBD) segments shared among individuals carrying each specific variant. This approach made it possible to characterize variants as founders by identifying a high proportion of IBD segments shared at the variant's location among unrelated individuals. This indicates that the DNA segment originated from a single or few ancestors who passed it down to their descendants many generations back. Among the variants characterized as founders in this study, some were already described in the literature and are known to be more frequent in the region. However, this study also uncovered novel founder variants that had not been previously reported. The approach taken in this study is innovative; rather than relying on declared symptoms from patients, as clinicians have traditionally done, I analyzed genetic data to precisely identify variants that are highly prevalent in the population. Thus, this project highlights new variants that could be important for consideration in public health discussions or population studies. This may lead to improved diagnosis of rare diseases in SLSJ and, consequently, better support in daily life. This could be achieved through prevention efforts, similar to those already in place for four diseases considered to be of founder origin, for which carrier testing is offered.

Table des matières

Liste des tableaux

Liste des figures

Liste des abréviations

ADN……Acide désoxyribonucléique

ARN……Acide ribonucléique

ARNm……Acide ribonucléique messager

ARNt……Acide ribonucléique de transfert

IBD……Identique par descendance (identity-by-descent)

QcP……Province de Québec

SLSJ……Saguenay–Lac-Saint-Jean

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

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

MAF……Fréquence de l'allèle minoritaire (Minor Allele Frequency)

MRCA……Ancêtres communs les plus proches (Most Recent Common Ancestor)

SNP……Polymorphisme nucléotidique simple (Single nucleotide polymorphism)

Remerciements

Je tiens à remercier mon directeur de recherche Simon Girard qui m'a accueillie avec bienveillance dans son laboratoire et m'a permis de travailler sur un projet fort intéressant. Je souhaite remercier Claudia Moreau également pour m'avoir guidée tout au long de cette année et qui m'a été d'une aide sans limite. De plus, j'aimerais remercier Laurence Gagnon pour m'avoir épaulée et instruite alors que je découvrais le monde de la bio-informatique. Merci d'avoir été d'une patience sans égale. Je tiens aussi à remercier tous mes collègues du Genopop avec qui j'ai pu partager mon aventure au labo et découvrir ensemble le Québec.

De plus, j'aimerais remercier du plus profond de mon cœur mes proches et plus particulièrement Martin qui m'a soutenue et encouragée chaque jour et chaque minute de ma maîtrise. Merci à ma famille d'avoir été présente malgré la distance qui nous séparait.

Enfin, j'aimerais remercier tous les participants du projet c'est-à-dire aussi bien les personnes travaillant à la clinique qui nous ont aidé dans la réalisation de celui-ci grâce à leurs données et expertise. Mais aussi à tous les individus qui constituent les cohortes présentées dans ce projet et sur lesquelles nous nous sommes basées pour nos analyses. Sans leur participation, ce projet n'aurait pas été possible, je tiens donc à les remercier tous.

Avant-propos

Le mémoire ici présent est structuré selon cinq sections. La première partie est intitulée introduction et a pour but de présenter toutes les notions de base nécessaires à la compréhension des parties suivantes et donc du projet global. Ainsi, je vais aborder les concepts de génétique ainsi que les analyses possibles avec des outils spécifiques. De plus, je vais mettre en lien le peuplement de la province de Québec avec les implications contemporaines présentes dans la population.

Ensuite, lors de la deuxième partie nommée chapitre 1, je vais présenter les données et méthodes utilisées lors de ce projet.

De plus, le chapitre 2 présentera l'article scientifique écrit lors de ma maîtrise sur le projet.

La quatrième partie portera sur la discussion des sections précédemment présentées. Mais aussi, elle permettra d'aborder les perspectives et limites du projet.

Pour finir, la dernière section permettra de conclure sur l'ensemble de mon projet et d'avoir un retour avec du recul sur ce travail.

Introduction

Les caractéristiques d'une population sont intrinsèquement liées à la diversité génétique de cette dernière. Il est donc primordial de prendre en considération la structure génétique d'une région donnée afin de mettre en évidence ces spécificités génétiques en lien avec son peuplement. Pour cela, il faut connaître les mécanismes sous-jacents en génétique et démographie afin de mieux comprendre les caractéristiques de la population étudiée.

1. Introduction au concept de génétique

Au commencement de la génétique et sans les outils contemporains, celle-ci se définissait par l'étude de traits spécifiques transmis à la descendance. Notamment, ce fut le moine Gregor Mendel qui mit en avant « l'existence de facteurs discrets qui transmettent l'information du développement d'un parent à ses descendants » et cela par l'étude de différentes souches de petits pois en 1865[1]. Aujourd'hui et grâce aux outils modernes, la génétique permet de mieux comprendre les mécanismes biologiques complexes qui font ce que nous sommes ainsi que l'évolution des êtres vivants ou des populations au cours du temps.

1.1. Le fonctionnement cellulaire

L'être humain est composé de nombreuses cellules permettant le fonctionnement de systèmes vitaux comme l'oxygénation des organes ainsi que des systèmes nécessaires à la reproduction comme la production de gamètes[2]. Cela est rendu possible en partie grâce aux protéines permettant l'exécution de fonctions diverses comme le maintien de l'intégrité de la cellule ou encore des transferts intercellulaires[3].

Les protéines sont constituées d'acides aminés reliés entre eux par différents types de liaisons permettant diverses configurations 3D. La forme d'une protéine joue un rôle très important dans la fonctionnalité de celle-ci. Les protéines sont produites lors de la traduction d'une séquence d'acide ribonucléique messager (ARNm) donnée. Ainsi, les ARN de transfert (ARNt) vont apporter aux ribosomes des acides aminés qui seront liés dans un ordre précis et donné par l'ARNm[3].

L'ARNm est formé lors de la transcription de l'acide désoxyribonucléique (ADN). Cela peut être fait dans le noyau, et dans ce cas l'ARNm est transporté dans le cytoplasme, ou bien dans l'organite appelé mitochondrie qui présente aussi de l'ADN qui lui est spécifique[4].

L'ADN est une molécule avec une structure tridimensionnelle stable dans l'espace et dans le temps dont la représentation est une hélice à double brin. La structure de l'ADN n'a été mise en avant que très récemment par les chercheurs Watson et Crick en 1953 grâce notamment au travail de Rosalind Franklin[1]. C'est une séquence contenant les nucléotides suivants : A, T, C, G. Ceux-ci sont composés du sucre désoxyribose, d'une base (adénine, cytosine, thymine et guanine) et d'un phosphate. Cette séquence est ordonnée avec un sens de lecture précis de telle façon que 3 nucléotides correspondent à un acide aminé spécifique. Ainsi une certaine séquence d'ADN complète peut correspondre à une protéine avec une fonction précise. Dans le cas de l'ARNm, le sucre est remplacé par un ribose, la base thymine par l'uracile U et la molécule est monobrin ce qui la rend plus instable[4]. Les molécules d'ADN forment le génome qui est visible sous sa forme la plus condensée par 22 paires de chromosomes autosomiques et de deux chromosomes sexuels X,Y dans le cas d'un homme et X,X dans le cas d'une femme. Celui-ci est présent dans chaque noyau de cellule nucléée.

Certaines régions du génome sont codantes et vont mener à la formation de protéines. D'autres régions du génome ne le sont pas et permettent d'autres fonctions. Par exemple, les régions télomériques ou les centromères permettent de structurer le chromosome. Mais aussi, certaines régions donnent lieu à l'expression d'ARN avec divers rôles comme les microARN qui régulent l'expression des gènes[4].

On peut considérer le génome comme un manuel d'instructions dont les cellules s'aident pour réaliser les fonctions vitales à l'humain. Pour un même génome, on peut avoir différents niveaux d'expressions affectant le phénotype résultant, c'est-à-dire des traits que l'on peut voir, en partie en fonction de l'environnement auquel doit s'adapter l'organisme. Cette capacité à

répondre à l'environnement est primordial pour la survie d'une espèce. En effet, des caractéristiques spécifiques, qui peuvent être basées sur le phénotype par exemple, pourront permettre ou non aux individus de prospérer et ainsi d'avoir une descendance.

1.2. La sélection naturelle

De ce fait, les individus vont changer au cours des générations et cela grâce à des modifications au niveau de leur génome. Ces modifications de l'ADN sont nécessaires à l'adaptation d'une espèce à son environnement. En effet, ces changements peuvent entraîner une modification phénotypique, se traduisant par l'apparition d'un trait qui est soumis à la pression de la sélection naturelle.

Notamment, un trait héritable augmentant les chances de reproduction d'un individu, car celui-ci est plus adapté à son environnement, permettra à ce dernier de plus se reproduire et former ainsi une descendance elle-même fertile[5]. Les descendants de cet individu vont hériter de ce trait et permettre la diffusion de ce dernier par la reproduction et ainsi de suite tant que le trait est favorable à la reproduction des individus dans un environnement donné[6]. Il est possible que le trait entraîne une diminution de la capacité de survie ou de reproduction d'un individu. Dans ce cas, sous la pression de la sélection naturelle, ce trait tendrait à disparaître au cours des générations. On appelle ces modifications de l'ADN des mutations et il en existe diverses formes.

1.3. Le concept de mutation

On sait qu'une grande proportion du génome est partagée entre tous les êtres humains. En effet, lorsqu'on prend deux individus aléatoirement dans le monde, leurs génomes ne vont différer que de 0.1%[7]. Ainsi, ce ne sont que ces différences qui font la diversité de l'espèce mais aussi qui lui permettent de s'adapter à son environnement au cours du temps. Or, ces différences apparaissent lors d'un phénomène aléatoire appelé la mutagénèse, ce qui signifie l'apparition de mutations[8], appelées aussi variants. En effet, lors de la vie d'une cellule, des changements de séquence peuvent apparaître spontanément malgré la stabilité de la molécule d'ADN. Cela est notamment possible lors de la division d'une cellule mère en deux cellules filles, soit au moment

de la formation des gamètes qu'on appelle méiose. Pour les cellules germinales, cela entraîne la transmission de la mutation de novo chez le zygote. Cela peut se traduire par un changement d'un seul nucléotide à un endroit précis du génome, appelé locus, ou bien par de plus grandes variations lors d'insertions ou de délétions[9,10]. Comme on peut le voir sur la figure 1 ci-dessous, un changement de nucléotide appelé substitution peut mener à l'apparition d'un autre acide aminé que celui de référence. Cela peut entraîner une mauvaise conformation de la protéine et ainsi la rendre défectueuse[11]. De plus, un changement de nucléotide peut aussi mener à la formation d'un codon stop qui entraîne l'arrêt de la traduction de l'ARNm en protéine. Ensuite, les autres types de mutations telles que les insertions et les délétions mènent à un décalage lors de la lecture de l'ARNm. Cela peut avoir de lourdes conséquences sur la fonctionnalité de la protéine car potentiellement tous les acides aminés suivant la mutation seront changés. Ainsi les mutations présentées dans ce paragraphe mènent à des changements au niveau phénotypique qui peuvent permettre l'adaptation d'une espèce à son environnement mais aussi entraîner des maladies dites génétiques.

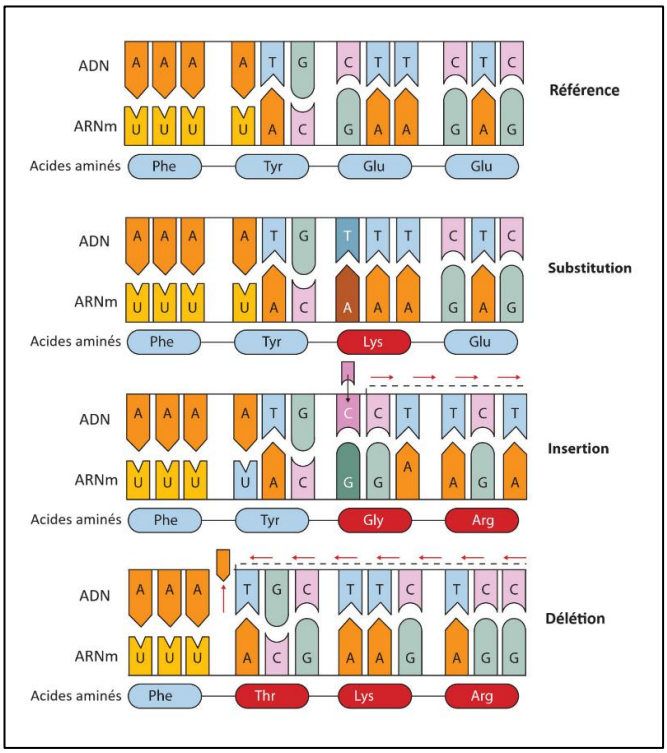

Figure 1 : Représentation des divers types de mutations.

Les 3 types sont représentés : Substitution, Insertion et délétion. Tiré de 19.5 Mutations and Genetic Diseases | The Basics of General, Organic, and Biological Chemistry. https://courses.lumenlearning.com/suny-orgbiochemistry/chapter/19-5-mutations-and-genetic-diseases/ (accessed 2024-10-16).

1.4. Maladies génétiques rares

On parle de maladie génétique lorsqu'une mutation sur un locus entraîne un dysfonctionnement au niveau d'un gène ou d'une région responsable de sa régulation[12]. Or ces copies du locus, qui peuvent être mutées ou non, sont héritées des parents qui donnent chacun une de leurs deux copies. Ainsi, si une mutation est présente ou apparaît au niveau de cellules permettant la production des gamètes aussi appelées les cellules germinales, alors cette dernière sera transmise à la descendance. Les symptômes peuvent se manifester lorsqu'une personne présente les 2 copies de la mutation à un locus donné. Cela signifie que la personne présente la mutation sur chacun des deux chromosomes de la paire. Dans ce cas, on parle d'une personne homozygote présentant une maladie récessive. Dans le cas d'une maladie récessive, une personne ne présentant qu'un seul exemplaire de la mutation ne sera pas atteinte de la maladie, on dit qu'elle est porteuse.

Lorsqu'une maladie génétique se déclare chez un individu présentant une seule copie de la mutation, soit sur un des chromosomes formant une paire, on parle de maladie génétique dominante[4].

Ainsi, dans le cas d'une maladie récessive, lorsque les deux parents sont porteurs de l'allèle responsable de la maladie, il y a un risque pour que l'enfant soit atteint et donc présente les deux copies. Dans le cas d'une maladie dominante, une seule copie est suffisante pour être malade. Ainsi, il suffit qu'un seul des parents soit atteint c'est-à-dire porteur de l'allèle pour qu'il y ait un risque d'avoir un enfant atteint aussi.

Il est de coutume de catégoriser les maladies génétiques selon la prévalence de ces dernières. En effet, la prévalence est un outil très utile qui indique le nombre de personnes atteintes d'une

maladie dans une population donnée et à un temps donné, en prenant en compte aussi bien les cas déjà recensés que les nouveaux.

Aujourd'hui, il est difficile de donner une définition globale et précise d'une maladie génétique rare. En effet, au niveau mondial, il existe différentes définitions selon les organisations. Par exemple, depuis 1999, le règlement de l'Union Européenne décrit une maladie rare comme une maladie touchant moins de 5 personnes sur 10 000[13]. Mais aussi, la loi américaine sur les médicaments orphelins datant de 1983 considère une maladie comme rare si celle-ci touche moins de 200 000 personnes dans le pays[14]. De plus, ces maladies sont nombreuses et présentent des mécanismes complexes et variés. Du fait de la rareté de la maladie, on dénombre peu de patients atteints. Ainsi, établir un diagnostic est difficile car la maladie est peu décrite ou peu étudiée. C'est notamment pour cela qu'aucune définition faisant consensus n'existe pour le moment[15]. Ainsi, afin de mieux caractériser ce qu'est une maladie rare, il est important d'estimer correctement la proportion d'individus atteints. Pour cela, nous pouvons utiliser des modèles mathématiques dans le but de se rapprocher d'une estimation de la présence de mutations pathogènes à partir d'un échantillon de la population afin de mieux dénombrer les individus atteints et améliorer leur prise en charge. En effet, si on conclut qu'un variant pathogène est fréquent dans une population, la maladie liée à ce variant ainsi que ces symptômes peuvent être surveillés et ainsi améliorer les diagnostics des patients.

1.5. Principes de Hardy-Weinberg

Ainsi, on peut calculer la fréquence attendue des génotypes dans une population donnée grâce au principe de Hardy-Weinberg[16]. Dans ce cas, il est possible de calculer la fréquence de l'allèle minoritaire nommé MAF (Minor Allele Frequency). On considère qu'une mutation est rare lorsque celle-ci a une MAF inférieure à 0.01 soit 1% ou 0.05 soit 5% selon les sources[17,18].

Le principe de Hardy-Weinberg part des hypothèses suivantes :

- Les organismes sont diploïdes, ce qui signifie que dans le génome, les chromosomes sont par paire.
- La reproduction est sexuée.
- Les générations ne se superposent pas.
- Le gène étudié présente deux allèles.
- Les fréquences alléliques sont identiques chez les hommes et les femmes.
- Le choix du partenaire est aléatoire.
- La taille de la population est infinie.
- Les flux migratoires sont négligeables.
- L'apparition de mutations peut être ignorée.

Ces hypothèses sont posées afin d'appliquer un modèle mathématique. Cependant les conditions réelles ne permettent pas de satisfaire ces dernières. Ainsi, notre application est théorique et permet seulement d'avoir une approximation.

Ainsi, lorsqu'on considère un gène avec les allèles suivants, A majoritaire et a minoritaire, on obtient :

- AA : p^2, la fréquence des homozygotes dans la population pour l'allèle majoritaire.
- Aa : $2pq$, la fréquence des hétérozygotes dans la population.
- aa : q^2, la fréquence des homozygotes dans la population pour l'allèle minoritaire.

On obtient la relation mathématique suivante selon les définitions de p et q : $p + q = 1$. On peut donc calculer la fréquence génotypique avec la relation suivante : $p^2 + 2pq + q^2 = 1$.
Ces relations mathématiques sont notamment utilisées afin de déterminer des indicateurs permettant de caractériser une population. Par exemple, avec la formule de Hardy-Weinberg, on peut calculer le taux de porteurs, c'est-à-dire le nombre de personnes qui portent une copie de l'allèle muté dans une population. Pour cela, on utilise la fréquence des hétérozygotes dans la population et on a :

$$taux\ de\ porteurs = \frac{1}{2pq}$$

On parle de l'équilibre de Hardy-Weinberg car les fréquences des allèles A et a ne changent pas au cours du temps selon cette théorie. Or, il existe des pressions qui s'exercent au cours du temps sur les populations qui vont justement modifier la fréquence de leurs allèles. En effet, comme nous l'avons vu plus tôt, nous savons que la sélection naturelle dans une large population et sur une longue période entraîne l'augmentation en fréquence de variants permettant une meilleure adaptation de l'individu à son environnement. Cependant, il existe d'autres mécanismes exerçant des pressions sur les populations et ainsi leurs traits spécifiques. En effet, il est possible de constater une augmentation en fréquence de certains allèles sans que la sélection naturelle y soit favorable lorsque certains paramètres sont réunis.

1.6. La dérive génétique et l'effet fondateur

Dans le cas d'une population de base assez grande et suivant le principe de Hardy-Weinberg, lorsqu'un petit groupe d'individus se sépare de la population initiale afin de migrer, son bagage génétique est défini par les individus composant ce nouveau groupe. Ainsi, lors de la prochaine génération, le nombre limité d'allèles fait qu'un allèle spécifique peut être plus représenté que d'autres dans la descendance simplement par le processus du hasard[16]. Cette pression du hasard sur la répartition des allèles dans la descendance se nomme la dérive génétique. Cela peut donc entraîner de grands changements dans les fréquences des allèles de la population et ainsi convertir un allèle rare en un allèle plus fréquent. Cependant, d'autres allèles peuvent être aussi perdus par changement dû au hasard.

Lorsqu'un sous-groupe d'individus se sépare d'une population donnée afin d'en fonder une nouvelle, le phénomène de goulot d'étranglement de population se produit. Cela se traduit par une perte instantanée d'hétérozygotie dans la nouvelle population. Puis, après un certain nombre de générations, il est possible de constater que la fréquence de certains allèles très présents dans la population source a diminué de façon significative dans le pool génétique de la

nouvelle population. Et inversement, il se peut que la fréquence de certains allèles considérés rares dans la population source augmente drastiquement dans la nouvelle après plusieurs générations. De cette façon, le goulot d'étranglement associé à la dérive génétique ainsi que l'isolement relatif vont entraîner ce que l'on appelle l'effet fondateur[16].

À la suite d'un effet fondateur, la fréquence de mutations peut augmenter de façon notable dans la population ou devenir négligeable comme mentionné auparavant. Parmi les mutations dont la fréquence a augmenté, certaines peuvent être pathogènes et ainsi jouer sur la capacité de reproduction de l'individu, ce qui ne serait pas arrivé avec le processus de sélection naturelle. Nous verrons par la suite que c'est le cas lors du peuplement de la province de Québec. Nous verrons notamment les conséquences de cet effet. Cependant, pour être capable de mettre en évidence de telles conclusions, il faut pouvoir étudier et décrire la génétique des populations contemporaines. Les notions présentées ensuite nous permettront d'accéder à ce type d'analyse.

2. Étude de la variation de la fréquence de certaines mutations par la génétique

Aujourd'hui, grâce aux données auxquelles nous avons accès et aux méthodes d'analyses en génétique, il est possible de démontrer qu'un variant est plus fréquent dans une population grâce à l'effet fondateur.

