UQAC
Université du Québec
à Chicoutimi

Déceler la structure fine de la population du Saguenay–Lac-Saint-Jean par la génétique et la généalogie

par Gilles-Philippe Morin

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

Saguenay, Canada

Résumé

On parle de structure génétique fine lorsque la distribution des origines ancestrales et des variants génétiques diffère au sein d'une population. Lorsqu'une telle structure corrèle avec un trait phénotypique, elle pose des défis quant aux études d'association pangénomique (GWAS) en introduisant des facteurs de confusion que les méthodes de correction standard peuvent ne pas corriger complètement. Ce mémoire vise à caractériser la structure génétique fine de la population du Saguenay–Lac-Saint-Jean (SLSJ), une région à effet fondateur souvent présumée « homogène ».

En intégrant les données génotypiques de la cohorte CARTaGENE aux registres généalogiques du fichier de population BALSAC, nous démontrons une forte concordance entre la parenté réalisée (génétique) et attendue (généalogique), avec des coefficients de corrélation de Pearson de 0,78 pour l'ensemble de la cohorte et de 0,83 pour le sous-ensemble d'individus originaires du SLSJ.

Un algorithme hybride combinant les approches de Karigl et de Kirkpatrick a été développé et implémenté dans la bibliothèque GeneaKit, permettant le calcul de plus de 3 milliards de coefficients de parenté en seulement deux minutes. Cette avancée méthodologique a rendu possible l'analyse de l'ensemble de la population du SLSJ en s'appuyant sur 26 445 proposants issus de familles distinctes, mariés dans la région entre 1935 et 1960 et avec des généalogies suffisamment complètes.

Nos résultats révèlent une structure fine détectable jusqu'au niveau municipal, suivant un gradient est-ouest façonné par les contributions différentielles des fondateurs charlevoisiens. La Malbaie et Les Éboulements ont principalement contribué au Saguenay, tandis que Baie-Saint-Paul a davantage contribué au Lac-Saint-Jean. Les centres urbains au sud de la rivière Saguenay présentent une haute diversité.

Ces découvertes remettent en question le paradigme d'homogénéité régionale et suggèrent qu'une structure fine similaire pourrait exister dans de nombreuses autres populations humaines, particulièrement celles ayant connu des histoires démographiques comparables. Ces résultats ont des retombées importantes pour la médecine de précision, les scores de risque polygénique et les programmes de dépistage génétique.

Mots-clés : structure fine de population, effet fondateur, génétique des populations, généalogie, coefficient de parenté, Saguenay–Lac-Saint-Jean, GWAS, scores de risque polygénique

Abstract

Fine-scale genetic structure occurs when the distribution of ancestry and genetic variants differs in a population. When such a structure correlates with a phenotypic trait, it poses challenges for genome-wide association studies (GWAS) by introducing confounding factors that standard correction methods may not fully address. This thesis aims to characterize the fine-scale genetic structure of the Saguenay–Lac-Saint-Jean (SLSJ) population, a founder population often presumed to be "homogeneous".

By integrating genotype data from the CARTaGENE cohort with genealogical records from the BALSAC population register, we demonstrate strong concordance between realised (genetic) and expected (genealogical) kinship, with Pearson correlation coefficients of 0.78 for the entire cohort and 0.83 for the SLSJ-origin subset.

A hybrid algorithm combining the Karigl and Kirkpatrick approaches was developed and implemented in the GeneaKit library, enabling the computation of over 3 billion kinship coefficients in just two minutes. This methodological advance made it possible to analyze the entire SLSJ population, based on 26,445 probands from distinct families, married in the region between 1935 and 1960, and with sufficiently complete genealogies.

Our results reveal fine-scale structure detectable down to the municipal level, following an east–west gradient shaped by differential contributions from Charlevoix founders. La Malbaie and Les Éboulements primarily contributed to the Saguenay area, while Baie-Saint-Paul had greater contribution on Lac-Saint-Jean. Urban centres south of the Saguenay River display high diversity.

These findings challenge the paradigm of regional homogeneity and suggest that similar fine-scale structure may exist in many other human populations, particularly those with comparable demographic histories. These results have important implications for precision medicine, polygenic risk scores, and genetic screening programs.

Keywords: fine-scale population structure, founder effect, population genetics, genealogy, kinship coefficient, Saguenay–Lac-Saint-Jean, GWAS, polygenic risk scores

Table des matières

Liste des figures

Liste des abréviations

Abréviation	**Définition**
ADN	acide désoxyribonucléique
ACP	analyse en composantes principales (*Principal Component Analysis*, PCA)
ARN	acide ribonucléique
ARSACS	ataxie récessive spastique de Charlevoix-Saguenay
cM	centimorgan
CaG	CARTaGENE
CNV	variation du nombre de copies (*Copy Number Variation*)
f	coefficient de consanguinité
F_{ST}	index de fixation
GWAS	étude d'association pangénomique (*Genome-Wide Association Study*)
HDBSCAN	*Hierarchical Density-Based Spatial Clustering of Applications with Noise*
IBD	identique par descendance (*Identical by Descent*)
IBS	identique par état (*Identical by State*)
LD	déséquilibre de liaison (*Linkage Disequilibrium*)
MDS	positionnement multidimensionnel (*Multidimensional Scaling*)
PCoA	analyse en coordonnées principales (*Principal Coordinates Analysis*)
PRS	score de risque polygénique (*Polygenic Risk Score*)
SLSJ	Saguenay–Lac-Saint-Jean
SNP	polymorphisme d'un seul nucléotide (*Single Nucleotide Polymorphism*)
UMAP	*Uniform Manifold Approximation and Projection*
WGS	séquençage du génome entier (*Whole Genome Sequencing*)
φ	coefficient de parenté (phi)

Remerciements

Je tiens d'abord à exprimer ma profonde gratitude envers mon directeur de recherche, le professeur Simon Girard, titulaire de la Chaire de recherche du Canada en génétique et généalogie. Son encadrement débordant d'enthousiasme et sa confiance en mes capacités ont été déterminants tout au long de ce projet. Je le remercie de m'avoir accueilli au sein du laboratoire Genopop et de m'avoir permis de m'épanouir dans un domaine à l'intersection de mes intérêts.

Je souhaite également remercier mon codirecteur, le professeur Amadou Barry, pour ses conseils avisés en biostatistique et son regard complémentaire sur mes travaux. Sa rigueur méthodologique a grandement enrichi cette recherche.

Ma reconnaissance va évidemment à Claudia Moreau, dont l'expertise, la patience infatigable face à mes nombreuses questions et à mes milliers de messages sur Slack ont été inestimables. Ses commentaires toujours pertinents et ses remises en question constructives m'ont poussé à approfondir mes réflexions, à améliorer continuellement la qualité de mes analyses et à pousser toujours plus loin nos recherches.

Je remercie sincèrement l'équipe du projet BALSAC pour l'accès aux données généalogiques qui constituent le fondement de cette étude, ainsi que la cohorte CARTaGENE et ses participants pour les données génotypiques essentielles à la validation de nos résultats. Sans ces ressources uniques au monde, ce projet n'aurait pas été possible.

Mes remerciements s'adressent également à mes collègues du laboratoire Genopop, pour les échanges scientifiques (et d'autres beaucoup moins scientifiques) stimulants, l'entraide quotidienne et les moments inoubliables qui ont rendu ces années de recherche aussi enrichissantes sur le plan humain que scientifique. Un merci particulier à Joanie pour avoir réorienté mon parcours en me faisant découvrir le labo de Simon !

Je tiens à souligner le soutien financier de l'Unité mixte de recherche INRS-UQAC en santé durable, qui a rendu possible la réalisation de ce projet. Je remercie également l'Alliance de recherche numérique du Canada pour l'accès aux ressources de calcul haute performance indispensables au traitement de données à grande échelle.

Enfin, je souhaite remercier ma mère pour son soutien indéfectible et ses encouragements constants tout au long de ce parcours, et mon père, qui partageait cet intérêt pour la généalogie et l'histoire de notre région et de nos ancêtres.

Avant-propos

Ce mémoire s'inscrit dans le cadre d'une maitrise en santé durable à l'Université du Québec à Chicoutimi. Il porte sur la caractérisation de la structure génétique fine de la population du Saguenay–Lac-Saint-Jean, une région à effet fondateur trop souvent présumée génétiquement homogène.