2.1. Principe de la recombinaison

Afin de mieux comprendre les concepts introduits ici, il faut d'abord connaître certains mécanismes comme celui de la recombinaison. Celui-ci joue un rôle important lors de la reproduction sexuée des organismes. Ici, nous nous intéresserons au cas de l'être humain.

2.1.1. La reproduction sexuée

Ainsi, afin que la reproduction sexuée soit possible, il faut d'abord la création de gamètes. Notamment, pour la femme les gamètes sont les ovules et chez les hommes, ce sont les spermatozoïdes. Ces derniers sont des cellules haploïdes, c'est-à-dire que les chromosomes ne

sont plus par paire. En effet, dans toutes les autres cellules nucléées du corps, soit les cellules dites somatiques, nous avons 44 chromosomes autosomes et 2 chromosomes sexuels assemblés par paires. Or lors de la formation des gamètes, ceux-ci sont produits par un processus appelé méiose. La méiose consiste en deux divisions cellulaires[19]. Au moment de la première division, les paires de chromosomes vont être séparées comme on peut le voir dans la figure 2 lors de la première anaphase. Ensuite, lors de la deuxième division, ce sont les chromosomes, constitués de deux chromatides qui vont se diviser. Ainsi, dans un gamète on trouve seulement une chromatide d'une paire de chromosomes tandis que dans les cellules somatiques il y a les deux chromosomes formant une paire. Ce brassage génétique se produit sur tous les chromosomes autosomiques et les chromosomes sexuels qui vont être répartis aléatoirement au sein des gamètes[20]. Ainsi il y a une très grande diversité de cellules germinales due à l'appariement indépendant des chromosomes. Cependant, il existe un autre processus lors des divisions cellulaires qui permet d'augmenter encore cette diversité.

2.1.2. La recombinaison

Ainsi, on appelle recombinaison le phénomène conduisant à l'apparition, dans une cellule ou dans un individu, de gènes ou de caractères héréditaires dans une association différente de celle observée chez les cellules ou individus parentaux. Cela est possible par le processus de crossing-over lors duquel les chromosomes d'une même paire s'alignent lors de la méiose afin d'être séparés par la suite, ce qui rend possible un transfert de segments d'ADN entre les chromosomes[21]. En effet, comme on peut le voir sur la figure 2 ci-dessous, une seule paire de chromosome est représentée et celle-ci va s'aligner au centre de la cellule. Lors de cet alignement, il est possible qu'un échange ait lieu comme on peut le constater sur la figure mentionnée. Cela signifie qu'une partie du chromosome de la paire va être échangée avec une autre partie du chromosome conjugué. Enfin, à la suite d'étapes de séparation des chromosomes homologues puis des chromatides, on obtient des gamètes uniques et nécessaires à la reproduction sexuée. En effet, grâce à ce processus de séparation des

chromosomes qui permet une répartition aléatoire, mais aussi grâce à la recombinaison, il existe des possibilités très nombreuses de gamètes. Notamment, ici, on représente seulement une paire, mais lors d'une méiose tous les chromosomes autosomiques peuvent être concernés par ce mécanisme, cela augmente donc les combinaisons possibles pour la formation du gamète.

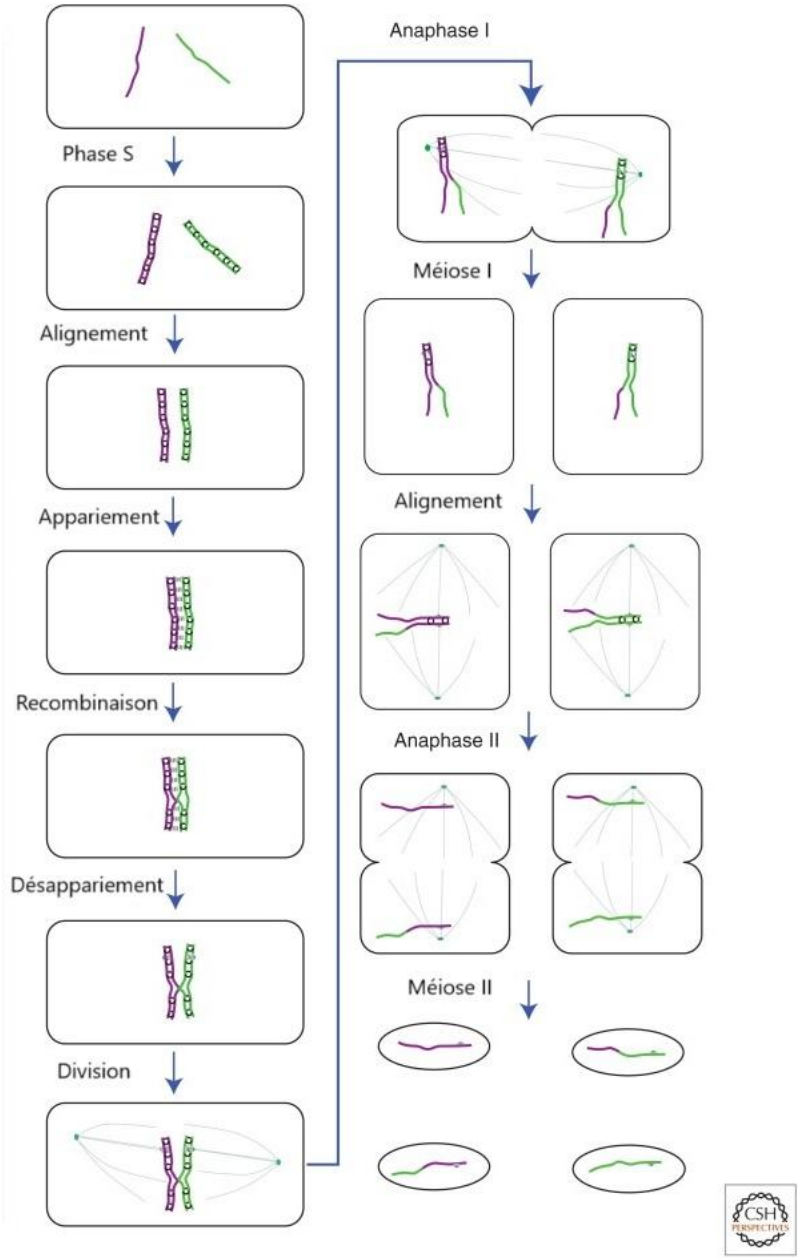

Figure 2 : Représentation de la recombinaison.

Tiré de (avec autorisation) : Hunter N., Meiotic Recombination: The Essence of Heredity. Howard Hughes Medical Institute, Department of Microbiology & Molecular Genetics, Department of Molecular & Cellular Biology, Department of Cell Biology & Human Anatomy, University of California Davis, Davis, California 95616 by Cold Spring Harbor Perspectives in Biology by Cold Spring Harbor Laboratory Press.

Cependant, comme expliqué auparavant, certaines mutations peuvent apparaître lors de la méiose et ainsi être transmises au zygote au moment de la fécondation. Or, il est possible de caractériser l'origine commune d'un variant parmi des individus après plusieurs générations à la suite de l'effet fondateur grâce au processus abordé lors de cette partie.

2.2. Utilisation des segments identiques-par-descendance

Afin d'identifier la transmission du segment d'ADN porteur d'un variant d'intérêt par un ancêtre commun, on utilise un outil appelé le segment identique-par-descendance ou IBD. Celui-ci est une séquence d'ADN qu'on retrouve chez des individus qui l'ont hérité d'un ancêtre commun[22]. En effet, un variant n'est jamais transmis seul, celui-ci est toujours transmis avec d'autres variants formant un segment d'ADN. Comme on peut le voir sur la figure 3 ci-dessous, un ancêtre fondateur présente une mutation représentée par une étoile rouge à un locus donné. Lors de la reproduction sexuée, une copie de chaque chromosome de chaque parent va être transmise à la descendance. Or, comme expliqué auparavant, cette copie est créée lors de la méiose et peut être le résultat de la recombinaison. Ainsi, on retrouve dans la descendance des segments hérités des parents qui sont le fruit d'un brassage génétique. Par la suite, après N générations, il est possible de retrouver ces segments d'ADN hérités de l'ancêtre fondateur chez des individus. En effet, comme on peut le voir sur la figure 3, les homozygotes 1, 2 et 3 présentent le variant fondateur contenu dans le segment IBD hérité de l'ancêtre commun. Or, dépendamment de la valeur de N, les individus peuvent être plus ou moins apparentés. Par exemple, si N est petit, alors le segment IBD aura plus tendance à être long, en lien avec le nombre de recombinaisons qu'il y a eu entre les générations séparant l'ancêtre commun et les individus porteurs du segment. Dans le cas où N est grand, le segment IBD va être généralement court et les individus partageant celui-ci ne seront que très peu apparentés. Cela traduit un ancêtre commun très éloigné pour lequel on doit remonter de nombreuses générations dans la généalogie pour le retrouver.

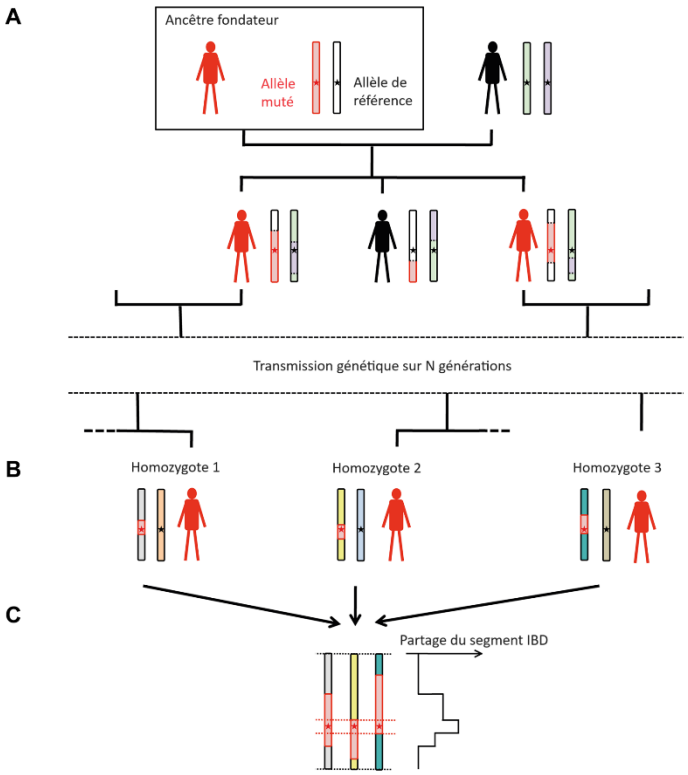

Figure 3 : Principe de l'IBD.

A. Transmission de segments IBD à la descendance. **B.** Résultats de la recombinaison après N générations. **C.** Comparaison des segments IBD de différents individus. Tiré de Letouzé, E., Sow, A., Petel, F., Rosati, R., Figueiredo, B. C., Burnichon, N., Gimenez-Roqueplo, A.-P., Lalli, E., de Reyniès, A., & Mailund, T. (2012). Identity by descent mapping of founder mutations in cancer using high-resolution tumor SNP data. *PLoS ONE, 7*(5). Creative commons license abbreviation.

C'est ainsi que l'on peut faire le lien entre la présence de segment IBD parmi des individus qui ne sont pas de parenté proche et l'effet fondateur. Or, afin de mettre en évidence ce type de relation, il faut d'abord bien caractériser la structure de la population étudiée. Ainsi, dans le prochain paragraphe, nous allons mettre en évidence des méthodes afin de représenter la diversité allélique d'une population dans le but de l'analyser.

3. Représentation de la structure de la population

3.1. Simplification des données

Afin d'étudier les relations entre les génomes des individus d'une population donnée, il est possible de procéder à une analyse par composante principale (PCA). En effet, les données génétiques sont complexes et on cherche donc à en simplifier l'expression tout en gardant l'information[23]. Le but de ce processus est de réduire toutes les dimensions de nos données génétiques afin de pouvoir les visualiser et ainsi d'obtenir une représentation des distances

relatives entre les individus de la population. Chaque point représente un individu et plus deux points sont proches sur ce graphique, plus les deux individus en question sont proches génétiquement. Il faut donc déterminer les dimensions qui nous donnent le plus d'informations sur nos données. Par exemple, sur la figure 4 ci-dessous, on peut voir à gauche les données complexes à simplifier. Plusieurs axes traversent ces données et on peut voir à droite dans la figure que l'axe représenté par une ligne pleine explique la plus grande variance de nos données. Ensuite, l'axe avec des pointillés larges est le 2ème axe représentant le mieux les données. On peut ainsi sélectionner seulement ces deux axes, appelés composantes principales.

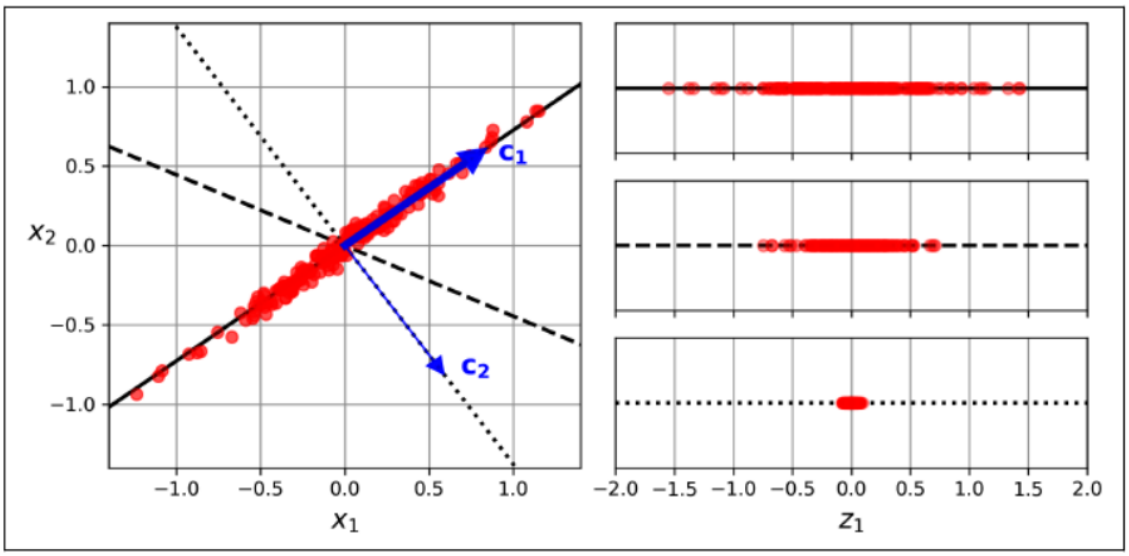

Figure 4 : Exemple de réduction des dimensions.

Tiré (avec autorisation) de Machine learning avec Scikit-learn - Mise en oeuvre et cas concrets, d'Aurélien GÉRON

3.2. Représentation de la diversité allélique

Après avoir sélectionné les composantes principales qui représenteront au mieux les données, on peut appliquer une approximation et projection uniforme de variétés (Uniform manifold approximation and projection – UMAP)[24,25]. Cette projection permet d'avoir une représentation de nos données selon deux dimensions avec un nuage de points. Dans celui-ci, chaque point représente les données génétiques d'un individu. Ainsi, l'UMAP qui fonctionne telle que la PCA, permet d'obtenir une projection de nos données tout en conservant les proximités respectives

de chaque donnée grâce à l'utilisation des composantes principales sélectionnées en amont. Une fois cette étape réalisée, nous pouvons former des groupements. L'intérêt de prétraiter les données avec la PCA avant d'effectuer une UMAP est d'avoir une exploration structurelle de la population plus efficace pour ce type de données. De ce fait, il est possible de mettre l'accent sur la structure des données tout en préservant la structure globale. Ainsi, cela va nous permettre de créer de manière précise des regroupements d'individus.

3.3. Méthode de formation de regroupements

Parmi nos nuages de points obtenus grâce aux méthodes de PCA et UMAP, on cherche à reconstituer des regroupements entre les individus d'une cohorte. Ainsi, on va former des groupes incluant les individus les plus proches génétiquement. Cela permet d'avoir une meilleure représentation des différentes origines de chacun des individus selon leurs données génétiques. Pour faire cela, on utilise la méthode de formation de regroupements appelée k-means clustering[23]. Cette méthode repose sur un algorithme capable de rassembler en regroupements des données. Prenons l'exemple de cette figure 5 qui représente des données tracées selon les axes x_1 et x_2 :

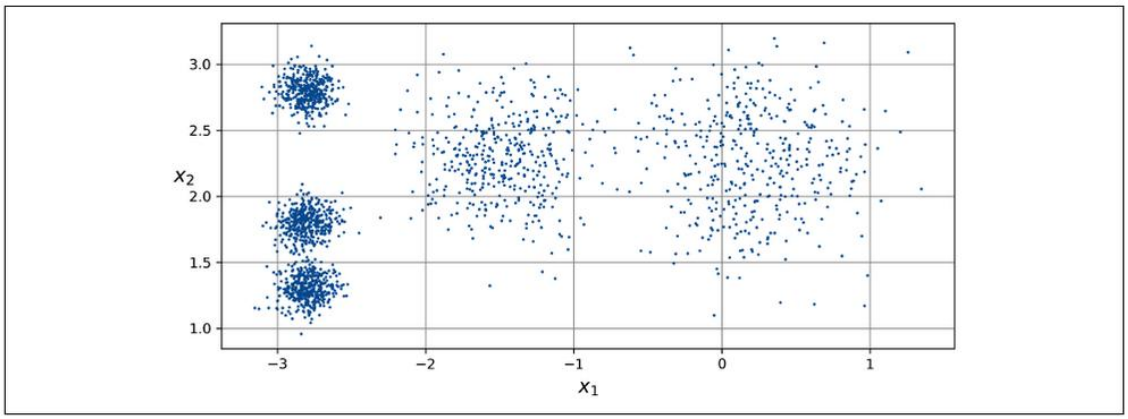

Figure 5 : Nuage de points sans regroupements.

Tiré (avec autorisation) de Machine learning avec Scikit-learn - Mise en oeuvre et cas concrets, d'Aurélien GÉRON

Spontanément, nous pouvons voir qu'il y a dans cette figure cinq regroupements distincts. Cependant, en tracer les limites nous serait impossible sans l'utilisation de l'algorithme. En effet,

grâce à ce dernier, on obtient les limites des différents regroupements ce qui nous permet de bien les délimiter. On a ainsi 5 regroupements pour lesquels les limites sont claires et qui représentent la diversité de nos données. En voici le résultat dans la figure 6 :

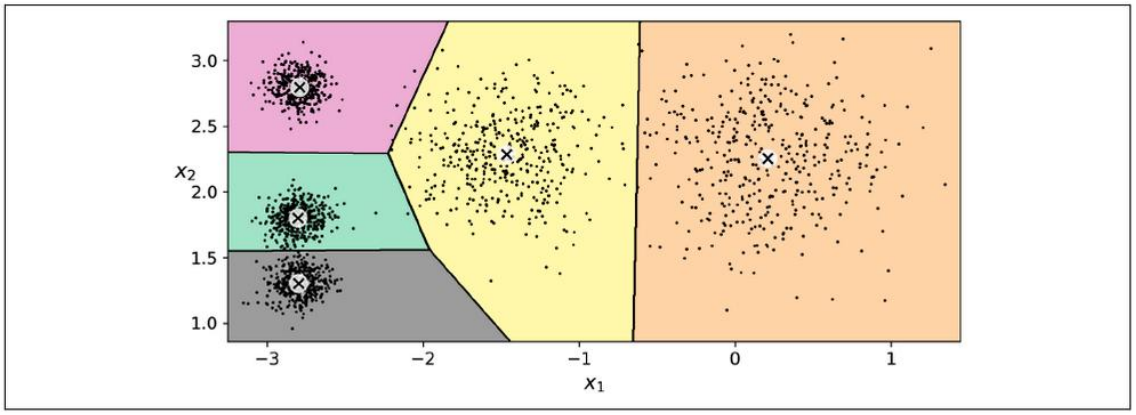

Figure 6 : Représentation des regroupements selon le k-means clustering.

Tiré (avec autorisation) de Machine learning avec Scikit-learn - Mise en oeuvre et cas concrets, d'Aurélien GÉRON

Ainsi, tous les concepts et outils présentés lors de cette première partie vont nous permettre de mieux comprendre les retombées actuelles du peuplement du SLSJ. Nous allons de cette manière pouvoir faire un lien entre le peuplement de cette région et la présence de maladies rares. Afin de comprendre tous les tenants et aboutissants, il faut commencer par l'histoire du peuplement de la province et donc détailler l'arrivée des premiers Français qui vont être à la base du peuplement de cette dernière.

4. Lien entre peuplement de cette population et étude de maladies génétiques

Nous avons le privilège d'avoir accès à des données concernant une population unique dont le peuplement peut être raconté à travers l'étude des caractéristiques génétiques de celle-ci.

4.1. L'établissement de la population au Québec

4.1.1. Le début du peuplement

Pour contextualiser, l'histoire du Québec commence avec une première migration au 17ème siècle, entre 1608 et 1699, d'un groupe d'environ 14 000 migrants français[26]. Or, nombre de ces individus ont pu soit migrer ailleurs, soit revenir en France par la suite ou encore rester

célibataires toute leur vie. Si l'on considère une migration pionnière, c'est-à-dire durant laquelle les individus se sont installés et ont fondé une famille au Québec, on ne compte plus que 5 000 individus[26].

Comme l'indiquent Girard et al., 1995[27], "Les régions portuaires de la côte atlantique, pensons à Saint-Malo, Rouen, Le Havre, Granville, Nantes, Bordeaux ou La Rochelle […] ont permis à la France de concrétiser sa prise de possession en Nouvelle-France". Ce sont notamment les régions de Poitou-Charentes, Normandie ainsi que la région parisienne qui fournissent le plus de pionniers[26].