Structure du mémoire

Le présent mémoire est organisé en quatre chapitres. Le **Chapitre 1** établit le cadre théorique nécessaire à la compréhension de l'étude. Il aborde les concepts fondamentaux de la génétique des populations, les défis posés par la stratification dans les études d'association pangénomique, le contexte démographique et génétique du Québec, ainsi que les innovations algorithmiques développées pour surmonter les contraintes computationnelles inhérentes à l'analyse de données à grande échelle. Le **Chapitre 2** présente l'article scientifique intitulé « *Fine-scale structure of a whole regional population through genetics and genealogies* », qui est présentement en cours de publication dans la revue *Nature Communications*. Cet article constitue le cœur du mémoire et présente les résultats principaux de la recherche. Le **Chapitre 3** propose une discussion élargie des résultats, examine leurs retombées pour la santé publique et la médecine de précision, identifie les limites de l'étude et ouvre des perspectives pour la recherche future. Enfin, le **Chapitre 4** conclut le mémoire par une synthèse des contributions principales.

Contribution de l'auteur

L'auteur de ce mémoire a réalisé l'ensemble des analyses présentées dans l'article scientifique, incluant le développement et l'implémentation de l'algorithme hybride de calcul des coefficients de parenté au sein de la bibliothèque **GeneaKit**. Ce développement algorithmique permet le calcul de plus de trois milliards de coefficients de parenté en quelques minutes, là où les méthodes existantes auraient nécessité des ressources inaccessibles.

L'auteur a également conçu et réalisé les visualisations par UMAP, effectué les analyses de contribution génétique attendue des fondateurs, et rédigé le manuscrit. Les coauteurs ont contribué à l'interprétation des résultats, à la révision du manuscrit et à la supervision générale du projet.

Statut de l'article

L'article « *Fine-scale structure of a whole regional population through genetics and genealogies* » est en cours de publication dans la revue *Nature Communications*. Au moment du dépôt de ce mémoire, l'article est disponible sous licence *Creative Commons*

Attribution - Utilisation non commerciale - Pas d'Œuvre dérivée 4.0 International (CC BY-NC-ND 4.0) à l'adresse suivante : https://www.nature.com/articles/s41467-026-70175-y.

Science ouverte

Dans un souci de reproductibilité et de science ouverte, la bibliothèque GeneaKit est librement disponible sous licence MIT sur GitHub (https://github.com/Genopop/geneakit). Le code d'analyse utilisé pour cette étude est également accessible publiquement sous licence MIT (https://github.com/Genopop/slsj_structure).

Chapitre 1 : Introduction

1.1 Les fondements de la génétique

1.1.1 L'ADN

L'**acide désoxyribonucléique** (ADN) encode l'information génétique chez tous les êtres vivants. Cette molécule est composée de quatre nucléotides distincts : l'adénine (A), la thymine (T), la cytosine (C) et la guanine (G). L'ADN adopte une structure en double hélice [1] où les deux brins sont maintenus ensemble par des liaisons hydrogène entre des paires de bases qui se complémentent : l'adénine se lie toujours avec la thymine, et la cytosine avec la guanine.

Lors de la division cellulaire, l'ADN est regroupé en **chromosomes**. Chez l'humain, le génome nucléaire est organisé en 23 paires de chromosomes [2] : 22 paires d'autosomes et une paire de chromosomes sexuels (XX chez la femme, XY chez l'homme). Pour chaque paire, un chromosome est hérité de la mère et l'autre du père. La mitochondrie, responsable entre autres de la respiration cellulaire, possède également son propre ADN (mitochondrial) qui est transmis exclusivement par voie maternelle.

1.1.2 Des gènes aux variants

Un **gène** correspond à une séquence de nucléotides qui code (par transcription) pour une molécule d'acide ribonucléique (ARN), dont certaines molécules vont ensuite produire une protéine (par traduction). Les copies d'un même gène, qui peuvent différer

dans leur séquence de nucléotides, sont appelées des **allèles**. Chaque individu possède deux copies de chaque gène autosomique (une héritée de chaque parent) et peut donc porter des allèles identiques (homozygote) ou différents (hétérozygote) à un locus donné.

La variation génétique peut prendre plusieurs formes. Les **polymorphismes d'un seul nucléotide** (*Single Nucleotide Polymorphisms,* SNPs) sont des substitutions ponctuelles où un nucléotide est remplacé par un autre. Habituellement, un variant est qualifié de « polymorphisme » lorsque dans la population il représente plus de 1 % des allèles au locus donné, alors qu'autrement il est qualifié de variant rare [3]. D'autres types de variations existent également, notamment les insertions et délétions de séquences (indels), les inversions et les variations du nombre de copies (*Copy Number Variants,* CNV) [3]. Ces variations trouvent leur origine dans des **mutations *de novo*** qui apparaissent spontanément chez un individu sans avoir été héritées d'aucun des deux parents. Ces mutations, lorsqu'elles surviennent dans la lignée germinale, peuvent être transmises aux générations suivantes et sont une source de diversité et de sélection [4].

1.1.3 La transmission du matériel génétique

La **méiose** est le processus de division cellulaire de la reproduction sexuée. Elle génère des gamètes (ovules ou spermatozoïdes) qui ne contiennent qu'un seul chromosome par paire. Deux mécanismes majeurs contribuent à la diversité génétique lors de ce processus.

La **recombinaison génétique** survient lorsque les chromosomes homologues échangent des segments d'ADN. Plus deux positions sont éloignées sur un chromosome, plus la probabilité augmente qu'une recombinaison survienne. Cette distance génétique se mesure en centimorgans (cM), où 1 cM correspond approximativement à une probabilité de 1 % de recombinaison par génération [5]. Cette distance correspond à peu près à un million de paires de bases de nucléotides [6]. L'**assortiment indépendant** fait en sorte que chaque gamète reçoit une combinaison aléatoire des chromosomes parentaux. Ainsi, bien qu'un enfant hérite de la moitié du matériel génétique de chaque parent, les segments qui sont transmis varient entre frères et sœurs.

1.1.4 Le déséquilibre de liaison

Lorsque deux variants génétiques situés à proximité sont transmis ensemble plus souvent que ne le ferait le hasard, on dit qu'ils sont en **liaison génétique** (*genetic linkage*) [7]. Ce phénomène vient du fait que la recombinaison se fait plus rare entre des positions qui sont proches [8]. À l'échelle de la population, lorsque des allèles à deux locus différents sont observés ensemble plus souvent que ce à quoi on s'attendrait, on dit qu'il y a **déséquilibre de liaison** (*linkage disequilibrium,* LD) [7]. Le déséquilibre de liaison est important en génétique des populations [9] notamment parce qu'il permet d'inférer, à partir d'un nombre limité de variants qu'on a génotypés, l'état de variants non directement observés par ce qu'on appelle l'**imputation** [10].

1.1.5 Identité par état et identité par descendance

Deux concepts permettent de caractériser la similarité génétique entre individus. On parle d'**identité par état** (*Identical by State,* IBS) lorsque deux segments d'ADN présentent la même séquence de nucléotides, peu importe leur origine. Cette identité peut résulter du hasard ou d'une ascendance très lointaine [11].

L'**identité par descendance** (*Identical by Descent,* IBD) s'observe quand deux segments d'ADN sont identiques parce qu'ils ont été hérités d'un ancêtre commun récent. L'IBD permet de retracer les liens familiaux et la structure généalogique récente d'une population [12]. La recombinaison fait en sorte qu'au fil des générations, les segments IBD qui sont partagés entre descendants d'un même ancêtre deviennent de plus en plus courts, ce qui permet d'estimer l'ancienneté du lien généalogique [13].

1.2 Les mécanismes d'évolution et la diversité génétique

La génétique des populations étudie la variation des fréquences alléliques au sein des populations et les mécanismes qui modifient cette variation au fil du temps et dans l'espace [14]. Quatre mécanismes majeurs façonnent cette évolution.

1.2.1 La sélection naturelle

La **sélection naturelle** se produit lorsqu'un allèle influence le potentiel reproductif des individus qui en sont porteurs. Un allèle qui confère un avantage en termes de survie ou de reproduction va avoir tendance à augmenter en fréquence dans la population,

tandis qu'un allèle délétère va tendre à diminuer de fréquence. Ce mécanisme constitue le moteur principal de l'adaptation des populations à leur environnement [15, 16].

1.2.2 La mutation

La **mutation** correspond à tout changement dans la séquence d'ADN d'un individu par rapport à celle de ses parents. Ces modifications peuvent survenir dans les cellules germinales (mutations transmissibles) ou somatiques (non transmissibles). Les mutations peuvent introduire de nouveaux allèles dans une population [17].