Tout d'abord, ces derniers se sont installés dans la province de Québec (QcP) et plus précisément dans la vallée du Saint-Laurent au niveau des côtes avec les ports à Québec en 1608, puis à Trois-Rivières (située dans la région de la Mauricie) en 1634 et enfin à Montréal en 1642[28] dont les localisations géographiques sont indiquées sur la figure 7.

Ensuite, à la fin du 17ème siècle, la population au niveau du Saint-Laurent s'étant agrandie, des migrations en direction de l'est, avec le Bas-Saint-Laurent et la Gaspésie, et en direction du Nord avec Charlevoix se font[29] (voir figure 7). Maintenant que l'histoire du commencement de cette province est connue, nous allons nous intéresser plus spécifiquement au lien entre ce peuplement et les retombées en termes de génétique qu'on constate de nos jours. Pour cela, il nous faut l'accès à des données spécifiques et précises sur le déroulé du peuplement dans cette province.

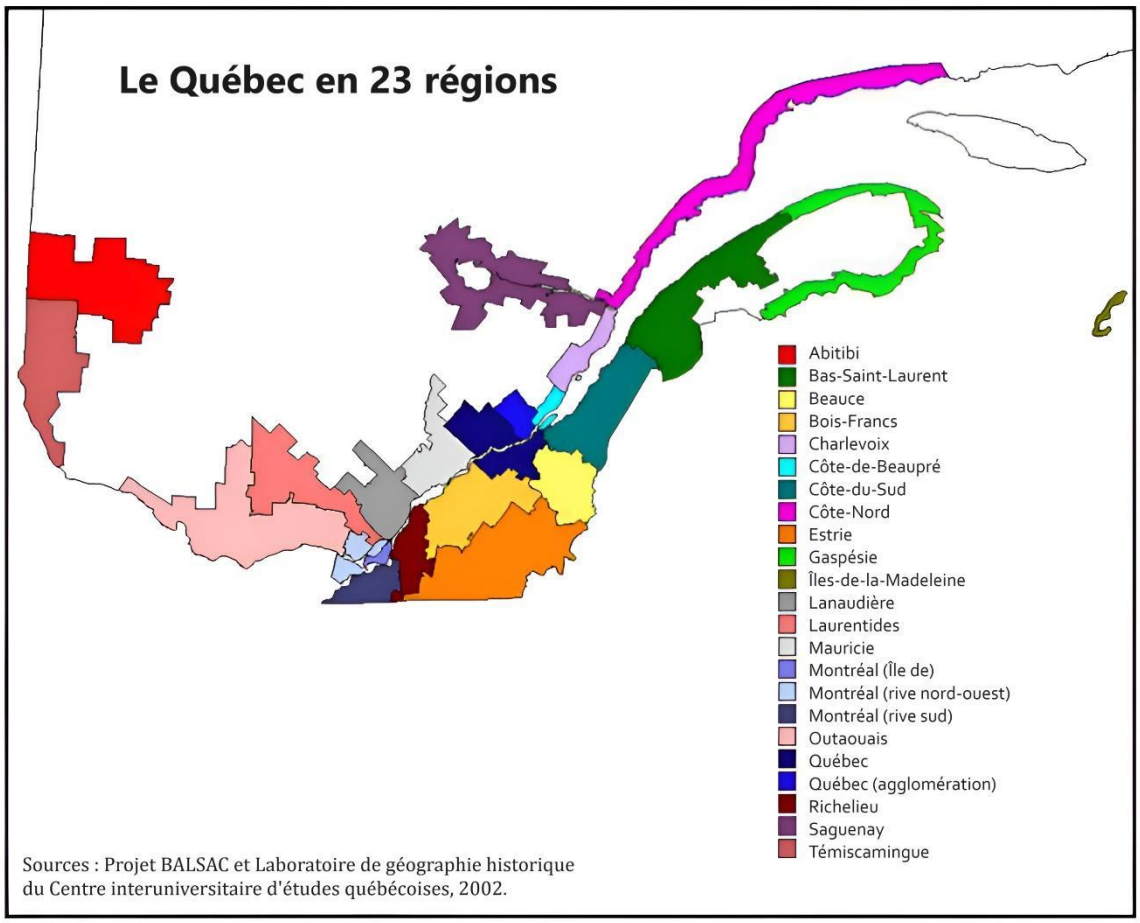

Figure 7 : Carte représentant les régions.

Tiré (avec autorisation) du Projet BALSAC et Laboratoire de géographie historique du Centre interuniversitaire d'études québécoises, 2002.

4.1.2. Mise en évidence d'un effet fondateur au Québec grâce à des données précieuses

Les conditions nécessaires à l'établissement d'un effet fondateur sont réunies pour la population du Québec. En effet, comme cité plus tôt, cette population établie au 17ème siècle se base sur un nombre limité de migrants. De plus, ces derniers ont eu de nombreuses descendances[26]. En effet, les couples avaient de nombreux enfants et cela peut s'expliquer par une très grande fécondité des femmes, des mariages à des âges précoces, mais aussi les conditions propices de l'environnement du Saint-Laurent[30] et la pression de l'Église. Cela a eu pour effet l'établissement de certaines mutations transmises à travers les générations et augmentant en fréquence au cours du temps. En effet, Charbonneau, 2020, affirme que « 1500 hommes et 1100 femmes sont aujourd'hui à l'origine des deux tiers du génome des Canadiens français ». Nous allons ainsi

pouvoir étudier l'histoire de l'établissement de cette population dans la province ainsi que dans la région du Saguenay–Lac-Saint-Jean, située dans la partie nord de la province, en lien avec la génétique de cette dernière. Cela est notamment possible grâce à une base de données précieuse. En effet, le fichier de population BALSAC (https://balsac.uqac.ca/) a pour objectif de recueillir tous les actes civils du Québec créés au cours des siècles. Cela a permis de retracer de nombreuses lignées généalogiques au sein de la population depuis le début du peuplement au 17ème siècle jusqu'à aujourd'hui sur l'ensemble du territoire. Ce fichier de population est unique et existe grâce à la collecte et à l'informatisation des actes civils. Ainsi, il permet d'avoir des informations sur les caractéristiques de la population en termes de migrations et de croissance. Cela a été possible grâce aux calculs par exemple de l'apparentement entre les individus ainsi que de la consanguinité. Mais aussi, la quantification des flux de personnes lors de migrations a permis d'avoir accès à des informations précieuses. En ce qui concerne l'apparentement, celui-ci mesure le degré de parenté entre 2 individus et donc la probabilité qu'ils partagent des allèles identiques par transmission de leurs ancêtres communs documentés. Il peut donc être calculé par l'établissement de la relation entre les individus dans la généalogie. Contrairement à l'apparentement, la consanguinité calcule la probabilité qu'un seul individu possède deux copies identiques provenant du même ancêtre. Ainsi, on s'intéresse dans ce cas plus précisément à la structure génétique d'une seule personne. C'est sur ce fichier que nous allons nous baser pour étudier les spécificités de la région du SLSJ. Et notamment grâce aux liens qui vont pouvoir être faits avec les données génétiques. Par exemple, l'apparentement peut être explicité par l'étude des segments IBD et la consanguinité peut être liée avec le concept de maladie récessive.

4.1.3. Migration vers la région de Charlevoix

Tout d'abord, nous allons nous intéresser plus particulièrement à l'établissement de la population à Charlevoix puisque, comme nous le verrons par la suite, cette dernière sera importante lors du peuplement du SLSJ. Or, on peut déjà constater que cette région est unique par la façon dont elle s'est formée. En effet, 91.3% des individus s'étant installés dans la région

étaient originaires d'une autre région du Québec. De plus, les migrations qui ont permis l'établissement dans cette région sont à très fort caractère familial. Ainsi, ce sont des familles qui se sont établies à Charlevoix et ses terres. Or, on constate dans cette région des coefficients de consanguinité parmi les plus élevés de ceux qui ont pu être mesurés dans des populations humaines[29]. On ne parle pas ici de consanguinité proche puisque ce ne sont pas des unions entre oncle et nièce ou tante et neveu. En effet, l'Église catholique a fait pression pour que ce type d'union ne soit pas possible. Cependant, on parle de consanguinité éloignée avec des mariages entre individus ayant des ancêtres communs. Ce phénomène, couplé avec la forte croissance de la population, va mener à l'établissement de certaines mutations létales dans la population. En effet, la population va passer d'environ 1 000 habitants à la fin du 17ème siècle à plus de 17 000 un siècle plus tard et cela grâce aux conditions propices telles que l'agriculture et l'exploitation forestière. Ainsi, la fréquence de certaines mutations pathogènes va être augmentée par ce fort accroissement de la population. Cela contribue à l'amplification de l'effet fondateur dans la continuité de celui initial québécois. Nous verrons par la suite que cela constitue la fondation de la présence importante de maladies rares au SLSJ.

4.2. Peuplement du SLSJ

4.2.1. Premières migrations en direction du SLSJ

Ainsi, la région du SLSJ étant au 17ème siècle protégée pour la traite de la fourrure, appelée « La traite de Tadoussac »[31], les individus ne pouvaient pas s'y installer. Suite à des pressions de la part des citoyens de Charlevoix afin d'accéder à ces terres plus vastes et avec un potentiel dans l'industrie du bois, cette dernière s'est ouverte au peuplement. On constate dans le premier tiers du 19ème siècle l'établissement de familles, principalement venues de Charlevoix, dans la région[29]. En effet, toutes les paroisses de Charlevoix ont contribué au peuplement du Saguenay et 71% des personnes ayant migré de ces paroisses se sont installées définitivement au Saguenay. De plus, il a été mis en évidence que ces migrations se faisaient principalement par familles et non pas par individus isolés. Ainsi, les familles venant de Charlevoix ont eu de nombreux enfants et ont pu s'installer définitivement sur les terres. C'est donc ces mouvements

migratoires familiaux importants de Charlevoix qui ont permis le transfert des traits spécifiques de Charlevoix dans la région du SLSJ.

En 1852, la région du Saguenay comptabilise 5 200 habitants dont 80% sont issus de l'immigration en provenance d'autres régions du Québec. Ce rythme de migration va continuer et croître jusqu'en 1911. Ainsi, on ne peut pas dire que la région est isolée géographiquement. Cependant, ces migrations importantes ne vont pas apporter autant d'éléments de diversification qu'elles devraient. Et cela s'explique par l'origine des migrants comme nous l'avons vu précédemment. De plus, les autres migrants venant de régions diverses ne se sont pas toujours établis définitivement au SLSJ face au noyau déjà existant des familles venues de Charlevoix qui s'étaient déjà installées de façon pérenne. En effet, ces familles ont pu profiter de leurs avances économiques et sociales afin d'avoir de nombreux descendants et donc contribuer au bassin génétique du SLSJ de manière significative. Ainsi, on peut voir que ce peuplement du SLSJ pose les bases propices pour l'intensification de l'effet fondateur. Nous allons voir ensuite comment le développement de cette région a continué de contribuer à l'amplification de l'effet fondateur[29].

4.2.2. Accroissement de la population

La croissance de la population au SLSJ va être très soutenue. En effet, cette dernière va passer de 5 000 individus en 1852 à 50 000 en 1911. Au début, la forte croissance s'explique par des migrations, comme nous avons pu le voir dans le paragraphe précédent, de familles venant de Charlevoix ainsi que, dans une moindre proportion, des individus isolés. Cependant, ce flux migratoire va rapidement s'inverser au cours du temps. En effet, on constate de l'émigration et notamment de ces individus isolés qui n'ont pas d'attache familiale. Ensuite, on constate une fécondité extrêmement élevée et une mortalité relativement faible au SLSJ jusqu'au 20ème siècle. Cela va contribuer à un accroissement rapide de la population. En effet, la population saguenéenne a été multipliée par 25 entre 1861 et 1961[32]. De ce fait, la relocalisation sur ces terres permet aux familles d'avoir de nombreux enfants qui constitueront la main-d'œuvre

nécessaire à la mise en valeur de ces dernières. Notamment, il y a assez de terres disponibles pour en céder à chaque enfant pour qu'eux-mêmes s'établissent. Ainsi, ce système de reproduction familiale va favoriser la concentration locale de mêmes variants. On parle d'effet multiplicateur[32]. Celui-ci va jouer sur l'intensité de l'effet fondateur et on peut voir de nos jours les conséquences de ce dernier.

4.3. Indicateurs contemporains de l'effet fondateur

On peut donc considérer un effet fondateur régional dans la vallée du Saint-Laurent et ayant une influence dans la région de Charlevoix et celle du SLSJ. On sait que la région de Charlevoix a fourni de nombreuses familles qui ont pris racines au SLSJ et qui ont pu s'établir dès son ouverture, leur donnant un avantage au niveau économique et démographique[29]. Cela a contribué à un changement de fréquence de certains variants au cours du temps et permet d'expliquer le fait que l'on retrouve des maladies rares communes aux deux régions, mais aussi certaines spécifiques à chacune, mettant en évidence la prolongation de l'effet fondateur[29]. Il est possible de mettre en évidence l'effet fondateur grâce aux outils génétiques et généalogiques. En effet, il a été mis en avant la contribution génétique très importante de certains ancêtres éloignés dans la généalogie grâce à l'analyse des ancêtres communs les plus proches (Most recent common ancestor, MRCA)[33]. Sachant que la contribution génétique se définit par l'apport d'allèles d'un individu dans une population par transmission à sa descendance, on peut voir ici que seulement quelques ancêtres ont eu une importante contribution à la composition du pool génétique de la population d'aujourd'hui. De plus, il a été montré que les segments IBD partagés entre les individus du SLSJ sont plus courts que d'autres populations de la province mettant en évidence des différences de structures dues à l'effet fondateur[33]. Mais aussi cela indique et confirme un apport important de segments d'ADN par des ancêtres plus éloignés.

Ensuite, grâce à l'étude des généalogies, il a été montré que la consanguinité proche au SLSJ est comparable à d'autres régions en Europe ou sur le continent nord-américain[29]. Mais aussi, il a

été mis en évidence que la consanguinité proche au SLSJ étaient parmi les plus faibles lorsqu'on la compare à la consanguinité proche d'autres régions du Québec qui ne présentent pourtant pas de maladies génétiques rares[34]. Ainsi, ce n'est pas un facteur à prendre en compte pour expliquer cette fréquence d'allèles plus importante mais c'est bien l'effet fondateur qui en est la cause. En effet, il a été mis en évidence qu'un seul individu fondateur pouvait être à l'origine de l'augmentation de fréquence d'un variant donné[35,36]. De plus, on peut voir que la dérive au SLSJ a été augmentée. Cela signifie que les individus sur le front de la vague ont eu une contribution génétique au pool du SLSJ bien plus importante que ceux en retrait de la vague. Cela s'explique notamment par une plus grande fécondité des femmes présentes sur le front de la vague[37]. On parle ainsi de dérive augmentée qui a pu aussi jouer sur l'intensité de l'effet fondateur. Enfin, il a été montré que le changement de fréquence chez les Canadiens-Français suite à l'effet fondateur avait touché des variants plus délétères que d'autres[38]. Cela pourrait s'expliquer par l'expansion significative de la population qui aurait pu mener à une sélection moins efficace. Tous ces éléments constituent les bases afin de mieux comprendre la présence plus importante de maladies rares dans cette province et en particulier au SLSJ.

4.4. Ce que l'on sait des maladies génétiques rares au SLSJ

De nombreux articles scientifiques[39–43] ont référencé diverses maladies génétiques rares dans le monde comme plus fréquentes dans la région du SLSJ. Ces publications ont été possible grâce au travail de cliniciens qui ont pu diagnostiquer des individus présentant des symptômes et atteints par ces maladies. On a notamment l'exemple des quatre maladies des tests de porteurs offerts à la population qui sont indiquées dans le tableau 1 suivant avec les taux de porteurs associés[42] :

Tableau 1 : Les quatre maladies du test de porteurs.

Maladie	Taux de porteurs associé
Tyrosinémie de type 1	1/20
Neuropathie sensitivomotrice héréditaire avec ou sans agénésie du corps calleux	1/23
Ataxie récessive spastique de Charlevoix-Saguenay	1/18
Acidose lactique congénitale	1/26

Un taux de porteurs de 1/18 signifie que sur 18 personnes, une sera porteuse d'une copie de la mutation responsable de ces maladies. Il est important de noter que ces quatre maladies sont récessives.

Chapitre 1 - Données, méthodes et objectifs

Dans ce projet, nous essayons d'apporter une nouvelle approche dans le domaine médical afin de déterminer autrement la présence de maladies génétiques fréquentes dans la population. Ainsi, nous pourrons comparer ce que nous trouvons lors de nos analyses avec ce qui est référencé dans la littérature. Cela permettra de confirmer nos résultats si nous trouvons les maladies déjà décrites, mais aussi de trouver peut-être de nouvelles maladies grâce à cet angle d'analyse nouveau.

1. Outils d'analyses

L'utilisation d'outils développés en génétique permet, dans mon mémoire, d'avoir une nouvelle vision sur les données à notre disposition et d'en tirer des conclusions sur l'effet du peuplement sur les génomes contemporains. Ces analyses se basent sur des individus non atteints de pathologies, ce qui nous démarque de la démarche traditionnelle. En effet, au lieu d'identifier des maladies liées à l'effet fondateur ainsi que les mutations associées grâce à l'examen de patients atteints de ces dernières qui viennent consulter, notre projet se base sur la population globale sans se limiter aux individus atteints. De ce fait, nous allons nous baser sur une cohorte populationnelle pour ce projet. Ainsi dans cette partie, les types de données utilisées seront introduits, ainsi que la cohorte sur laquelle je me base pour ce projet. Finalement, un résumé des analyses sera explicité afin d'en comprendre le déroulement.

1.1. Types de données utilisées

Afin de déterminer la diversité allélique d'une population donnée, il nous faut avoir recours à des outils de génotypage. Ces derniers permettent de déterminer la présence de polymorphisme nucléotidique simple (SNP) dans la population. On parle de SNP lorsqu'on constate différentes copies chez différents individus à un même locus. Lors de ce projet, nous avons utilisé deux techniques permettant d'obtenir les données souhaitées : le séquençage entier du génome (Whole Genome Sequencing - WGS) et les données imputées par l'utilisation de génotypages.

La technique de WGS permet, comme l'indique son nom, de séquencer chaque nucléotide présent et cela à travers tout le génome[44]. Cette technique est donc très coûteuse malgré les avancées technologiques permettant de réduire le prix[45]. Mais aussi, elle demande un certain temps de traitement des données. Cependant, cette technique est plus précise puisqu'elle permet de séquencer aussi bien les régions codantes que les régions non codantes. Ainsi, il est plus facile dans une population donnée de mettre en évidence des variants rares spécifiques à celle-ci.

Ensuite, l'utilisation de données imputées repose sur la combinaison de génomes séquencés entièrement dans le but de les utiliser comme références[46] et de génotypes. Les génotypes acquis lors du génotypage correspondent à des régions précises qui vont être déterminées par séquençage à travers tout le génome. Une fois les génotypes constitués, il est possible de comparer ceux-ci avec les génomes de référence qui ont été séquencés entièrement. Lorsqu'il y a une correspondance de nucléotides entre les deux, il est possible de former les données imputées par combinaison des génotypes et des WGS[47]. Par exemple, comme on peut le voir sur la figure 8 ci-dessous, nous avons à notre disposition une séquence d'ADN dont certaines bases ont été déterminées, on parle ici d'un génotype. Le but est de combler les nucléotides manquants à l'aide des génomes de référence séquencés et numérotés de 1 à 3. Or, on peut voir que seul le génome de référence 2 correspond entièrement aux bases du génotype. Ainsi, il est possible de compléter les nucléotides manquants afin d'obtenir nos données imputées.

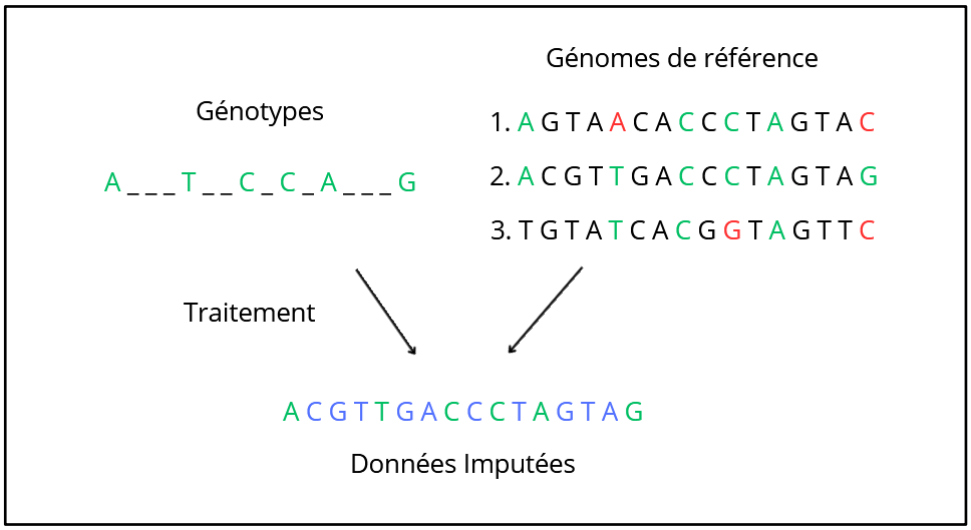

Figure 8 : Principe des données imputées.