1.2.3 La migration ou flux génétique

La **migration**, ou flux génétique, survient lorsque des individus se déplacent d'une population à une autre, ce qui introduit de nouveaux allèles ou modifie les fréquences alléliques de la population d'accueil [18, 19].

1.2.4 La dérive génétique et l'effet fondateur

La **dérive génétique** correspond aux fluctuations aléatoires des fréquences alléliques d'une génération à l'autre, indépendamment des avantages ou désavantages sélectifs. Son impact est inversement proportionnel à la taille de la population : dans les petites populations, la dérive peut conduire à la fixation ou à la disparition d'allèles par pur hasard [16].

Un cas particulier de dérive génétique survient lors de l'**effet fondateur,** lorsqu'une nouvelle population est établie par un petit nombre d'individus [20]. Ces fondateurs emportent avec eux une fraction non représentative de la diversité génétique de la population source, ce qui peut conduire à l'enrichissement d'allèles rares, y compris ceux associés à des maladies héréditaires [21]. Similairement, un **goulot d'étranglement** (*bottleneck*) survient lorsqu'une population subit une réduction drastique de sa taille sans fonder une nouvelle population, par des évènements cataclysmiques tels qu'une épidémie ou une catastrophe naturelle, ce qui entraine une perte de diversité génétique [22].

1.3 Le génotypage et les études d'association pangénomique

1.3.1 Techniques

Le **génotypage** consiste à identifier les variants génétiques présents chez un individu. Les puces de génotypage permettent d'analyser simultanément jusqu'à plusieurs millions de SNPs répartis sur l'ensemble du génome [23]. Ces puces exploitent le déséquilibre de liaison en ciblant stratégiquement des SNPs qui permettent d'inférer, par imputation, l'état d'autres variants qui ne sont pas directement génotypés [24].

Le **séquençage du génome entier** (*Whole Genome Sequencing,* WGS) offre une approche exhaustive en déterminant la séquence de nucléotides de tout le génome. Cette technique permet d'identifier tous les variants, y compris les variants rares mal capturés par les puces de génotypage [25]. Toutefois, son cout plus élevé limite encore son application aux grandes cohortes.

1.3.2 Les études d'association pangénomique

Les **études d'association pangénomique** (*Genome-Wide Association Studies,* GWAS) sont particulièrement utiles pour comprendre la génétique des traits complexes, qui sont influencés par plusieurs gènes ou par une interaction avec l'environnement. Contrairement aux études qui testent des hypothèses préétablies sur des gènes spécifiques, les GWAS analysent des marqueurs distribués sur l'ensemble du génome sans apriori sur les régions potentiellement impliquées [26].

Le plan d'expériences de base compare les fréquences alléliques entre les cas (individus atteints d'une maladie) et les témoins (individus sains) pour les **traits binaires**, ou calcule la corrélation entre les génotypes et des mesures continues pour les **traits quantitatifs** [26]. L'analyse statistique procède par tests d'association variant par variant, typiquement par régression logistique pour les traits binaires ou par régression linéaire pour les traits quantitatifs [26]. Lors de ces analyses, il est crucial de prendre en compte des covariables telles que l'âge, le sexe et les origines ancestrales, car les résultats des GWAS y sont très sensibles [26].

1.4 Structure de population et stratification de population

1.4.1 Quelle est la différence entre les deux ?

La **structure de population** désigne les différences dans les fréquences alléliques entre des sous-populations, qui résultent de l'ascendance généalogique et de l'histoire démographique. Elle peut être quantifiée par l'indice de fixation (F_{ST}) [27], qui

mesure la proportion de la diversité génétique attribuable à la différenciation des populations [28, 29].

La **stratification de population** correspond à la corrélation qui peut se produire entre la structure de population et le phénotype à l'étude, ce qui provoque des associations trompeuses dans les études génétiques. La distinction clé est que si une structure de population existe, elle ne cause un biais de stratification que lorsque les cas et les témoins sont échantillonnés à partir de sous-populations qui diffèrent à la fois dans leurs fréquences alléliques et dans la prévalence du trait [30].

1.4.2 Exemples de confusion

Du côté théorique, on retrouve le scénario hypothétique du « gène des baguettes asiatiques » (*chopsticks gene*), décrit notamment par Hamer et Sirota [31]. Ce scénario invite à imaginer une étude de variants génétiques associés à l'utilisation de baguettes dans un échantillon qui mélange des individus d'ascendance est-asiatique et européenne. Les SNPs qui présentent des fréquences différentes entre les deux groupes apparaitraient associés à l'utilisation des baguettes parce que l'ascendance et les différences culturelles confondent l'analyse. Ces mêmes auteurs soulignent que cette confusion peut autant provoquer des faux positifs que des faux négatifs [31].

Du côté pratique, un exemple réel provient de Knowler et al. [32], qui ont rapporté une association entre un haplotype (c'est-à-dire un ensemble de variants sur un seul

chromosome) et le diabète de type 2 chez les peuples autochtones Pima et Papago. Les auteurs ont démontré que l'association était entièrement fausse, due à une confusion par le mélange entre ascendances autochtone et européenne. Après un ajustement pour les proportions de mélange d'ascendances, le signal disparait complètement [33].

1.4.3 Impact sur les scores de risque polygénique

Les **scores de risque polygénique** (*Polygenic Risk Scores,* PRS) cumulent les effets de milliers de variants identifiés par GWAS pour prédire le risque de présenter une maladie ou la valeur d'un trait quantitatif [34]. Ces scores sont particulièrement vulnérables aux biais de stratification : même de petits biais par variant peuvent s'accumuler en erreurs considérables lorsqu'ils sont sommés sur l'ensemble du génome [35].

Les travaux de Sohail et al. [36] et Berg et al. [37] ont démontré que des conclusions initiales suggérant une adaptation polygénique sur la taille humaine à travers les populations européennes étaient largement attribuables à une stratification non contrôlée. En parallèle, Haworth et al. [38] ont montré que la structure géographique demeure détectable dans une grande cohorte homogène même après ajustement pour le centre d'étude et les composantes génétiques principales (voir ci-dessous), ce qui soulève des questions sur la validité des comparaisons de PRS entre populations.

1.4.4 Méthodes de correction

Plusieurs approches permettent de contrôler la stratification de population dans les GWAS :

Le **contrôle génomique** (*genomic control*), développé par Devlin et Roeder [39], calcule un facteur d'inflation (λ) à partir de SNPs qui ont un effet nul sur le phénotype (c'est-à-dire les manifestations observables ou mesurables) et ajuste les statistiques de test en conséquence.

L'**analyse en composantes principales** (ACP), implémentée notamment dans le logiciel EIGENSTRAT [40, 41], identifie les axes de variation génétique (vecteurs propres) et les inclut comme covariables dans les modèles d'association. Cette approche demeure la plus répandue, incluant typiquement les 10 à 20 premières composantes principales. Toutefois, comme l'ont démontré Zaidi et Mathieson [42], l'impact de la stratification est également façonné par l'histoire démographique, un facteur souvent négligé même après ajustement pour 100 composantes principales de variants communs.

Les **études de type familial,** comme le test de déséquilibre de transmission (*Transmission Disequilibrium Test,* TDT) [43], évitent complètement le problème de la stratification en n'utilisant que la variation génétique à l'intérieur même de la famille, au détriment d'une moins grande puissance [26], soit une moins bonne capacité à identifier un vrai signal.

1.5 La généalogie

1.5.1 Fondements de la généalogie

La **généalogie** offre une perspective historique complémentaire aux données génétiques, permettant d'étendre l'analyse à des populations entières et sur plusieurs siècles. Un arbre généalogique représente graphiquement les liens de filiation, avec les ancêtres en haut et les descendants en bas, une génération par rangée.

À chaque génération, le nombre d'ancêtres **double théoriquement**. Ainsi, si un individu est seul à la génération 0 (2^0), à la génération 40, soit environ 1 200 ans d'histoire démographique à raison de 30 ans par génération, on compterait théoriquement plus de 1 000 milliards (2^{40}) d'ancêtres, soit un nombre supérieur au nombre d'êtres humains ayant vécu sur Terre [44]. Cette impossibilité mathématique implique que l'arbre se referme nécessairement : les mêmes ancêtres apparaissent plusieurs fois dans une généalogie [45].

1.5.2 Consanguinité et parenté

Lorsqu'un même individu apparait dans les ancêtres des deux parents d'une personne, on parle de **consanguinité.** Le coefficient de consanguinité (*f*) [46] mesure la probabilité qu'un individu ait hérité de deux copies identiques par descendance d'un même allèle ancestral. Plus les parents d'une personne partagent d'ancêtres en commun et plus ces ancêtres sont récents, plus le coefficient de consanguinité est élevé.