Si l'on compare ces deux techniques qui permettent d'analyser l'ADN, on constate que la méthode de séquençage entier du génome demande plus de temps et est plus onéreuse[45]. Ensuite, ces techniques sont très coûteuses en espace de stockage sachant que le génome entier d'un humain est de 3,2 Giga paires de bases en moyenne entre les hommes et les femmes[48]. De plus, lorsque l'on travaille avec des données imputées, il y a un risque de passer à côté de mutations rares[49]. En effet, comme expliqué dans le paragraphe précédent, les données imputées se basent sur la correspondance des génotypes et les WGS de référence. Or, ces derniers proviennent de panels mondiaux et non de populations spécifiques. Ainsi, des mutations spécifiques à une population précise mais rares dans le reste du monde seront moins bien déterminées lors de la correspondance entre les génotypes d'une population précise et des WGS issus de panels mondiaux. De ce fait, la détermination de mutations rares est moins probable lors d'utilisation de WGS de référence pour une population spécifique. Cependant, l'avantage des données imputées est qu'on peut avoir une plus grande cohorte pour les raisons évoquées précédemment. Contrairement aux données imputées, les données de séquençage entier du génome, puisque la technique consiste à déterminer chaque SNP de tout le génome, permettent d'identifier plus systématiquement les mutations rares dans une population. Cependant, dû au coût et au temps que les analyses prennent, les cohortes sont généralement de taille inférieure. Ainsi, la méthode d'imputation est très utilisée pour de très grandes cohortes

pour lesquelles on veut étudier des mutations considérées communes voire peu communes. Alors que les données séquencées sont analysées dans le contexte d'une plus petite cohorte pour des mutations rares.

Ainsi, lors de mon projet, j'ai eu la chance d'avoir accès aux données de séquençage entier du génome de 2,184 personnes ainsi qu'aux génotypages de 29,337 individus de la population d'intérêt et ces dernières nous ont permis de générer des données imputées plus précises en ce qui concerne des mutations spécifiques à cette population. Sachant que notre intérêt se porte sur les mutations rares, nous avons commencé par étudier les données de séquençage. Or, voulant augmenter le nombre d'individus de notre cohorte, nous avons décidé de travailler avec les imputations. Ainsi, nous avons utilisé les génomes séquencés de notre cohorte ainsi que les données de génotypages afin de générer les données imputées. Cela permet de pallier au problème de non-identification de mutations rares. En effet, si nous utilisons des génomes de références basés sur la population mondiale, les mutations plus fréquentes dans la population du SLSJ qui sont très rares dans les autres populations auraient pu être manquées. Pour résumer, nous utilisons lors de nos analyses les données imputées grâce à l'utilisation de données de séquençage entier de la population afin de déterminer les mutations rares dans la population en ayant tout de même un nombre d'individus élevé. Ainsi, cette méthode d'analyse est possible grâce à la cohorte populationnelle à laquelle nous avons accès nommée CARTaGENE.

1.2. Banque de données

La plateforme CARTaGENE a été créée afin de soutenir la recherche en santé au Québec[50]. Ainsi, grâce à cette base de données, nous avons accès aux données de séquençage entier du génome de 2 184 individus et aux génotypes de 29 337 individus[51]. Le recrutement de ces derniers s'est fait dans 6 villes réparties dans la province de Québec : Montréal, Sherbrooke, Québec, Saguenay, Gatineau et Trois-Rivières. La collecte de données s'est déroulée entre 2009 et 2015. Or, sachant que des individus peuvent migrer tout au long de leur vie, une personne habitant Montréal peut venir d'une autre région comme le SLSJ. Or, pour notre projet, nous avons besoin

de regroupement d'individus selon leur proximité génétique et non pas selon leur lieu de vie. Nous avons donc utilisé les outils présentés précédemment comme la PCA et l'UMAP en finissant avec la méthode de formation de regroupements k-means afin de traiter les données. Cela nous a permis de former des groupes d'individus selon leur proximité génétique et donc de leur origine respective afin de mieux identifier des mutations spécifiques à une dite région. Cette banque de données traitée par les outils génétiques à notre disposition et combinée avec des méthodes d'analyses nous a permis de mettre en évidence des variants dont les fréquences ont été modifiées au cours du peuplement et cela en lien avec l'effet fondateur.

1.3. Identification des variants fondateurs

Ici, nous nous intéressons aux variants qui ont été introduits par peu, voire un unique individu, et qui ont subi un changement de fréquence considérable au cours des générations. C'est l'effet fondateur dont nous avons abordé la théorie dans les paragraphes précédents. Ce dernier se traduit par une augmentation considérable de la fréquence d'un variant considéré rare dans la population de base. Néanmoins, il existe une autre hypothèse afin d'expliquer une fréquence importante d'un variant. En effet, un variant peut être plus fréquent car celui-ci a été introduit par de nombreux individus au cours du temps. Cependant, nous nous intéressons dans cette étude aux variants dits fondateurs car nous pouvons les caractériser ainsi selon nos méthodes d'analyse et les mettre en lien avec le peuplement de la région. On considère donc un variant comme fondateur selon deux critères. Premièrement, ce dernier doit être retrouvé chez plusieurs individus dans la population contemporaine. Mais aussi, un nombre limité d'ancêtres communs entre les individus doit être à l'origine de la transmission de ce variant chez ces derniers. Ainsi, ici on ne se base pas seulement sur la fréquence du variant dans la population comme cela a été fait auparavant. En effet, nous avons vu qu'il pouvait y avoir plusieurs raisons expliquant la fréquence élevée d'un variant dans la population. Dans notre cas, nous considérons les variants fréquents dans la population mais aussi dont l'origine peut être établie à peu d'individus, voire un seul. Pour cela, on postule que le variant pathogène en question, ainsi

que le segment d'ADN entourant le variant, vont être transmis de générations en générations. C'est pourquoi nous allons essayer de retrouver ce segment transmis au cours du temps grâce à nos outils d'analyse.

Ainsi, grâce à une méthode déjà décrite et ayant permis de décrire le lien entre les segments IBD et la structure de la population[33,52,53], on peut générer ces segments IBD partagés entre des individus donnés et procéder à leur analyse. Pour cela, on regarde s'il y a une forte proportion de partage de segments IBD au niveau du variant. Si c'est le cas, cela traduit une origine commune ultérieure au peuplement entre les porteurs du variant. Cependant, on regarde aussi s'il y a de fortes proportions de partage dans d'autres parties du génome et sur le génome entier. En effet, cela traduirait une parenté proche entre les individus et non l'influence de l'effet fondateur.

Ainsi, grâce à ces données et les outils d'analyses présentés précédemment, nous avons pu étudier l'influence de l'effet fondateur sur la population du SLSJ. Pour ce faire, nous avons suivi des étapes précises.

1.4. Flux d'analyses

Afin d'étudier la présence de maladies rares au SLSJ, différentes étapes ont été suivies lors de ce projet et vont être présentées dans ce paragraphe. Tout d'abord, la première étape de mon projet est d'avoir une représentation qualitative de nos données génétiques afin d'en déterminer les différentes structures. Notamment, le but est de former des regroupements d'individus selon leur proximité génétique afin d'en faire ressortir les spécificités grâce à la PCA et l'UMAP. Une fois nos populations bien définies grâce à la méthode des k-means clustering, on cherche à mettre en avant la présence de variants plus fréquents dans celles-ci en les comparant à une référence mondiale. Pour ce faire, on détermine la fréquence d'apparition de chacun des variants présents dans nos données et on sélectionne seulement ceux qui sont plus fréquents au Québec. Ensuite, pour chacun de ces variants, on cherche à déterminer si sa

présence est le résultat de l'effet fondateur. Pour cela, on utilise les segments IBD qui nous permettent d'identifier l'origine ancestrale commune du variant présent chez les individus. Lorsque cela est fait, le but est d'associer chacun de ces variants à la maladie afin de comparer nos résultats aux informations présentes dans la littérature mais aussi dans les cliniques. Cela nous mène à la partie objectifs de mon projet.

2. Objectifs

Le travail des cliniciens au cours des années a permis la mise en évidence de la présence de maladies rares liées à l'effet fondateur au SLSJ[39–43]. Cela a été possible grâce à l'étude des symptômes de personnes atteintes qui ont pu être diagnostiquées. Cependant, connaissant la complexité des maladies rares, la compréhension des symptômes en lien avec une maladie n'est pas toujours facile et un individu peut attendre de nombreuses années avant d'avoir un diagnostic final clair. Ainsi, l'objectif de ce projet porte sur l'étude des génomes et la détermination de variants potentiellement pathogènes présents au SLSJ.

En d'autres mots, l'objectif principal de mon projet est de démontrer la présence de variants rares dans la population du SLSJ et que cela est lié à son peuplement. Cela nous permettra de mettre en avant la présence de variants déjà répertoriés et considérés comme fondateurs dans nos analyses. De plus, on cherche à déterminer s'il est possible de trouver des variants qui n'ont jamais été référencés auparavant au SLSJ qui sont pourtant fondateurs et ainsi plus fréquents dans la population et associés à cet effet. Ainsi, les connaissances acquises lors de cette étude permettront d'accélérer le diagnostic des patients atteints de maladies rares car il y aura plus d'informations sur ces dernières. La finalité de ce projet est de travailler conjointement avec les services de génétique de la région du SLSJ afin de tendre vers la mise à disposition de tests de porteurs comme il en existe déjà pour la population. Ces tests de porteurs seront ainsi basés sur nos analyses qui nous permettent de déterminer avec précision les taux de porteurs dans la population et ainsi quelles maladies mettre en avant pour ces derniers. Cela assure la meilleure prise en charge possible avec les outils que nous avons aujourd'hui.

Chapitre 2 : Rare diseases load through the study of a regional population

Avant-propos

Cet article a été rédigé lors de ma maîtrise avec le soutien de mon directeur de recherche et de son laboratoire Genopop. Celui-ci répond aux objectifs du mémoire énoncés précédemment dans le Chapitre 1. En effet, dans cet article, nous avons mis en application toutes les notions abordées en introduction et toutes les méthodes décrites afin de déterminer et détailler la présence de maladies génétiques rares dans la région du Saguenay–Lac-Saint-Jean.

Cet article intitulé *Rare diseases load through the study of a regional population* a deux aspects principaux. Le premier est analytique avec l'étude des génomes pour la détermination de variants fondateurs. Le deuxième est clinique avec la détermination de certains de ces variants chez des patients. C'est ainsi que les cliniciens ont contribué à l'article. En ce qui concerne ma contribution lors de la rédaction de cet article, j'ai travaillé sur tous les aspects analytiques de ce projet avec l'identification des variants fondateurs et les représentations graphiques avec les figures ainsi que les tableaux de cette partie.

Au moment de la rédaction de ce mémoire, cet article est publié sur un serveur de prépublication nommé MedRxiv le 30 octobre 2024 et avec le DOI suivant : https://doi.org/10.1101/2024.10.29.24316346. De plus, il a été soumis au journal PLOS Genetics en date du 06 mars 2025. Une révision est actuellement en préparation.

Rare diseases load through the study of a regional population

Élisa Michel 1,2, Claudia Moreau 1,2, Laurence Gagnon 1,2, Josianne Leblanc 3, Jessica Tardif 3, Lysanne Girard 4, Jean Mathieu 5,6, Cynthia Gagnon 5,6, Mathieu Desmeules 5,6,8, Jean-Denis Brisson 5,6,9, Luigi Bouchard 3,4, Simon L. Girard 1,2,10,11

1. Département des sciences fondamentales, Université du Québec à Chicoutimi, Saguenay, Québec, Canada.
2. Centre Intersectoriel en Santé Durable (CISD), Université du Québec à Chicoutimi, Saguenay, Québec, Canada.
3. Département clinique de médecine de laboratoire du Centre intégré universitaire de santé et services sociaux (CIUSSS) du Saguenay–Lac-St-Jean, Saguenay, Québec, Canada.
4. Département de biochimie et de génomique fonctionnelle, Faculté de médecine et des sciences de la santé, Université de Sherbrooke, Saguenay, Québec, Canada.
5. Groupe de recherche interdisciplinaire sur les maladies neuromusculaires (GRIMN), CIUSSS du Saguenay–Lac-Saint-Jean, Saguenay, Québec, Canada.
6. Faculté de médecine et des sciences de la santé, Université de Sherbrooke, Saguenay, Québec, Canada.
7. Centre de recherche et d'innovation du CIUSSS du Saguenay–Lac-St-Jean, Saguenay, Québec, Canada.
8. Clinique de pédiatrie du Saguenay, Saguenay, Québec, Canada.
9. Clinique des maladies neuromusculaires (CMNM), CIUSSS du Saguenay–Lac-St-Jean, Saguenay, Québec, Canada.
10. Projet BALSAC, Université du Québec à Chicoutimi, Saguenay, Québec, Canada.

11. Centre de recherche CERVO, Université Laval, Québec, Québec, Canada.

✉email: simon2_girard@uqac.ca

Résumé

Cet article porte sur l'étude des maladies génétiques rares présentes dans la province de Québec et a pour but d'établir un lien concret entre l'effet fondateur caractérisant le peuplement de celle-ci avec la fréquence élevée de certains variants pathogènes. Pour cela, nous nous basons sur les données de CARTaGENE, une plateforme de recherche publique qui regroupent les données de séquençage entier du génome de 2 184 individus et les données imputées de 29 337 individus, tous provenant de la province. Après la détermination de la fréquence des variants pathogènes et l'étude des segments identiques par descendance, il a été mis en avant 80 variants fondateurs dont 38 étaient déjà répertoriés. Parmi les nouveaux variants déterminés lors de cette étude, 2 répondent aux critères pour être inclus dans les tests de porteurs déjà existants et mis à disposition de la population. De plus, cette étude permet de mettre en avant la proportion plus importante de variants présents dans la population du Saguenay–Lac-Saint-Jean par rapport au reste du Québec. Cela peut être justifié par un effet fondateur plus marquant au SLSJ que dans le reste de la province, et qui s'explique par son expansion très rapide et le goulot d'étranglement lors du peuplement de la région. Pour conclure, le travail présenté ici a mis en évidence des variants qui n'avaient jamais été répertoriés avant et qui ont une fréquence non négligeable dans la population. Ainsi, la détermination de ces variants tend à aider les cliniciens pour le diagnostic des patients et donc améliorer la qualité de vie de ces derniers avec l'apport de traitement approprié suite à un diagnostic plus rapide et précis.

Abstract

Rare genetic diseases impact many people worldwide and are challenging to diagnose. In this study, we introduce a novel regional population cohort approach to identify pathogenic variants

that occur more frequently within specific populations and are of clinical interest. We utilized a cohort from Quebec, including the Saguenay–Lac-Saint-Jean region, which is known for its founder effect and higher frequency of certain pathogenic variants. By analyzing both the frequency of these variants and their origin through shared identical-by-descent segments, we validated 38 variants previously reported as being more common due to the founder effect. Additionally, we identified 42 unreported founder variants in Quebec or the Saguenay–Lac-Saint-Jean, some with carrier rates estimates as high as 1/22. We also observed a greater deleterious mutational load for the studied variants in individuals from the Saguenay–Lac-Saint-Jean compared to other urban Quebec regions. These findings were brought to the clinic where 12 pathogenic variants were detected in patients, including 3 that are responsible for very severe diseases and could be considered for inclusion in a carrier test for the Saguenay–Lac-Saint-Jean population. This study highlights the potential underestimation of rare disease prevalence and presents a population-based approach that could aid clinicians in their diagnostic efforts and patients' management.

Abbreviations: SLSJ (Saguenay–Lac-Saint-Jean), UQc (Urban Quebec regions), QcP (Quebec province), WGS (Whole-genome sequencing), IBD (Identical-by-descent), CR (Carrier rate), MAF (Minor allele frequency), RFD≥10% (Relative frequency difference of at least 10%), LD (Linkage disequilibrium), CaG (CARTaGENE cohort)

Introduction

Rare diseases are thought to collectively affect as much as 10% of the population[1]. There are more than 10,000 rare diseases described in Orphanet[2] and most of them are of genetic origin. Diagnosis remains a significant challenge for patients living with a rare disease. Despite the growing accessibility of genome sequencing technologies in precision medicine efforts for rare diseases diagnosis[3], these patients often experience prolonged diagnostic odyssey due to insufficient knowledge about their specific condition and the diversity of symptoms observed

for a given disease. It becomes increasingly important to improve the diagnostic yield of rare diseases and to shorten the diagnostic odyssey of patients[4]. Understanding population health disparities is an essential component of equitable precision health efforts.

In certain populations, the prevalence of some rare diseases may increase due to demographic events such as founder effects. It is the case in Quebec, a province in eastern Canada, predominantly settled by people of French origin starting in the early 1600s[5]. The initial European founder effect was followed by subsequent regional founder effects, notably the well-characterized one observed in Charlevoix and Saguenay–Lac-Saint-Jean (SLSJ) regions[6]. Consequently, many rare diseases are more frequent in SLSJ than elsewhere in the world[7–10]. In SLSJ, most people are aware of the higher risk of transmission of some rare diseases and a carrier test is offered to the populations of Charlevoix, SLSJ and Haute-Côte-Nord for 4 of these diseases[8,11]. Nevertheless, numerous rare diseases still lack a known genetic etiology and diverse manifestations of diseases across patients further complicate clinical diagnosis. Traditionally, founder effects have been analyzed using a bottom-up approach, starting with the phenotypes of patients observed in clinical settings and linking them to genes that are specific to each individual. Often, medical geneticists and the healthcare system would gain valuable insights from obtaining a comprehensive overview of variants that are more frequent in the population and potentially associated with rare diseases. This study focuses on addressing this need.

More specifically, we aimed to describe potentially pathogenic variants that have an increased frequency in SLSJ due either to the founder effect or simply due to many introductions in the population. We conducted a comprehensive screening to identify pathogenic variants with higher frequency in SLSJ. Since the SLSJ population has been extensively studied over the past 40 years, we expected to identify many previously reported variants, thereby validating our findings. In fact, we successfully replicated and confirmed the majority of known founder variants in the SLSJ population and systematically documented their carrier rates. However, we

also identified several variants that may be causal of rare diseases and were not previously documented in SLSJ. As the SLSJ healthcare system features a single entry point for all residents, it simplifies the process of locating patients with newly identified pathogenic variants. A thorough investigation of these newly discovered variants revealed clear diagnoses in the phenotypes of several patients.

Furthermore we report for the first time the global load of rare variants in a single population and assess how the founder effect was pivotal in increasing that load. In the context of rare diseases, a large number of populations remain poorly characterized and we believe that our study highlights the need for regional genetic programs to better understand and diagnose the variety of rare diseases affecting one population.

Results

We detected 1,302 potentially pathogenic rare variants (Supplementary Table 1, see methods) in the whole genome sequencing (WGS) of 1,852 individuals within the Quebec province (QcP) that reached at least 10% relative frequency difference (RFD≥10%, see methods) compared to gnomAD non-Finnish Europeans. To improve carrier rate estimates, we imputed 29,353 individuals from CARTaGENE[12] using the WGS as reference (see methods) and looked at these 1,302 variants. Whenever a variant was present in the imputed data, we used the imputed variant frequency, otherwise, we relied on the WGS variant frequency. The clustering on a Uniform Manifold Approximation and Projection (UMAP) performed on the imputed data identified 3,589 and 21,472 individuals who were genetically related to the SLSJ and the urban Quebec areas respectively (UQc, see methods). Noticeably, 540 (42%) variants with an RFD≥10% are absent from the SLSJ region and 17 (1.3%) only are absent from the UQc, although many variants are more frequent in the SLSJ region (Fig.1). Accordingly, we observed a lower proportion of individuals from the SLSJ that do not carry any potentially pathogenic variant (chi^2 p-value < $2e^{-16}$) while a higher proportion carry 2 or more (chi^2 p-value < $2e^{-16}$) (Fig.2).

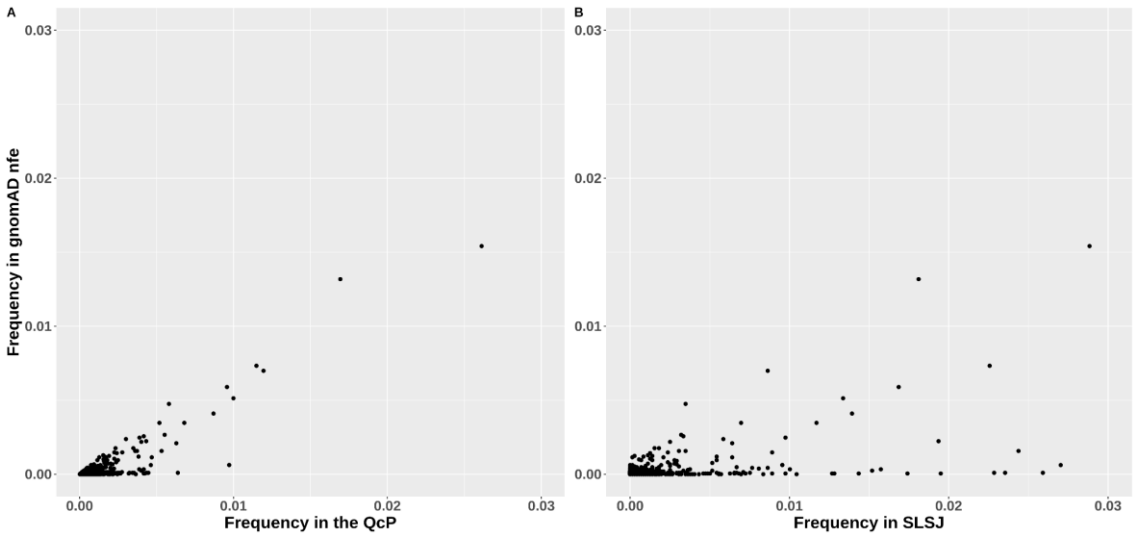

Fig.1: Frequencies of rare variants with RFD≥10% compared to gnomAD in A) QcP and B) SLSJ.

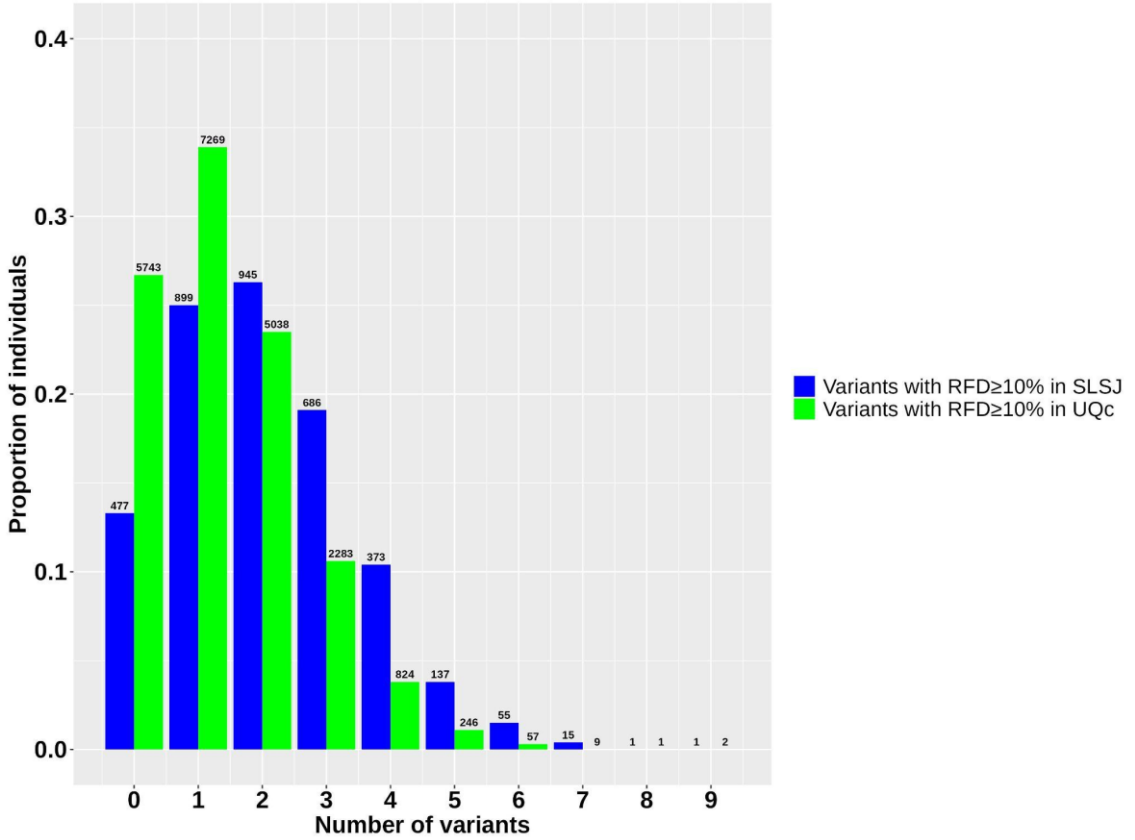

Fig.2: Proportion of individuals carrying variants with RFD≥10%.

Previously reported and newly discovered variants

Previous literature reviews focussed on the Charlevoix-SLSJ founder effect identified 72 variants[7–10]. Among them, 42 were present in our data (Table 1 and Supplementary Table 1). Supplementary Table 2 provides details on the 30 previously reported variants that were either absent in our data or had an RFD<10%, as well as information on some variants that were not considered in our analysis. Noticeably, there is a great correlation between the carrier rates (CR) previously reported and the ones calculated herein (Supplementary Fig.1). Moreover, some carrier rates were reassessed in the CIUSSS laboratory using a subset of 1,000 randomly selected samples with the appropriate consent, and the newly calculated rates fall within the range of those reported in this study (Supplementary Table 3).

In the present study, to be classified as founder, the variant must be present with a carrier rate of at least 1/200 and the proportion of pairs of carriers sharing segments identical-by-descent (IBD) around the variant should be at least 0.5 (see methods). Among the 1,302 rare variants with RFD≥10%, 80 variants met these criteria and are considered as founders either in the QcP, UQc or SLSJ. We reviewed the literature on these 80 variants, examining not only the reviews focusing on the Charlevoix-SLSJ founder effect[7–10], but also case reports within the QcP population. While these reports did not primarily emphasize the founder effect, we still classified the variants therein as previously reported[13–27]. 38 of the founder variants were already documented whereas 42 were not reported in the Quebec population (Table 2 and Supplementary Table 1).

Table 1: Variants previously reported.

				QcP		UQc		SLSJ			
Gene	**Nucleotide**	**Disease name (ClinVar ID)**	**Data type**	**Count**	**CR**	**Count**	**CR**	**Count**	**CR**	**Status**	**Reported CR in SLSJ**
FAH	c.1062+5G>A	Tyrosinemia type I (11870)	Imputed data	232	1/108	38	1/565	194	1/18	F	1/20[7]
SACS	c.8844del	Charlevoix-Saguenay spastic ataxia (5512)	Imputed data	319	1/79	133	1/163	185	1/20	F	1/22[7]

SLC12A6	c.2436+1del	Agenesis of the corpus callosum with peripheral neuropathy (436730)	Imputed data	218	1/115	54	1/398	164	1/22	F	1/23[7]
PDZD7	c.2672AGA[1]	Hearing loss, autosomal recessive 57 (44131)	Imputed data	210	1/120	72	1/298	138	1/26	F	1/32[8]
CYP27B1	c.262del	Vitamin D-dependent rickets, type 1A (1664)	Imputed data	209	1/120	84	1/256	125	1/29	F	1/29[7]
INVS	c.1078+1G>A	Infantile nephronophthisis (660098)	Imputed data	130	1/193	27	1/795	103	1/35	F	1/33[8]
LRPPRC	c.1061C>T	Congenital lactic acidosis, Saguenay-Lac-Saint-Jean type (3110)	Imputed data	122	1/205	30	1/716	92	1/39	F	1/26[7]
AIRE	c.1616C>T	Polyglandular autoimmune syndrome, type 1 (68218)	Imputed data	133	1/188	42	1/511	91	1/39	F	1/39[8]
LPL	c.701C>T	Hyperlipoproteinemia, type I (1527)	Imputed data	103	1/243	33	1/651	70	1/51	F	1/40[7]
TTC7A	c.1001+3_1001+6del	Gastrointestinal defects and immunodeficiency syndrome 1 (50608)	Imputed data	131	1/191	67	1/320	64	1/56	F	1/49[8]
GNPTAB	c.3503_3504del	Mucolipidosis type II (2771)	Imputed data	100	1/251	38	1/565	62	1/58	F	1/39[7]
TTI2	c.950A>T	Severe intellectual disability-short stature-behavioral abnormalities-facial dysmorphism syndrome (1691272)	Imputed data	106	1/236	46	1/467	60	1/60	F	1/45[8]
CTNS	c.414G>A	Cystinosis (4443)	WGS	6	1/309	1	1/1,538	5	1/63	F	1/39[7]
HJV	c.959G>T	Hemochromatosis type 2A (2365)	Imputed data	78	1/321	23	1/934	55	1/65	F	1/70[8]
JUP	c.902A>G	Naxos disease (222662)	Imputed data	81	1/309	29	1/740	52	1/69	F	1/52[8]
CFTR	c.489+1G>T	Cystic fibrosis (38799)	Imputed data	96	1/261	45	1/477	51	1/70	F	1/15[7]
MPI	c.884G>A	MPI-congenital disorder of glycosylation (14349)	Imputed data	60	1/418	10	1/2,147	50	1/72	F	1/71[8]
MAN1B1	c.1075G>T	Rafiq syndrome (1691271)	Imputed data	74	1/339	25	1/859	49	1/73	F	1/62[8]
PEX6	c.802_815del	Peroxisome biogenesis disorder 4A (Zellweger) (555443)	Imputed data	70	1/358	25	1/859	45	1/80	F	1/55[8]
PLPBP	c.370_373del	Epilepsy, early-onset, vitamin B6-dependent (503895)	Imputed data	67	1/374	26	1/826	41	1/88	F	1/71[8]
LAMA3	c.8941C>T	Junctional epidermolysis bullosa gravis of Herlitz (449049)	Imputed data	56	1/448	16	1/1,342	40	1/90	F	1/71[8]

LDLR	c.259T>G	Hypercholesterolemia, familial, 1 (3685)	Imputed data	24	1/1,044	5	1/4,294	19	1/189	F	1/120[7]
SLC25A15	c.553TTC[3]	Hyperornithinemia-hyperammonemia-homocitrullinuria syndrome (5992)	Imputed data	177	1/142	158	1/136	19	1/189	F	NA[9]
CFTR	c.1521_1523del	Cystic fibrosis (7105)	Imputed data	848	1/30	718	1/30	130	1/28	NF	1/15[7]
CFTR	c.1364C>A	Cystic fibrosis (7111)	Imputed data	26	1/964	12	1/1,789	14	1/256	NF	1/15[7]
CFTR	c.617T>G	Cystic fibrosis (7190)	Imputed data	138	1/182	124	1/173	14	1/256	NF	1/15[9]
CFTR	c.579+1G>T	Cystic fibrosis (38494)	Imputed data	44	1/570	37	1/580	7	1/513	NF	NA[10]
FAH	c.47A>T	Tyrosinemia type I (11865)	Imputed data	6	1/4,177	6	1/3,579	0	NA	NF	NA[10]
FAH	c.1090G>T	Tyrosinemia type I (11867)	Imputed data	13	1/1,928	13	1/1,652	0	NA	NF	NA[10]
LPL	c.644G>A	Hyperlipoproteinemia, type I (1522)	Imputed data	46	1/545	38	1/565	8	1/449	NF	1/40[7]
LPL	c.829G>A	Hyperlipoproteinemia, type I (1539)	Imputed data	14	1/1,790	12	1/1,789	2	1/1,794	NF	1/40[7]
CTNS	c.473T>C	Cystinosis (21439)	Imputed data	13	1/1,928	6	1/3,579	7	1/513	NF	1/39[9]
WNK1	c.3301C>T	Neuropathy, hereditary sensory and autonomic, type 2A (5166)	WGS	4	1/463	3	1/513	1	1/314	NF	NA[9]
CAPN15	c.1838C>T	Oculogastrointestinal-neurodevelopmental syndrome (1074293)	Imputed data	50	1/511	34	1/632	16	1/239	NF	1/124[8]
HEXA	c.805+1G>A	Tay-Sachs disease (3938)	Imputed data	22	1/1,139	7	1/3,067	15	1/239	NF	NA[9]
BRCA2	c.8537_8538del	Breast-ovarian cancer, familial, susceptibility to, 2 (9328)	Imputed data	18	1/1,392	13	1/1,652	5	1/718	NF	NA[9]
BRCA1	c.4327C>T	Breast-ovarian cancer, familial, susceptibility to, 1 (17675)	Imputed data	15	1/1,671	14	1/1,534	1	1/3,589	NF	NA[9]
GJB6	c.31G>A	Hidrotic ectodermal dysplasia syndrome (5544)	Imputed data	9	1/2,785	9	1/2,386	0	NA	NF	NA[9]
PAH	c.896T>G	Phenylketonuria (613)	Imputed data	26	1/1,044	23	1/1,022	3	1/1,196	NF	NA[10]
PAH	c.1A>G	Phenylketonuria (586)	Imputed data	16	1/1,566	14	1/1,534	2	1/1,794	NF	NA[9]
PAH	c.117C>G	Phenylketonuria (605)	Imputed data	7	1/3,580	7	1/3,067	0	NA	NF	NA[10]
PAH	c.1045T>C	Phenylketonuria (615)	WGS	1	1/1,852	1	1/1,538	0	NA	NF	NA[10]

SLSJ: Saguenay–Lac-Saint-Jean, UQc: Urban Quebec regions, QcP: Quebec province, WGS: Whole-genome sequencing, CR: Carrier rate, gray: nonfounder diseases according to our criteria.

Table 2: Novel founder variants found in this study.

					QcP		UQc		SLSJ	
Inheritance	**Gene**	**Nucleotide**	**Disease name (ClinVar ID)**	**Data type**	**Count**	**CR**	**Count**	**CR**	**Count**	**CR**
AD/AR	DNAH8	c.8635_8636del	Primary ciliary dyskinesia (2037549)	Imputed data	224	1/115	55	1/390	169	1/22
AR	CNGA1	c.947C>T	Retinitis pigmentosa 49 (16932)	Imputed data	217	1/119	78	1/275	139	1/27
AR	CTU2	c.881C>A	Dysmorphic facies, renal agenesis, ambiguous genitalia, microcephaly, polydactyly, and lissencephaly (2067774)	Imputed data	194	1/129	81	1/265	113	1/32
AR	TMEM107	c.*759C>T	Leukoencephalopathy with calcifications and cysts (265788)	Imputed data	195	1/129	125	1/172	70	1/51
AR	ENPP1	c.583T>C	ENPP1-related disorder (2580630)	WGS	7	1/309	1	1/1,538	6	1/52
AD/AR	RGS9	c.895T>C	Leber congenital amaurosis (5862)	Imputed data	112	1/224	54	1/398	58	1/62
AR	TRIOBP	c.1933C>T	Autosomal AR nonsyndromic hearing loss 28 (620162)	Imputed data	112	1/224	64	1/336	48	1/75
AR	UROS	c.217T>C	Cutaneous porphyria (3750)	Imputed data	84	1/298	36	1/596	48	1/75
AR	ASPM	c.8191_8192del	Microcephaly 5, primary, autosomal AR (21613)	Imputed data	79	1/317	32	1/671	47	1/76
AR	PYGM	c.148C>T	Glycogen storage disease, type V (2298)	Imputed data	151	1/166	109	1/197	42	1/85
AR	CEP290	c.7220_7223del	Meckel syndrome, type 4|Bardet-Biedl syndrome 14 (418123)	Imputed data	70	1/358	30	1/716	40	1/90
AR	DONSON	c.1047-9A>G	Microcephaly, short stature, and limb abnormalities (431414)	Imputed data	51	1/491	12	1/1,789	39	1/92
AD	PKD1	c.9829C>T	Polycystic kidney disease, adult type (192320)	Imputed data	86	1/291	47	1/457	39	1/92
AR	ETFA	c.495_496del	Multiple acyl-CoA dehydrogenase deficiency (459956)	Imputed data	96	1/261	61	1/352	35	1/103
AR	DNAH9	c.1733del	DNAH9-related disorder (3013954)	Imputed data	54	1/464	20	1/1,074	34	1/106
AR	MOCOS	c.2326C>T	Xanthinuria type II (253162)	Imputed data	57	1/456	24	1/895	33	1/116

AR	CCDC40	c.961C>T	Primary ciliary dyskinesia 15 (216118)	Imputed data	35	1/716	7	1/3,067	28	1/128
AR	SLC26A4	c.1001+1G>A	Pendred syndrome (4819)	Imputed data	52	1/482	25	1/859	27	1/133
AD/AR	EIF2AK4	c.1153dup	Familial pulmonary capillary hemangiomatosis (101527)	Imputed data	34	1/737	8	1/2,684	26	1/138
AR	TSHB	c.373del	Isolated thyroid-stimulating hormone deficiency (437070)	Imputed data	40	1/627	14	1/1,534	26	1/138
AR	DYNC2I2	c.1312_1313del	Short-rib thoracic dysplasia 11 with or without polydactyly (665979)	Imputed data	183	1/137	158	1/136	25	1/144
AR	PHKB	c.1257T>A	Glycogen storage disease IXb (13620)	Imputed data	30	1/835	6	1/3,579	24	1/150
Unknown	CDK5RAP2	c.2202+1G>A	not provided (1066422)	Imputed data	100	1/251	77	1/279	23	1/156
AD/AR	KCNJ1	c.472G>A	Bartter syndrome (2506156)	Imputed data	36	1/696	13	1/1,652	23	1/156
AR	ASAH1	c.410A>G	Spinal muscular atrophy-progressive myoclonic epilepsy syndrome (375548)	Imputed data	95	1/267	73	1/298	22	1/163
Unknown	DCAF6	c.2240G>A	Cerebral visual impairment and intellectual disability (224814)	Imputed data	83	1/302	61	1/352	22	1/163
AR	PKHD1	c.6793C>T	Autosomal AR polycystic kidney disease (1946278)	Imputed data	129	1/194	107	1/201	22	1/163
AR	TYR	c.572del	Tyrosinase-negative oculocutaneous albinism (99570)	Imputed data	63	1/411	41	1/551	22	1/163
AR	ALMS1	c.11648_11649insGTTA	Alstrom syndrome (550627)	Imputed data	93	1/269	72	1/298	21	1/171
AR	ERCC2	c.2164C>T	Cerebrooculofacioskeletal syndrome 2 (16792)	Imputed data	24	1/1,044	3	1/7,157	21	1/171
Unknown	RAD50	c.3779del	Hereditary cancer-predisposing syndrome (185537)	Imputed data	27	1/928	6	1/3,579	21	1/171
AR	SLC45A2	c.264del	Oculocutaneous albinism type 4 (242518)	Imputed data	77	1/325	56	1/383	21	1/171
AR	RMRP	n.71A>G	Metaphyseal chondrodysplasia, McKusick type (14208)	Imputed data	125	1/200	105	1/204	20	1/179
AD	CHEK2	c.247del	Hereditary cancer-predisposing syndrome (142851)	Imputed data	29	1/864	10	1/2,147	19	1/189
AR	NPHS1	c.2071+2T>C	Finnish congenital nephrotic syndrome (56460)	Imputed data	38	1/660	19	1/1,130	19	1/189
Unknown	PKLR	c.1091G>A	PKLR-related disorder (1456959)	Imputed data	97	1/258	78	1/275	19	1/189
AR	ACY1	c.575dup	Aminoacylase 1 deficiency (800812)	Imputed data	58	1/448	40	1/565	18	1/199

AD/AR	CAPN3	c.2115+1G>A	Autosomal AR limb-girdle muscular dystrophy type 2A (555599)	Imputed data	26	1/964	8	1/2,684	18	1/199
AR	GMPPB	c.79G>C	Autosomal AR limb-girdle muscular dystrophy type 2T (60546)	Imputed data	47	1/533	29	1/740	18	1/199
AR	NDUFV1	c.1162+4A>C	Mitochondrial complex I deficiency, nuclear type 1 (372716)	Imputed data	52	1/482	34	1/632	18	1/199
AR	RSPH3	c.859+1G>T	Primary ciliary dyskinesia 32 (2980542)	Imputed data	35	1/716	17	1/1,263	18	1/199
Unknown	LARS1	c.2500A>T*	not specified (3117894)	Imputed data	124	1/202	109	1/197	15	1/239

AD: Autosomal dominant, AR: Autosomal recessive, SLSJ: Saguenay–Lac-Saint-Jean, UQc: Urban Quebec regions, QcP: Quebec province, WGS: Whole-genome sequencing, CR: Carrier rate. * Founder variant only in UQc.

Founder variants' regional carrier rates and individuals' mutation load

We then compared the carrier rates between the SLSJ and UQc (Fig.3). Most of the already reported founder variants are at higher CR than the newly identified ones, but some of the latters are as high as 1/22 in the SLSJ (Table 2). Carrier rates are generally higher in the SLSJ compared to the UQc. Specifically, the count of variants with carrier rates higher than 1/200 is 8 times higher in SLSJ than in the UQc (3 times higher when considering all variants with an RFD≥10% regardless of whether they are founder variants) (Fig.4). Consequently, the number of individuals who carry at least one potentially pathogenic founder variant is higher in the SLSJ than in the UQc (chi^2 p-value < $2e^{-16}$) (Fig.5). In fact, for the variants already reported in the literature, 50% of the SLSJ and only 11% of the UQc individuals carry at least one variant. Notably, when the newly identified variants are added, these percentages reach 66% and 18%, respectively (chi^2 p-value < $2e^{-16}$).