Lorsqu'un même ancêtre apparait dans les origines de deux individus distincts, on parle de **parenté.** Le coefficient de parenté (φ), formalisé par Wright [46] et développé par Malécot [47], correspond à la probabilité qu'un allèle tiré au hasard chez un individu *i* soit identique par descendance à un allèle tiré au même locus chez un individu *j*. Plus deux personnes ont d'ancêtres en commun et plus ces ancêtres sont récents, plus le coefficient de parenté est élevé. Les valeurs théoriques incluent $\varphi = 0{,}25$ pour des parents-enfants ou des frères et sœurs, $\varphi = 0{,}125$ pour des demi-frères et demi-sœurs, et $\varphi = 0{,}0625$ pour des cousins et cousines.

Un **proposant** (ou probant) désigne un individu qui est tout en bas de la généalogie à l'étude, tandis qu'un **fondateur ou** une **fondatrice** désigne un individu dont aucun parent n'est connu dans cette même généalogie. En génétique médicale, les proposants sont aussi qualifiés de **cas index**, ce qui désigne les individus atteints d'une affection qui sont à l'origine d'une enquête génétique [48]. La **complétude généalogique** mesure, pour une génération donnée, la proportion d'ancêtres connus par rapport au nombre d'ancêtres attendus [49]. Cette complétude constitue la principale limitation des études généalogiques, diminuant inévitablement avec la profondeur de la génération.

1.5.3 Concordance entre généalogie et génétique

Les données généalogiques permettent de calculer la **parenté attendue** basée sur les relations familiales documentées. Les données génomiques permettent quant à elles d'estimer la **parenté réalisée** à partir des segments IBD effectivement partagés. La concor-

dance entre ces deux mesures valide l'utilisation des généalogies comme indicateur fiable de la structure génétique.

Les données sur la population québécoise d'origine canadienne-française ont permis des **comparaisons directes** entre parenté généalogique et génétique [50-54]. En particulier, Gauvin et al. [51] ont étudié un petit échantillon avec des généalogies s'étendant jusqu'aux fondateurs et fondatrices du XVII^{e} siècle, combinées à des génotypes. Ils ont démontré que la longueur totale des segments IBD partagés expliquait 85 % de la variance des coefficients de parenté généalogique, ce qui montre une forte correspondance entre les prédictions basées sur les généalogies et le partage génétique réalisé.

1.6 Mesurer la parenté

1.6.1 Le problème de l'échelle

Bien que la théorie du coefficient de parenté soit bien établie, son application à une population entière pose un défi majeur. Pour N proposants, le calcul de la matrice de parenté complète nécessite d'évaluer $\boldsymbol{N(N-1)/2}$ **paires** [13], sans compter toutes les paires ancestrales dont elles sont dérivées. Pour une génération complète de plusieurs dizaines de milliers d'individus, ce calcul implique des milliards de paires, ce qui rend les algorithmes standards inapplicables.

1.6.2 Algorithmes existants

Plusieurs approches algorithmiques ont été développées pour calculer le coefficient de parenté sur des généalogies :

La **méthode de Karigl** [55] est une approche récursive qui calcule le coefficient de parenté paire par paire. Elle est exacte et facilement parallélisable, ce qui permet le calcul simultané de plusieurs paires. Toutefois, elle est lente car on doit remonter l'arbre généalogique pour chaque ancêtre commun, en recalculant les mêmes ancêtres à chaque apparition. Cette implémentation est disponible dans le progiciel GENLIB pour R [56].

La **méthode de Kirkpatrick** et al. [57] analyse la généalogie génération par génération. Elle est plus rapide, car elle évite la redondance des calculs en divisant la généalogie en sous-généalogies successives. Les proposants d'une sous-généalogie sont traités comme les fondateurs de la suivante, ce qui minimise ainsi le recalcul des ancêtres. Cependant, cette approche n'est pas conçue pour l'exécution en parallèle.