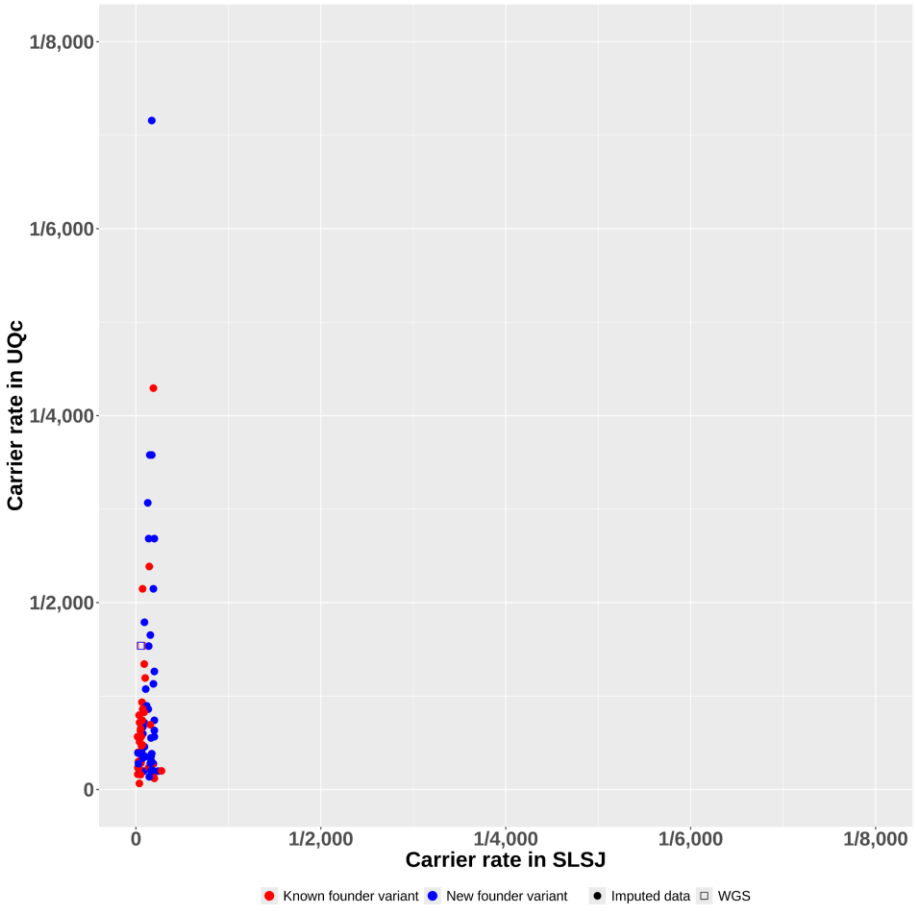

Fig.3: Carrier rates for founder variants. *Only variants classified as founders in SLSJ or UQc or QcP are shown here (80 variants). When available, the CR from the imputed data was used; otherwise, the CR from WGS data was utilized.*

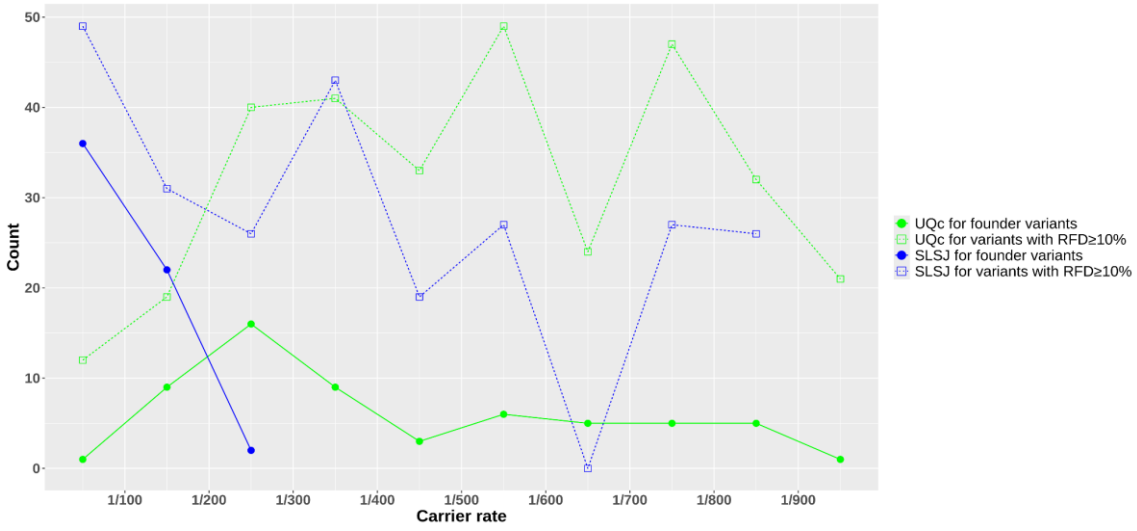

Fig.4: Number of variants in each carrier rate's class. *When available, the CR from the imputed data was used; otherwise, the CR from WGS data was utilized.*

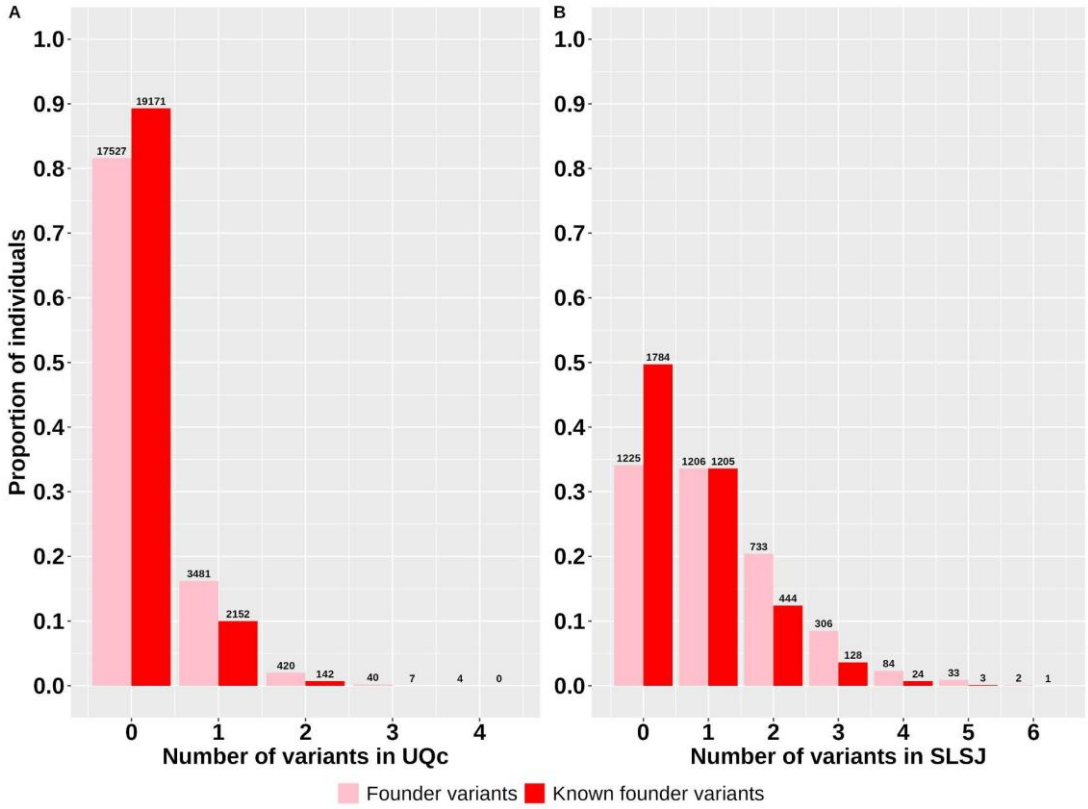

Fig.5: Proportion of individuals carrying founder variants in A) UQc and B) SLSJ. *Only rare variants classified as founders in SLSJ or UQc or QcP are shown here (80 variants).*

Clinical validation

To confirm that our method identifies clinically relevant variants in the SLSJ population, we requested clinical experts to examine their databases seeking variants that segregate within families of patients presenting the corresponding phenotype. Table 3 presents the variants found in patients from the Medical Genetics service and the CMNM in addition to CARTaGENE phenotypes provided. Of note, 3 of the variants identified herein (Alstrom syndrome (550627), Multiple acyl-CoA dehydrogenase deficiency (459956) and Joubert syndrome 9 (217607)) would be good candidates to include in an ongoing effort for designing a new carrier test for the SLSJ population in the Medical Genetics service of the CIUSSS of the SLSJ.

Table 3: Clinical information for variants found in patients with corresponding phenotypes.

Inheritance	Gene	Nucleotide	Heterozygotes	Homozygotes	Clinic
AD*	PRPH2	c.554T>C	1	0	Genetic
AR	CC2D2A**	c.4667A>T	-	1	Genetic + CMNM
AR	PDZD7	c.2107del	4 (compound)	0	Genetic
AR*	EIF2AK4	c.1153dup	0	1	Genetic
AR	SLC45A2	c.264del	-	1	Genetic
AR*	TYR	c.572del	2 (compound)	1	Genetic + CaG
AR	ALMS1**	c.11648_11649insGTTA	-	1	Genetic
AR	ETFA**	c.495_496del	-	1	Genetic
AR	UROS	c.217T>C	3 (compound)	0	Genetic
AR	SLC26A4	c.1001+1G>A	-	3	Genetic
AD	CHEK2	c.247del	10	0	Genetic + CaG
AR	PKD1	c.9829C>T	-	1	CaG

Variants in gray were reported, but not in SLSJ while other variants were not reported, *: Also have another inheritance mode, **: considered for a new carrier test in the SLSJ, Heterozygotes: Number of heterozygous patients (for dominant diseases), Homozygotes: Number of homozygous patients (for recessive diseases), CaG: CARTaGENE phenotypes

Discussion

In this study we aimed to identify potentially pathogenic variants found at higher frequency in the QcP and more specifically in the SLSJ region. Starting from 240,716 variants in ClinVar, we found 1,302 rare variants with RFD≥10% in Quebec compared to gnomAD non-Finnish Europeans (nfe). Among these 1,302 variants, we identified 80 that met our criteria to be

classified as founders, with 38 being previously reported in the QcP. Note that we classified a founder variant as already known if it was reported at least once in the literature, either in the QcP or SLSJ or among French-Canadians. Consequently, we do not differentiate between the variants documented in studies for which the focus was on founder variants in Quebec[7–10] and the ones identified in case reports over the years[13,15–27].

In addition to taking the high carrier rate into account, we examined the shared IBD segments around the variant to identify variants associated with the founder effect. By doing so, we ensure that the variant originated from a single ancestor and was spread through drift in the population due to the founder effect. Additionally, keeping only variants with a CR of at least 1/200 avoids small familial or more recent sporadic increases in frequency that would not be attributable to a population founder effect. We also propose that variants with a high frequency in the population, but without a majority of pairs sharing a surrounding IBD segment, are likely the result of multiple introductions rather than a single one. Within the 42 variants previously characterized as founders in the literature and found in the present study[7–10], 19 did not meet our criteria in any of the QcP, UQc or SLSJ groups primarily due to an insufficient number of carriers (Supplementary Table 1). Among these, 9 were documented alongside another founder variant for the same disease, leaving 10 variants without evidence of being real founder variants according to our criteria. Of note, all 15 variants associated with phenylketonuria identified in this study (4 of which were previously reported) had carrier rates below 1/200, which means that this condition was not classified as a founder disease in the present analysis. The same was observed for 5 other diseases (highlighted in gray in Table 1).

Regarding diseases caused by multiple variants, the literature often inaccurately reports that these are all at the same carrier rate in the population. It is rather the summed CR of all variants associated with a given disease which is reported. Indeed, by aggregating the carrier rates of the 10 variants present in our data for Cystic Fibrosis (Supplementary Table 1), we arrive at a final

carrier rate of 1/16 which is very close to the one previously reported of 1/15[8]. However, each variant associated with one disease has its own CR and its own history. Among the 10 variants documented for Cystic Fibrosis[28], our analysis reveals that 9 are nonfounders, including the 2 most frequent which seem to originate from multiple introductions in the population since less than half of pairs share IBD at the variant's genomic location. The 7 other nonfounder variants are present with a carrier rate below 1/200. On the other hand, our findings suggest that 1 Cystic Fibrosis variant (CFTR c.489+1G>T) does meet our criteria. This variant is much more common in the SLSJ than elsewhere in Quebec indicating that it likely has risen in frequency during the Charlevoix-SLSJ regional founder effect. Thus, the increased prevalence of this disease in the SLSJ population can be partly attributed to multiple introductions of different genetic variants by various ancestors, with the Charlevoix-SLSJ founder effect being only one of several contributing factors.

In addition to confirming known founder variants, we also report for the first time 42 novel founder variants that, to our knowledge, have never been documented in the QcP. Some of these exhibit a high carrier rate, comparable to the most common known variants included in the carrier test offered to the population. These variants could potentially account for unreported rises in disease prevalence within the population, which suggests a potential underestimation of the overall prevalence of rare diseases in the SLSJ region as also reported in other populations with founder effects[29]. Indeed, adding the newly identified variants raises the proportion of individuals carrying at least one founder variant of 1.3 and 1.7 times in the SLSJ and in the UQc respectively. Establishing carrier rates plays a critical role in advancing precision medicine among populations with a founder effect[29]. In addition, it is a great proof-of-concept for larger initiatives to come in the field of precision medicine in regard to carrier frequency panels in larger populations.

Considering that we demonstrate an underestimation of the number of pathogenic variant carriers in SLSJ, which has been the focus of numerous studies on rare genetic diseases linked to the founder effect, we hypothesize that this phenomenon might also be present in other populations worldwide. Indeed, a rise in deleterious allele frequencies following range expansions has also been observed in other non-African populations[30]. Consequently, the number of individuals affected with a rare disease might be underestimated in many countries or local communities. Our population cohort's approach could be applied in other worldwide populations at low costs thus helping in enhancing and fastening the molecular diagnosis of patients.

When comparing pathogenic variant frequencies between gnomAD and the QcP, we observed almost 4 times more deleterious variants with an RFD ≥10% in the latter (Supplementary Fig.2) in line with the higher deleterious mutation load previously observed in rapidly expanding populations[31]. The present study also demonstrates the higher pathogenic mutational load of individuals from the SLSJ region compared to UQc not only for founder variants, but also for variants with an RFD ≥10%. It seems that the UQc group has a higher number of unique variants with an RFD ≥10%, while the SLSJ individuals are more likely to carry two or more variants. Indeed, 42% of variants with an RFD ≥10% in the QcP were lost in the SLSJ, although some of them might be too rare to be observed in the SLSJ due to the smaller sample size. As for the 80 founder variants, again a greater mutational load in addition to a higher number of variants with CR above 1/200 are observed in the individuals from the SLSJ compared with those from the UQc. It can be concluded that the higher mutation load in the SLSJ individuals is mainly caused by an overrepresentation of variants with a CR greater than 1/200. This is the result of the genetic bottleneck of the SLSJ followed by a very rapid population expansion, 5 times greater than the one observed in the whole Quebec for the same period[32], which represents one of the strongest regional founder effects in Quebec[6]. Some founders in this region contributed a lot to the present population[33] and therefore could have introduced an allele in the population that

would reach such a high frequency[34]. Moreover, it was demonstrated that the first SLSJ settlers had an increased fitness[35] which could have contributed to increasing deleterious allele frequencies[36,37] possibly due to increased drift and relaxed selection[38].

To validate that the variants identified in this study are associated with specific diseases, we searched clinical databases and CARTaGENE phenotypes for patients carrying those variants who have been diagnosed with the corresponding disease. Notably, 12 of the variants identified in this study were detected in patients from the CIUSSS of SLSJ clinics and/or in CARTaGENE. Those variants have not been previously reported in the SLSJ, although 3 of them (PRPH2 c.554T>C, CC2D2A c.4667A>T, PDZ7 c.2107del) were reported in the French-Canadian population.

This study has certain limitations. Firstly, the sample size of the WGS data may be insufficient to accurately estimate the frequency of variants in the population. Therefore, we chose to work with imputed data, which includes a significantly larger number of individuals. To achieve the most accurate representation of our data, especially given our focus on rare variants, we performed imputation using our WGS data rather than a global worldwide reference panel. However, we acknowledge that imputed data may not be as reliable as WGS or genotyping. Therefore, we compared the WGS data with the imputed data for the same individuals and excluded any imputed variants that were false positives or not present in the WGS data. This sometimes could affect comparisons of CR and cumulative CR for different variants due to differences in sample sizes between both data types. We also excluded any individual whose cluster in WGS did not match the one in the imputed data clustering based on the UMAP. Also, our definition of a founder variant is stringent with a CR of at least 1/200, especially for the SLSJ region where the sample size is smaller and this CR could not be reached for the WGS. As a result, some less common but genuine founder variants might be missed. Lastly, our conclusions regarding the mutation load apply only for the 1,302 variants with an RFD ≥10% in QcP as we did not assess the load on all variants with a confirmed pathogenicity in ClinVar.

In conclusion, this study demonstrates a greater mutation load for founder variants and for variants with an RFD ≥10% in the SLSJ region compared to urban Quebec areas due to its stronger founder effect. In fact, this load is driven by more numerous variants present at a higher frequency in the SLSJ population. In addition to confirming previously described pathogenic variants using a population cohort instead of a clinical approach, we found variants associated with diseases that were not yet described in Quebec or in the SLSJ. These findings might be crucial for clinicians to shorten the patients' diagnostic odyssey and reduce the economic burden associated with undiagnosed rare diseases. This could help improve the management of patients and, for some or them, enhance their quality of life and slow disease progression as appropriate treatments could be offered earlier. With this information, precision medicine can implement targeted genetic screening programs, allowing for early detection of inherited conditions that are more prevalent due to the founder effect. This enables tailored prevention strategies, personalized treatments, and risk-reduction measures that are specific to the genetics of the population. Additionally, pharmacogenomics can benefit from this knowledge by optimizing drug therapies based on genetic susceptibilities. Ultimately, understanding carrier rates in populations with a founder effect helps healthcare providers offer more precise and effective medical care, enhancing outcomes for both individuals and the community. Indeed we identified 30 patients carrying 12 causal variants that have not been previously reported in SLSJ. The underestimation of pathogenic mutational load might also happen in other populations as a result of range expansions and rare diseases might be much less rare than anticipated. In an era of precision medicine with at least 10% of the population affected by rare diseases, it is crucial to adopt new approaches to enhance and fasten the molecular diagnosis of rare diseases.

Data and Methods

This study was approved by the University of Quebec in Chicoutimi (UQAC) ethics board. The approval for the secondary use of anonymized samples coming from the provincial screening

testing was obtained from the CIUSSS of the SLSJ Direction of professional services. Written informed consent for the use of saliva samples were obtained from participants.

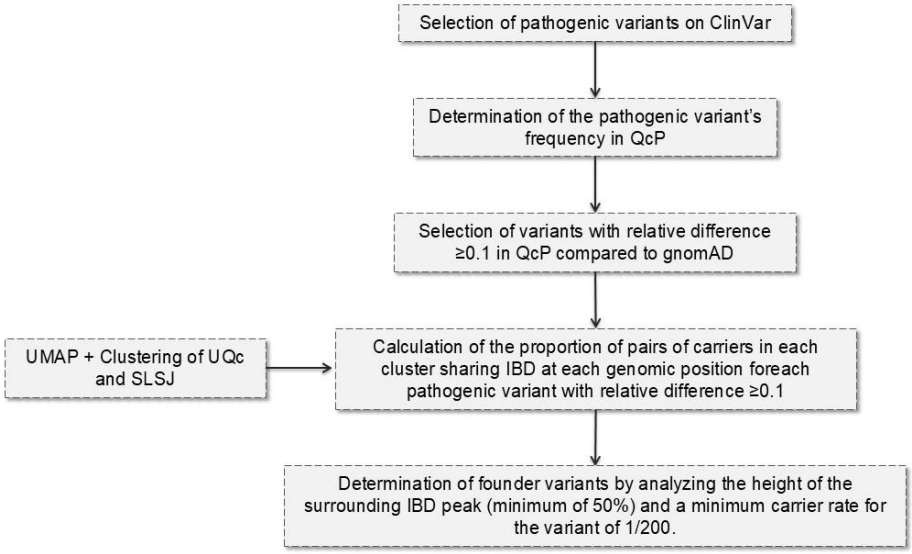

Fig.6: Simplified view of the flow of the analysis.

Cohort

The CARTaGENE cohort[39] (https://cartagene.qc.ca/) used in this study includes WGS of 2,184 and genotyping of 29,337 participants. Individuals aged between 40 and 69, residing in 6 distinct cities (Montreal, Quebec City, Trois-Rivières, Sherbrooke, Gatineau, Saguenay), were recruited between 2009 and 2015, regardless of their birthplace. The CARTaGENE cohort also includes a wide range of phenotypes, among which are the occurrence of a disease. The genotype and WGS data quality control are described here (https://cartagene.qc.ca/files/documents/other/Info_GeneticData3juillet2023.pdf). All genomic data were alignmed on the GRCh38 genome assembly.

Genotypes cleaning and imputation

To increase our sample size and achieve more accurate carrier rates, we imputed the 6 different CARTaGENE genotyping chips using SHAPEIT5[40] and IMPUTE5[41]. The individuals were genotyped

on different arrays (Omni 2.5, GSAv1 + Multi disease panel, GSAv1, GSAv2 + Multi disease panel, GSAv3 + Multi disease panel, GSAv2 + Multi disease panel + addon and Affymetrix Axiom 2.0) (https://cartagene.qc.ca/files/documents/other/Info_GeneticData3juillet2023.pdf), and were cleaned and merged as follows. Each dataset was cleaned separately using PLINK software v1.9[42], ensuring individuals with at least 95% genotypes among all SNPs were retained. At the SNP level, we retained SNPs with at least 95% genotypes among all individuals, located on the autosomes and in Hardy–Weinberg equilibrium $p > 10^{-6}$ (calculated on each dataset).

The imputation was performed on each genotyping batch separately using the 2,184 CARTaGENE WGS as reference to enhance our capacity to identify rare variants within our population. All batches were then merged and the final imputed dataset includes 29,353 individuals. A postimputation quality control filter was applied on each individual imputed batch to remove variants with an imputation quality score <0.3 for the PCA and UMAP.

UMAP and clustering according to individuals' origin

As a reminder, the CARTaGENE individuals' recruitment site was based on their current residence rather than their birth place, even though many individuals live in their place of birth. For the purpose of this study, we needed to identify the most related individuals based on genetics, regardless of where they were recruited. To do so, a PCA was performed using PLINK on the WGS SNPs with a minor allele frequency (MAF) of at least 5% and after removal of SNPs with more than 2% missing individuals and in LD. We retained only biallelic SNPs within the accessibility mask[43], resulting in a total of 90,073 remaining SNPs. We also filtered out individuals with more than 2% missing SNPs resulting in 2,166 individuals remaining. A UMAP[44] was then performed on the 3 first PCs (determined by the scree test) with the R umap library v0.9.2.0 (Supplementary Fig.3). This technique was proven efficient to reveal fine-scale population structure[45]. K-means clustering was then employed to create 3 clusters, aiming to retain as many individuals from the SLSJ as possible, given its limited sample size, especially for the WGS data.

We also intended to choose individuals with the strongest ancestry connection to the region. Based on the recruitment place (Supplementary Fig.3A), we could see that the majority of the CARTaGENE participants recruited from the SLSJ region belongs to the red cluster (Supplementary Fig.3B). In fact, the red cluster of the WGS data encompasses 90% of the individuals who were recruited from the SLSJ region. We identified 314 individuals originating from the SLSJ region (red cluster) who were recruited in different places and 1,538 individuals from the other urban Quebec regions (UQc) (green cluster), for a total of 1,852 for the QcP (green and red clusters). Clusters were also defined on imputed data as described above on pruned SNPs at 5% frequency or more keeping 5 PCs for the UMAP, leaving 3,589 individuals in the SLSJ (red cluster) and 21,472 in the UQc (green cluster), for a total of 25,061 in the QcP (Supplementary Fig.4). For the imputed data, the red cluster encompasses 84% of the individuals who were recruited from the SLSJ region. We ensured consistency of individuals in clusters between the WGS and imputed data by removing 27 samples that exhibited mismatches, likely due to sample mix-ups in one dataset or because they were at the boundaries of both clusters. This method ensures that individuals have a common genetic background and was shown to be helpful in uncovering rare variants with smaller sample sizes[46,47].

Selection of pathogenic variants with relative frequency difference ≥ 10%

Variants' classification was extracted from the ClinVar database version of June 24, 2024[48]. Only variants classified as: Pathogenic, Likely pathogenic, and conflicting (both pathogenic and likely pathogenic variants), as well as SNPs, insertions and deletions (indels), were included in the analysis whereas repeat expansions were excluded. Furthermore, variants with the following review status were removed: no assertion criteria provided, no classification provided, no classification for the individual variant. Additionally, we incorporated all variants referenced as founder variants in previous studies[7–10], regardless of their status on ClinVar. Therefore, we obtain a list of 240,716 variants.

We calculated the variants' frequency in the WGS and imputed data using PLINK v1.9 for the individuals originating from the SLSJ, UQc and QcP (both clusters) inferred by the clustering. Whenever a variant was present and was not a false positive in the imputed data, we used the imputed variant frequency, otherwise, we relied on the WGS variant frequency. Only WGS variants with less than 10% missing individuals were kept. Notably, the frequencies of variants show strong correlation between both data types (Supplementary Fig.5). The gnomAD frequencies for the non-Finnish Europeans (non_topmed_nfe) were directly extracted from gnomAD genomes v3.1.2[49]. To calculate the relative frequency difference (RFD) of a variant in the QcP compared to gnomAD nfe, we used the following formula:

$$RFD = \frac{freq_{QcP} - freq_{gnomAD}}{freq_{QcP}}$$

Knowing that:

- $freq_{QcP}$ corresponds to the frequency of the variant in the QcP population.
- $freq_{gnomAD}$ corresponds to the frequency of the variant in the gnomAD non-Finnish Europeans.

We fixed a minimum RFD threshold of 0.1 to make sure it encompasses all variants that could be of interest. We detected 1,304 potentially pathogenic variants in the WGS of 1,852 individuals within the QcP that reached RFD≥10% compared to gnomAD nfe. Since we are focussing on rare variants, we removed 2 variants with a MAF ≥ 5% leaving 1,302 rare variants with RFD≥10% in the QcP.

Estimation of carrier rate

We directly counted the number of heterozygotes for each variant and determined the CR by calculating the inverse of the frequency of the heterozygous individuals as follows:

$CR = \frac{1}{f_{hetero}}$.

Selection of founder variants

Furthermore, we established criteria for a variant to be called as founder. The number of individuals carrying the variant must be adequate to avoid false signals or misinterpretations while also being high enough to be clinically relevant; thus, the target CR was set to 1/200. Hence, we set the minimum threshold at 5 (1/63), 8 (1/192) and 10 (1/185) individuals carrying the variant for WGS of SLSJ, UQc and QcP respectively. However, for imputed data, we set the threshold to reach a CR of 1/200 which represents 18, 108 and 126 individuals for the SLSJ, UQc and QcP. Furthermore, to be called as founder, a variant must show a proportion of pairs of carriers sharing an IBD segment around the variant (see next section) of at least 0.5.

IBD segments inference and sharing

All cleaned genotyping batches (excluding the Affymetrix chip due to its poor SNPs intersection with other Illumina chips) were combined and only the intersecting common SNPs were kept. 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 148,200 SNPs and 28,358 individuals.

We then inferred IBD segments on phased genotypes using refinedIBD[50] version 17Jan20 and Beagle version 18May20. Subsequently, the segments were merged using the merge-ibd-segments 17Jan20.102 tool. We retained only IBD segments of 2Mb or longer and with a LOD score greater than 3.

We then examined the genome-wide IBD segments shared among pairs of individuals carrying a specific pathogenic variant with RFD≥10% and determined the proportion of pairs sharing IBD at each genomic position. We considered the variant as a founder variant if this proportion reached a minimum of 0.5, indicating that half of the pairs share IBD at the variant's location. Considering that a founder variant usually originates from a single ancestor in a population with a founder effect[34], and within a relatively recent time frame (with the first permanent

settlement in Quebec starting in 1608), this variant is often inherited with other variants in LD. Therefore, examining the IBD sharing around a variant is a dependable method to confirm its status as a true founder variant.

Calculation of genome-wide relatedness

To validate that our findings are not subject to potential high relatedness biases, we conducted genetic kinship calculations among individuals carrying the same variant. To do so, we assessed the total length of IBD segments shared between pairs of individuals divided by twice the length of the genome (to account for the diploid human genome). If the resulting percentage exceeded 50%, it suggested a potential first-degree relationship between the individuals. Detecting these relationships could reveal potential biases in our IBD analysis, particularly if individuals shared a recent common ancestor. In such cases, the IBD segments they share may not accurately represent a founder effect, but instead, direct familial transmission from the recent ancestor. For carriers of founder variants, the average whole genome IBD sharing between pairs of individuals with the same variant never exceeds 5%. This suggests that the observed proportion of individuals sharing IBD around the founder variants is not attributable to close relatedness.

Clinical data

The patient group consisted of individuals residing in SLSJ during the assessment, all of whom had genetic disorders. They were clinically evaluated at the Medical Genetics service and the CMNM of CIUSSS of SLSJ. Their DNA samples were analyzed in certified clinical molecular laboratories as part of the clinical testing and genetic evaluation process. A review of internal databases and the patients' medical records enabled the identification of patients who were homozygous, compound heterozygous, or heterozygous for autosomal recessive or dominant variants.

Experimental validation of carrier rates

To validate the estimated carrier rates derived from the WGS and imputed data, we randomly selected 1,000 individuals from the SLSJ who had consented to the storage of their anonymized DNA samples for research purposes. We chose 2 founder variants, DOK7 c.1124_1127dup and CTNS c.414G>A, for genotyping due to their high CR observed in the SLSJ in the present study (1/21 and 1/63, respectively). They were genotyped using custom TaqMan genotyping assays (catalog #4332072; Applied Biosystem Inc). We designed the assays with specific probes targeting each allele using the Primer Express software. We extracted the DNA from buccal swab samples using DNA extract all kit (catalog #4402616; Applied Biosystem Inc) following the manufacturer recommendations. In brief, 22.0 µl of Lysis solution was added to 7.0 µl of buccal swab emulsions, then incubated for 3 minutes at 95°C in a thermocycler. 22.0 µl of DNA stabilizing solution was then added to the mix. For amplification and detection, the manufacturer recommendations were followed. In brief, for each reaction 1.75 µl of sterile water, 3.10 µl of GTXpress Master Mix (catalog #4401892; Applied Biosystem Inc), 0.15 µl of TaqMan assay, and 1.2 µl of DNA. Analysis was carried out in a 96-well plate and samples were amplified on a 7500 Fast Real-Time PCR thermocycler (Applied Biosystems Inc). The amplification conditions were as follows: 1: 60°C for 1 min with fluorescence acquisition, 2: 95°C for 20 s, 3: 95°C for 3 s, and 60°C for 30 s with fluorescence acquisition (step 3 was repeated 40 times), 4: 60°C for 1 min with fluorescence acquisition. The genotypes were called using 7500 Software v2.0.1 (Applied Biosystem Inc) after visual inspection of the amplification data. For the first variant, 33 samples were excluded from the analysis due to unsuccessful PCR amplification (N=967), while for the second variant, 10 samples were excluded from the analysis (N=990).

Data availability

Quebec genotype, imputed and WGS data are available via an independent data access committee by the CARTaGENE cohort (https://cartagene. qc.ca/en/researchers/access-request.html).

Code availability

The code used for this study can be found in the following GitHub repository: https://github.com/ElisaMichel/Founder_variant_2024.

Acknowledgements

Funding for SLG was provided by the Canada Research Chair in Genetics and Genealogy. LB and CG were funded by the Research Chair in *Génétique et parcours de vie en santé.*

References

1. Haendel, M. *et al.* How many rare diseases are there? *Nat Rev Drug Discov* **19**, 77–78 (2020).
2. Orphadata – Orphanet datasets. https://www.orphadata.com/.
3. Tesi, B. *et al.* Precision medicine in rare diseases: What is next? *Journal of Internal Medicine* **294**, 397–412 (2023).
4. Bauskis, A., Strange, C., Molster, C. & Fisher, C. The diagnostic odyssey: insights from parents of children living with an undiagnosed condition. *Orphanet Journal of Rare Diseases* **17**, 233 (2022).
5. Charbonneau, H., Desjardins, B., Légaré, J. & Denis, H. The population of the St-Lawrence Valley, 1608-1760. *A Population History of North America* 99–142 (2000).
6. Gagnon, L., Moreau, C., Laprise, C., Vézina, H. & Girard, S. L. Deciphering the genetic structure of the Quebec founder population using genealogies. *Eur J Hum Genet* 1–7 (2023) doi:10.1038/s41431-023-01356-2.
7. 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).
8. Cruz Marino, T. *et al.* Portrait of autosomal recessive diseases in the French-Canadian

founder population of Saguenay-Lac-Saint-Jean. *American Journal of Medical Genetics Part A* **191**, 1145–1163 (2023).

9. Laberge, A.-M. *et al.* Population history and its impact on medical genetics in Quebec. *Clinical Genetics* **68**, 287–301 (2005).
10. Scriver, C. R. Human genetics: lessons from Quebec populations. *Annu Rev Genomics Hum Genet* **2**, 69–101 (2001).
11. Que pensent et que savent les 18-44 ans du Saguenay-Lac-Saint-Jean au sujet de la génétique et des maladies héréditaires? | Journées annuelles de santé publique (JASP). *Institut national de santé publique du Québec* https://www.inspq.qc.ca/jasp/que-pensent-et-que-savent-les-18-44-ans-du-saguenay-lac-saint-jean-au-sujet-de-la-genetique-et-des-maladies-hereditaires.
12. 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).
13. Müller, J. S. *et al.* Phenotypical spectrum of DOK7 mutations in congenital myasthenic syndromes. *Brain* **130**, 1497–1506 (2007).
14. Srour, M. *et al.* DOK7 mutations presenting as a proximal myopathy in French Canadians. *Neuromuscul Disord* **20**, 453–457 (2010).
15. Srour, M. *et al.* Mutations in C5ORF42 cause Joubert syndrome in the French Canadian population. *Am J Hum Genet* **90**, 693–700 (2012).
16. Brown, S. J. *et al.* Loss-of-function variants in the filaggrin gene are a significant risk factor for peanut allergy. *J Allergy Clin Immunol* **127**, 661–667 (2011).
17. Cruz Marino, T. *et al.* First glance at the molecular etiology of hearing loss in French-Canadian families from Saguenay-Lac-Saint-Jean's founder population. *Hum Genet* **141**, 607–622 (2022).
18. Chetaille, P. *et al.* Mutations in SGOL1 cause a novel cohesinopathy affecting heart and gut

rhythm. *Nat Genet* **46**, 1245–1249 (2014).

19. Ambalavanan, A. *et al.* De novo variants in sporadic cases of childhood onset schizophrenia. *Eur J Hum Genet* **24**, 944–948 (2016).
20. Dupré, N. *et al.* Clinical, electrophysiologic, and genetic study of non-dystrophic myotonia in French-Canadians. *Neuromuscul Disord* **19**, 330–334 (2009).
21. Boissel, S. *et al.* Genomic study of severe fetal anomalies and discovery of GREB1L mutations in renal agenesis. *Genet Med* **20**, 745–753 (2018).
22. Chan, E. M. *et al.* Mutations in NHLRC1 cause progressive myoclonus epilepsy. *Nat Genet* **35**, 125–127 (2003).
23. Coussa, R. G. *et al.* Genotype and Phenotype Studies in Autosomal Dominant Retinitis Pigmentosa (adRP) of the French Canadian Founder Population. *Invest Ophthalmol Vis Sci* **56**, 8297–8305 (2015).
24. Ebermann, I. *et al.* PDZD7 is a modifier of retinal disease and a contributor to digenic Usher syndrome. *J Clin Invest* **120**, 1812–1823 (2010).
25. Kitzler, T. M., Kachurina, N., Bitzan, M. M., Torban, E. & Goodyer, P. R. Use of genomic and functional analysis to characterize patients with steroid-resistant nephrotic syndrome. *Pediatr Nephrol* **33**, 1741–1750 (2018).
26. La Piana, R. *et al.* Spastic paraparesis and marked improvement of leukoencephalopathy in Aicardi-Goutières syndrome. *Neuropediatrics* **45**, 406–410 (2014).
27. Haj Salem, I. *et al.* Genetic and Epidemiological Study of Adult Ataxia and Spastic Paraplegia in Eastern Quebec. *Can J Neurol Sci* **48**, 655–665 (2021).
28. Bepari, K. K., Malakar, A. K., Paul, P., Halder, B. & Chakraborty, S. Allele frequency for Cystic fibrosis in Indians vis-a/-vis global populations. *Bioinformation* **11**, 348–352 (2015).
29. Mathijssen, I. B. *et al.* With expanded carrier screening, founder populations run the risk of being overlooked. *Journal of Community Genetics* **8**, 327 (2017).
30. Henn, B. M. *et al.* Distance from sub-Saharan Africa predicts mutational load in diverse

human genomes. *Proc Natl Acad Sci U S A* **113**, E440-449 (2016).

31. 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).
32. Pouyez, C. & Lavoie, Y. *Les Saguenayens: introduction à l'histoire des populations du Saguenay, XVIe-XXe siècles*. (Presses de l'Université du Québec, Sillery, 1983).
33. Bherer, C. *et al.* Admixed ancestry and stratification of Quebec regional populations. *Am J Phys Anthropol* **144**, 432–441 (2011).
34. Heyer, E. & Austerlitz, F. Update to Heyer's 'One founder/one gene hypothesis in a new expanding population' (1999). *Hum Biol* **81**, 657–662 (2009).
35. Moreau, C. *et al.* Deep human genealogies reveal a selective advantage to be on an expanding wave front. *Science* **334**, 1148–1150 (2011).
36. 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).
37. Peischl, S. *et al.* Relaxed Selection During a Recent Human Expansion. *Genetics* **208**, 763–777 (2018).
38. Gravel, S. When Is Selection Effective? *Genetics* **203**, 451–462 (2016).
39. Awadalla, P. *et al.* Cohort profile of the CARTaGENE study: Quebec's population-based biobank for public health and personalized genomics. *International Journal of Epidemiology* **42**, 1285–1299 (2013).
40. Hofmeister, R. J., Ribeiro, D. M., Rubinacci, S. & Delaneau, O. Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the UK Biobank. *Nat Genet* **55**, 1243–1249 (2023).
41. Rubinacci, S., Delaneau, O. & Marchini, J. Genotype imputation using the Positional Burrows Wheeler Transform. *PLoS Genet* **16**, e1009049 (2020).
42. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet* **81**, 559–575 (2007).

43. GRCh38 genome accessibility masks for 1000 Genomes data | 1000 Genomes. https://www.internationalgenome.org/announcements/genome-accessibility-masks/.
44. McConville, R., Santos-Rodríguez, R., Piechocki, R. J. & Craddock, I. N2D: (Not Too) Deep Clustering via Clustering the Local Manifold of an Autoencoded Embedding. in *2020 25th International Conference on Pattern Recognition (ICPR)* 5145–5152 (2021). doi:10.1109/ICPR48806.2021.9413131.
45. Diaz-Papkovich, A., Anderson-Trocmé, L., Ben-Eghan, C. & Gravel, S. UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. *PLOS Genetics* **15**, e1008432 (2019).
46. Diaz-Papkovich, A. *et al.* Topological stratification of continuous genetic variation in large biobanks. 2023.07.06.548007 Preprint at https://doi.org/10.1101/2023.07.06.548007 (2023).
47. Gagnon, L., Moreau, C., Laprise, C. & Girard, S. L. Fine-scale genetic structure and rare variant frequencies. Preprint at https://doi.org/10.1101/2024.02.02.578687 (2024).
48. Index of /pub/clinvar/vcf_GRCh38. https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh38/.
49. Chen, S. *et al.* A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625**, 92–100 (2024).
50. Browning, B. L. & Browning, S. R. Improving the Accuracy and Efficiency of Identity-by-Descent Detection in Population Data. *Genetics* **194**, 459–471 (2013).

Chapitre 3 : Discussion

Ce dernier chapitre va nous permettre de revenir sur les différents éléments présentés dans ce mémoire. Nous allons commencer par discuter des données utilisées ainsi que les méthodes et les objectifs introduits lors du chapitre 1. Ensuite, nous aborderons les résultats présentés dans l'article écrit lors de mon mémoire et qui constituent le chapitre 2. De plus, nous parlerons des limitations possibles aussi bien pour les méthodes que pour les conclusions faites précédemment. Enfin, nous finirons avec la mise en avant des perspectives concernant ce projet et à quoi celui-ci peut aboutir.

1. Retour sur les Chapitres

Le chapitre 1 aborde la méthodologie suivie lors de ce projet ainsi que les données sur lesquelles repose ce dernier. Ce chapitre est donc fondamental pour la bonne compréhension du chapitre suivant qui explicite le projet global de ma maîtrise. L'objectif de ce chapitre est donc d'introduire les méthodes et de les expliciter mais aussi de présenter les données utilisées lors de ce projet. Ainsi, ce chapitre aborde les cohortes populationnelles utilisées lors du projet et le type de données que sont les WGS et les données imputées. Afin de pallier aux problèmes liés aux données imputées et qui sont détaillés précédemment, nous avons utilisé le séquençage du génome entier comme référence afin de compléter le génotypage à notre disposition et ainsi de créer les données imputées. Cette astuce permet d'avoir des données robustes et un nombre suffisant d'individus pour considérer notre cohorte représentative de la population globale. En ce qui concerne les données utilisées lors de ce projet, celles-ci proviennent de la plateforme de recherche publique CARTaGENE. Le regroupement des individus dans cette base de données s'est fait selon le lieu de recrutement des individus et non pas selon leur ville d'origine. Or, dans cette étude nous essayons de mettre en évidence des variants spécifiques à chaque région selon l'impact de l'effet fondateur. Ainsi, cette façon de regrouper les individus ne nous convenait pas et nous avons donc utilisé les outils en génétique présentés dans le chapitre afin de former des

groupes d'individus selon leur proximité génétique. Nous avons donc utilisé l'analyse en composantes principales ainsi que l'UMAP pour avoir la meilleure représentation des données en prenant en compte le plus possible toute sa complexité. Après cela, nous avons formé les groupes d'individus selon leur proximité génétique avec la méthode k-means. Les résultats de ces groupements sont présentés dans le chapitre 2. La formation de ces groupes selon leur proximité génétique était nécessaire afin de déterminer les variants spécifiques aux régions étudiées dans ce projet. En effet, on peut voir dans les graphiques présentés précédemment que les regroupements sont nettement différents de ceux considérés de base selon le lieu de recrutement. Notamment, on constate de nombreuses personnes qui n'étaient pas considérées comme venant du SLSJ dont les génomes ont montré le plus de similitudes avec des personnes de cette région. Ensuite, est abordée l'identification des variants fondateurs grâce à l'utilisation des segments IBD. Ces derniers permettent de mettre en avant l'origine ancestrale commune de variants chez des individus peu apparentés et traduisant de ce fait l'effet fondateur. Leur utilisation nous a donc permis de caractériser des variants comme fondateurs avec une assurance forte et possible grâce à ces segments. Ainsi, l'introduction aux méthodes et données utilisées lors de ce projet est bien en accord avec les objectifs déterminés pour ce chapitre. Enfin, les objectifs pour le projet dans sa globalité sont aussi présentés dans ce chapitre. Or, on peut penser que l'identification des variants faite lors du projet a bien permis de répondre aux objectifs fixés lors de ma maîtrise. Ils sont notamment détaillés dans le chapitre suivant.

Le but du chapitre 2 était d'appliquer les méthodes décrites précédemment aux données auxquelles nous avions accès afin d'identifier les variants plus fréquents au SLSJ dû à l'effet fondateur. Ainsi, le chapitre 2 se compose de l'article écrit lors de ma maîtrise avec le soutien de l'équipe de mon laboratoire. Celui-ci aborde les méthodes détaillées du traitement des données génétiques ainsi que les analyses telles que la PCA et l'UMAP. Ces dernières nous ont permis d'avoir des populations définies selon leur proximité génétique associées aux données génétiques de celles-ci. Ainsi, les variants considérés comme fondateurs ont pu être déterminés

et les plus pertinents sont indiqués dans les tableaux de ce chapitre. On peut ainsi voir que l'on retrouve de nombreux variants déjà identifiés comme plus fréquents dans la région du SLSJ. Cependant, notre méthode permet d'affirmer qu'ils sont bien fondateurs, confirmant les hypothèses des cliniciens qui les ont répertoriés au cours des années. En plus des variants déjà identifiés, nos analyses ont permis de mettre en avant 42 variants qui n'avaient jamais été rapportés au SLSJ et qui sont, selon notre méthode, fondateurs. Ainsi, grâce à une collaboration avec les cliniciens, il a pu être décidé que 3 variants étaient éligibles pour les tests de porteurs. Cela a été décidé selon leur taux de porteur et leur pathogénicité. De ce fait, notre nouvelle approche, avec l'utilisation non pas des symptômes des patients mais de leur génome, permet d'apporter un nouveau point de vue sur les maladies génétiques rares. Le but final étant de faciliter le travail des cliniciens en simplifiant le diagnostic de ces maladies rares et ainsi améliorer la qualité de vie des patients en leur fournissant des traitements adaptés et plus rapidement. Enfin, avec cette nouvelle méthode, il a été possible de mettre en avant des variants dont la fréquence est élevée dans la population non pas par leur caractère fondateur mais parce que ces derniers ont été introduits par plusieurs personnes lors du peuplement. Cela se traduit par un faible partage des segments IBD au niveau du variant parmi les individus portant ce dernier. Ainsi, il est possible de déterminer si un variant a eu un changement de fréquence lors de l'effet fondateur ou bien lors de son introduction dans la population par plusieurs personnes. Pour conclure sur ce chapitre, les objectifs fixés initialement ont bien été atteints et cela a été possible grâce à l'identification des segments IBD et de la détermination de la fréquence des variants dans la région. De ce fait, des variants dont la fréquence a augmenté suite à l'effet fondateur ont été mis en évidence. Parmi ces derniers, on retrouve des variants qui ont déjà été décrits mais aussi d'autres dont la présence n'avait jamais été révélée auparavant et qui a été mise en avant grâce à notre méthode innovante. En effet, celle-ci se base sur l'étude des individus non malades d'une population afin de déterminer la présence de maladies génétiques rares. Cela apporte un nouveau point de vue car jusqu'à maintenant, les cliniciens se basent sur

les symptômes des patients atteints. Or comme expliqué précédemment, les maladies rares sont difficiles à diagnostiquer par le manque de cas cliniques. Ainsi, faire l'étude des variants plus fréquents dans une région permet d'apporter une nouvelle méthode afin de diagnostiquer les maladies rares et ainsi aider les patients en leur apportant un diagnostic précis et rapide.

2. Limitations

Les travaux qui ont été menés pendant ma maîtrise et qui sont ici exposés présentent certaines limitations. En effet, on peut dans un premier temps aborder la question de la taille de nos données en WGS qui pourrait être considérée comme insuffisante pour estimer correctement les fréquences des variants. Cependant, cette limitation a pu être compensée par l'utilisation des données imputées. Ainsi, nous avions les génotypes de nombreux individus et à partir de ces données, et en combinant avec les WGS, il a été possible d'avoir accès à des données imputées. Celles-ci étaient basées sur de nombreux individus, ce qui nous permet d'avoir des fréquences de variants robustes mais aussi elles ont été créées à partir des données de séquençage entier des génomes des individus de la province de Québec. Cela nous permet de pallier le problème des variants rares que nous aurions rencontré si nous avions pris des séquençages entiers de génome de panels de références mondiaux.

Dans un deuxième temps, une autre limitation est le seuil du taux de porteurs afin de caractériser un variant comme fondateur. Celui-ci a été fixé à 1/200 selon ce qui a été fait dans la littérature et celui-ci s'applique pour les données imputées. Or, sachant que le SLSJ est une région avec une population plus réduite, ce taux de porteur peut être restrictif et nous pouvons manquer des variants qui sont bien fondateurs mais dont la fréquence est inférieure à cause de la population plus petite de celle-ci par rapport au reste du Québec.

De plus, la formation des groupements d'individus selon leur proximité génétique peut présenter des limites. En effet, comme indiqué dans le chapitre 2, des individus ont dû être retirés des analyses car ils n'aboutissent pas aux mêmes résultats selon les données WGS ou les

données imputées. Ainsi, afin d'éviter tout biais, nous avons décidé de les retirer. Ce problème rencontré met en avant les limites de cette méthode de formation de groupements.

Enfin, une dernière limitation est l'usage de la base de données ClinVar pour la détermination de la pathogénicité de chaque variant dans notre étude. En effet, cette base de données étant une plateforme en ligne, elle s'actualise régulièrement et des variants que nous ne considérions pas comme pathogènes lors de notre étude pourront voir leur statut changer et être considérés par la suite comme bien pathogènes.

3. Perspectives

En ce qui concerne les perspectives de ce projet, celles-ci ont été abordées brièvement lors du chapitre 2 au niveau de la discussion de l'article. En effet, il a été introduit le fait que 3 variants peuvent être inclus dans les tests de porteurs proposés au SLSJ qui comptent aujourd'hui seulement 4 maladies explicitées dans le tableau 1. Ainsi, cette méthode peut mener à la découverte de nouvelles maladies génétiques rares qui sont importantes à prendre en compte et qui rentrent dans les critères pour être inclus dans les tests de porteurs. De plus, ici l'étude a été menée sur la province de Québec seulement. Or, il existe d'autres populations connues pour lesquelles l'effet fondateur a pu jouer sur la fréquence des variants et notamment augmenter la fréquence de variants délétères. C'est pourquoi, cette méthode peut être appliquée dans d'autres régions du monde et ainsi participer à une meilleure détection des maladies génétiques rares. Le but étant de faciliter les prises en charge des patients, sachant que les maladies rares sont difficiles à diagnostiquer, et ainsi fournir le meilleur traitement possible aux individus atteints.

Conclusion

Pour conclure, le travail effectué lors de ma maîtrise dans le laboratoire Genopop a permis de développer une méthode analytique mettant en évidence la présence de maladies génétiques rares en lien avec l'effet fondateur. Ainsi, en se basant sur des outils contemporains d'analyse génétique, notamment les segments identiques-par-descendance, il a été possible de mettre en avant des variants pathogènes dont la fréquence a augmenté tout au long du peuplement de la région et cela à cause de l'effet fondateur. Ainsi, un variant a été caractérisé de fondateur lorsque l'on pouvait constater une proportion de partage de ces segments au niveau du variant de plus de 50% parmi les individus porteurs de celui-ci. De plus, il a été établi un taux de porteurs limite, soit un nombre minimum de personnes porteuses du variant afin de le considérer dans notre étude. En faisant cela, nous avons pu mettre en avant des variants pour lesquels de nombreux individus partageaient des segments IBD au niveau du variant et traduisant l'origine ancestrale commune de ce segment pour ces individus dont la parenté est faible. Ceci reflète l'effet fondateur. Jusqu'à aujourd'hui, la détermination de maladies génétiques rares au SLSJ reposait sur le travail de diagnostic des cliniciens face aux patients atteints de symptômes. Or, comme dit précédemment, les maladies génétiques rares, par leur faible apparition sont difficiles à caractériser et un patient peut attendre de nombreuses années avant d'avoir un diagnostic précis. La nouvelle méthode que nous mettons en avant ici permet d'apporter une approche complémentaire et d'avoir une vision différente sur les possibles maladies présentes dans une région. Ainsi, lors de notre étude, nous avons retrouvé des variants déjà connus comme plus fréquents au SLSJ grâce au travail des cliniciens. Cela permet d'avoir une validation de notre méthode. De plus, nous avons mis en avant des variants qui n'avaient jamais été répertoriés avant dans la province de Québec et qui répondaient aux critères que nous avons fixés pour les caractériser de fondateurs. Parmi ces nouveaux variants, certains répondent aux conditions fixées par les cliniciens pour être présents dans les tests de porteurs

mis à disposition de la population. On peut donc constater que cette méthode a permis de porter un nouveau regard sur les maladies génétiques dans la région du SLSJ.

Bibliographie

(1) Griffiths, A. J. F.; Wessler, S. R.; Carroll, S. B.; Doebley, J. F.; Sanlaville, C.; Charmot-Bensimon, D. *Introduction à l'analyse génétique*, 6e édition.; De Boeck Université: Bruxelles, 2013.

(2) Alberts, B.; Johnson, A.; Lewis, J.; Morgan, D.; Raff, M. C.; Roberts, K.; Walter, P.; Darmon, M. *Biologie moléculaire de la cellule*, Sixième édition.; Lavoisier médecine-sciences: Paris, 2017.

(3) Alberts, B.; Hopkin, K.; Johnson, A.; Morgan, D. O.; Raff, M. C.; Roberts, K.; Walter, P. *Essential Cell Biology*, Fifth edition.; W.W. Norton & compagny: New York, 2019.

(4) Strachan, T.; Read, A. P. *Human Molecular Genetics*, 4th ed.; Garland Science: New York, 2011.

(5) Barrette, C. *Le miroir du monde: évolution par sélection naturelle et mystère de la nature humaine*; Éditions MultiMondes: Sainte-Foy, Québec, 2000.

(6) Darwin, C. *On the Origin of Species by Means of Natural Selection, or, The Preservation of Favoured Races in the Struggle for Life*, 4th impression.; World's classics; 11; no. 9-91021; H. Frowde, Oxford University Press: London, 1907.

(7) Haga, S. B. *The Book of Genes and Genomes*; Springer: New York, NY, UNITED STATES, 2022.

(8) Caporale, L. H. *The Implicit Genome*; Oxford University Press, Incorporated: Oxford, UNITED STATES, 2006.

(9) *19.5 Mutations and Genetic Diseases | The Basics of General, Organic, and Biological Chemistry*. https://courses.lumenlearning.com/suny-orgbiochemistry/chapter/19-5-mutations-and-genetic-diseases/ (accessed 2024-10-16).

(10) Passarge, E. *Atlas de poche de génétique*, 2e éd. complétée et mise à jour.; Atlas de poche; Flammarion: Paris, 2002.

(11) Erhabor, O. *Sickle Cell Disease*; IntechOpen: London, 2022.

(12) Hammer, G. D.; McPhee, S. J. *Pathophysiology of Disease: An Introduction to Clinical Medicine*, Seventh edition.; Lange medical book; McGraw-Hill Education Medical: New York, 2014.

(13) http://ec.europa.eu/health/ph_threats/non_com/rare_diseases_fr.

(14) Commissioner, O. of the. *Rare Diseases at FDA*. FDA. https://www.fda.gov/patients/rare-diseases-fda (accessed 2024-06-25).

(15) Wu, Z. H. *Rare Diseases*; IntechOpen: London, 2020.

(16) Hartl, D. L.; Clark, A. G. *Principles of Population Genetics*, 3rd ed.; Sinauer Associates: Sunderland, 1997.

(17) Goswami, C.; Chattopadhyay, A.; Chuang, E. Y. Rare Variants: Data Types and Analysis Strategies. *Ann Transl Med* **2021**, *9* (12), 961–961. https://doi.org/10.21037/atm-21-1635.

(18) Momozawa, Y.; Mizukami, K. Unique Roles of Rare Variants in the Genetics of Complex Diseases in Humans. *J Hum Genet* **2021**, *66* (1), 11–23. https://doi.org/10.1038/s10038-020-00845-2.

(19) Bartee, L.; Shriner, W.; Creech, C. *Principles of Biology*; Open Oregon Educational Resources, 2017.

(20) Génermont, J. Chapitre 3. Divisions cellulaires, réduction chromatique et seconde fonction universelle de la sexualité. *Sciences philosophie* **2014**, 73–86.

(21) Hunter, N. Meiotic Recombination: The Essence of Heredity. *Cold Spring Harb Perspect Biol* **2015**, a016618. https://doi.org/10.1101/cshperspect.a016618.

(22) Browning, B. L.; Browning, S. R. Improving the Accuracy and Efficiency of Identity-by-Descent Detection in Population Data. *Genetics* **2013**, *194* (2), 459–471. https://doi.org/10.1534/genetics.113.150029.

(23) Géron, A. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 3rd ed.; O'Reilly Media, Incorporated: Zurich, 2022.

(24) McConville, R.; Santos-Rodríguez, R.; Piechocki, R. J.; Craddock, I. N2D: (Not Too) Deep Clustering via Clustering the Local Manifold of an Autoencoded Embedding. In *2020 25th International Conference on Pattern Recognition (ICPR)*; 2021; pp 5145–5152. https://doi.org/10.1109/ICPR48806.2021.9413131.

(25) Diaz-Papkovich, A.; Anderson-Trocmé, L.; Gravel, S. A Review of UMAP in Population Genetics. *J Hum Genet* **2021**, *66* (1), 85–91. https://doi.org/10.1038/s10038-020-00851-4.

(26) Charbonneau, H.; Scholars Portal. *Naissance D'une Population: Les Français établis au Canada au XVIIe siècle*, Nouvelle édition revue et corrigée.; Ined Éditions ; Les Presses de l'Université de Montréal: Aubervilliers], [Montréal, Québec, 2020.

(27) Girard, C.; Perron, N.; Institut québécois de recherche sur la culture. *Histoire du Saguenay-Lac-Saint-Jean*; Collection Les régions du Québec; Institut québécois de recherche sur la culture: Québec, 1995.

(28) Charbonneau, H.; Desjardins, B.; Légaré, J.; Denis, H.; Centre interuniversitaire d'études québécoises. *La population française de la vallée du Saint-Laurent avant 1760*; Atlas historique du Québec; [Presses de l'Université Laval]: Sainte-Foy, 1996.

(29) Bouchard, G.; De Braekeleer, M. *Histoire d'un génôme: population et génétique dans l'est du Québec*; Presses de l'Université du Québec: Sillery/Qué., 1991.

(30) Haines, M. R.; Steckel, R. H. *A Population History of North America*; Cambridge University Press: Cambridge, Angleterre, 2000.

(31) Tremblay, V.; Société historique du Saguenay. *Histoire du Saguenay depuis les origines jusqu'a 1870*, 4e ed.; Publications de la Société historique du Saguenay; Société historique du Saguenay: Chicoutimi, 1984.

(32) Pouyez, C.; Lavoie, Y.; Bouchard, G. *Les Saguenayens: Introduction à l'histoire Des Populations Du Saguenay, XVIe-XXe Siècles*; Presses de l'Université du Québec: Sillery, Québec, 1983.

(33) Gagnon, L.; Moreau, C.; Laprise, C.; Vézina, H.; Girard, S. L. Deciphering the Genetic Structure of the Quebec Founder Population Using Genealogies. *Eur J Hum Genet* **2023**, 1–7. https://doi.org/10.1038/s41431-023-01356-2.

(34) 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* (2), 67–83. https://doi.org/10.3917/adh.108.0067.

(35) Heyer, E. One Founder/One Gene Hypothesis in a New Expanding Population: Saguenay (Quebec, Canada). *Hum Biol* **1999**, *71* (1), 99–109.

(36) Heyer, E.; Austerlitz, F. Update to Heyer's "One Founder/One Gene Hypothesis in a New Expanding Population" (1999). *Hum Biol* **2009**, *81* (5–6), 657–662. https://doi.org/10.3378/027.081.0614.

(37) Moreau, C.; Bhérer, C.; Vézina, H.; Jomphe, M.; Labuda, D.; Excoffier, L. Deep Human Genealogies Reveal a Selective Advantage to Be on an Expanding Wave Front. *Science* **2011**, *334* (6059), 1148–1150. https://doi.org/10.1126/science.1212880.

(38) Casals, F.; Hodgkinson, A.; Hussin, J.; Idaghdour, Y.; Bruat, V.; De Maillard, T.; Grenier, J.-C.; Gbeha, E.; Hamdan, F. F.; Girard, S.; Spinella, J.-F.; Larivière, M.; Saillour, V.; Healy, J.; Fernández, I.; Sinnett, D.; Michaud, J. L.; Rouleau, G. A.; Haddad, E.; Le Deist, F.; Awadalla, P. Whole-Exome Sequencing Reveals a Rapid Change in the Frequency of Rare Functional Variants in a Founding Population of Humans. *PLoS Genet* **2013**, *9* (9), e1003815. https://doi.org/10.1371/journal.pgen.1003815.

(39) Bchetnia, M.; Bouchard, L.; Mathieu, J.; Campeau, P. M.; Morin, C.; Brisson, D.; Laberge, A.-M.; Vézina, H.; Gaudet, D.; Laprise, C. Genetic Burden Linked to Founder Effects in Saguenay–Lac-Saint-Jean Illustrates the Importance of Genetic Screening Test Availability. *J Med Genet* **2021**, *58* (10), 653–665. https://doi.org/10.1136/jmedgenet-2021-107809.

(40) Laberge, A.-M.; Michaud, J.; Richter, A.; Lemyre, E.; Lambert, M.; Brais, B.; Mitchell, G. Population History and Its Impact on Medical Genetics in Quebec. *Clinical Genetics* **2005**, *68* (4), 287–301. https://doi.org/10.1111/j.1399-0004.2005.00497.x.

(41) Scriver, C. R. Human Genetics: Lessons from Quebec Populations. *Annu Rev Genomics Hum Genet* **2001**, *2*, 69–101. https://doi.org/10.1146/annurev.genom.2.1.69.

(42) Cruz Marino, T.; Leblanc, J.; Pratte, A.; Tardif, J.; Thomas, M.-J.; Fortin, C.-A.; Girard, L.; Bouchard, L. Portrait of Autosomal Recessive Diseases in the French-Canadian Founder Population of Saguenay-Lac-Saint-Jean. *American Journal of Medical Genetics Part A* **2023**, *191* (5), 1145–1163. https://doi.org/10.1002/ajmg.a.63147.

(43) De Braekeleer, M. Hereditary Disorders in Saguenay-Lac-St-Jean (Quebec, Canada). *Hum Hered* **1991**, *41* (3), 141–146. https://doi.org/10.1159/000153992.

(44) Primrose, S. B.; Twyman, R. M. *Principles of Genome Analysis and Genomics*, 3th ed.; Blackwell Pub.: Malden, Mass., 2003.

(45) Davey, J. W.; Hohenlohe, P. A.; Etter, P. D.; Boone, J. Q.; Catchen, J. M.; Blaxter, M. L. Genome-Wide Genetic Marker Discovery and Genotyping Using next-Generation Sequencing. *Nat Rev Genet* **2011**, *12* (7), 499–510. https://doi.org/10.1038/nrg3012.

(46) Li, Y.; Willer, C.; Sanna, S.; Abecasis, G. Genotype Imputation. *Annu. Rev. Genom. Hum. Genet.* **2009**, *10* (1), 387–406. https://doi.org/10.1146/annurev.genom.9.081307.164242.

(47) Marchini, J.; Howie, B. Genotype Imputation for Genome-Wide Association Studies. *Nat Rev Genet* **2010**, *11* (7), 499–511. https://doi.org/10.1038/nrg2796.

(48) Piovesan, A.; Pelleri, M. C.; Antonaros, F.; Strippoli, P.; Caracausi, M.; Vitale, L. On the Length, Weight and GC Content of the Human Genome. *BMC Res Notes* **2019**, *12* (1), 106. https://doi.org/10.1186/s13104-019-4137-z.

(49) Tam, V.; Patel, N.; Turcotte, M.; Bossé, Y.; Paré, G.; Meyre, D. Benefits and Limitations of Genome-Wide Association Studies. *Nat Rev Genet* **2019**, *20* (8), 467–484. https://doi.org/10.1038/s41576-019-0127-1.

(50) *Accueil | CARTaGENE*. https://cartagene.qc.ca/ (accessed 2024-01-16).

(51) Info_GeneticData3juillet2023.Pdf. https://cartagene.qc.ca/files/documents/other/Info_GeneticData3juillet2023.pdf (accessed 2024-03-27).

(52) Nait Saada, J.; Kalantzis, G.; Shyr, D.; Cooper, F.; Robinson, M.; Gusev, A.; Palamara, P. F. Identity-by-Descent Detection across 487,409 British Samples Reveals Fine Scale Population Structure and Ultra-Rare Variant Associations. *Nat Commun* **2020**, *11* (1), 6130. https://doi.org/10.1038/s41467-020-19588-x.

(53) Gagnon, L.; Moreau, C.; Laprise, C.; Girard, S. L. Fine-Scale Genetic Structure and Rare Variant Frequencies. February 7, 2024. https://doi.org/10.1101/2024.02.02.578687.

(54) Nguengang Wakap, S.; Lambert, D. M.; Olry, A.; Rodwell, C.; Gueydan, C.; Lanneau, V.;

Murphy, D.; Le Cam, Y.; Rath, A. Estimating Cumulative Point Prevalence of Rare Diseases: Analysis of the Orphanet Database. *Eur J Hum Genet* **2020**, *28* (2), 165–173. https://doi.org/10.1038/s41431-019-0508-0.

Certification éthique

Ce mémoire a fait l'objet d'une certification éthique au CER de l'UQAC. Son numéro de certificat est le suivant : 2021-560 - Étude génétique des fondateurs et fondatrices des populations d'origine européenne.