La **méthode indirecte de Colleau** [58] peut également être appliquée pour calculer les coefficients de parenté à grande échelle [59], mais elle incorpore toutes les générations de la généalogie, ce qui peut ne pas correspondre aux objectifs d'analyse. L'auteur a été informé de l'existence de cette méthode suite à la prépublication de l'article et cet algorithme n'a pas fait l'objet d'une comparaison exhaustive de sa performance par rapport à l'algorithme hybride développé dans le cadre de ce projet. Toutefois, il a été confirmé que

les coefficients produits par les deux algorithmes sont équivalents. La performance satisfaisante de l'algorithme hybride n'a pas justifié de revisiter le problème en profondeur à l'aide de la méthode indirecte de Colleau.

1.7 Visualisation de la structure de population

Une fois les matrices de parenté calculées, la visualisation de la structure de population requiert des techniques de réduction de dimension capables de révéler les structures complexes présentes dans ces données de haute dimension.

1.7.1 Méthodes linéaires : ACP et PCoA

L'**analyse en composantes principales** (ACP) est la méthode standard de réduction de dimension. Elle projette les données sur les axes de variance maximale (vecteurs propres), permettant de visualiser les principales sources de variation génétique. Depuis les travaux de Price et al. [41], l'ACP est devenue la méthode de référence pour corriger la stratification dans les GWAS. Toutefois, les premières composantes principales capturent principalement la structure ancienne et grossière, ce qui laisse la structure fine récente dans des composantes ultérieures où elle se confond avec le bruit statistique aléatoire [42].

L'**analyse en coordonnées principales** (*Principal Coordinates Analysis,* PCoA), également appelée MDS classique (*Multidimensional Scaling*) [60, 61], décompose des distances précalculées et cherche à les préserver lors de la projection en un nombre moindre de dimensions. Cette méthode est particulièrement utile lorsque les données sont

exprimées sous forme de matrice de distances, comme les matrices de parenté transformées en distances $(1 - \varphi)$.

1.7.2 Méthode non linéaire : UMAP

La **projection UMAP** (*Uniform Manifold Approximation and Projection*) est une technique non linéaire développée par McInnes et al. [62]. Contrairement à l'ACP, la projection UMAP préserve à la fois la structure locale (les relations de voisinage) et globale des données. Elle est particulièrement efficace pour révéler des gradients géographiques et des continuums génétiques subtils qui seraient écrasés par une projection linéaire [63].

La projection UMAP comporte une composante stochastique (c'est-à-dire, due au hasard) inhérente : différentes exécutions avec les mêmes paramètres peuvent produire des projections visuellement différentes [62]. Cette variabilité peut être atténuée par l'utilisation d'une graine aléatoire (*random seed*) fixe et d'une initialisation déterministe, comme la PCoA décrite ci-dessus. Toutefois l'interprétation des distances dans une projection UMAP requiert de la prudence : les distances entre groupes distincts ne sont pas nécessairement comparables aux distances au sein d'un même groupe, et peuvent être affectées par les données utilisées pour initialiser la projection [64].

1.7.3 Classification non supervisée

La méthode des ***k*-moyennes** est l'approche la plus utilisée de classification non supervisée : elle segmente l'ensemble des données en un nombre prédéfini de regroupe-

ments et optimise cette partition en minimisant la distance entre les données d'un même groupe [65] par rapport à un point central qu'on nomme centroïde [66].

HDBSCAN (*Hierarchical Density-Based Clustering of Applications with Noise*) est une méthode de classification non supervisée basée sur la densité [67]. Contrairement aux méthodes comme les *k*-moyennes qui forcent les données dans un nombre prédéfini de groupes, HDBSCAN identifie des groupes de densité variable et considère les points isolés comme du bruit.

1.8 Le contexte démographique et génétique du Québec

Les outils méthodologiques présentés précédemment, soit les matrices de parenté, la réduction de dimension et la classification non supervisée, trouvent une application particulièrement pertinente dans les populations à effet fondateur où l'histoire démographique récente façonne une structure génétique détectable. À cet égard, le Québec constitue un cas d'étude privilégié.

1.8.1 L'effet fondateur québécois

Le peuplement d'origine européenne du Québec se caractérise par une migration provenant principalement de France (1608-1760) impliquant un nombre limité de fondateurs, suivie d'un isolement relatif après la fin du régime français. Environ 80 % du pool génique chez les Québécois d'origine canadienne-française remonte à moins de 7 000 pionnières et pionniers européens établis au XVII^{e} siècle [49].

Cet effet fondateur initial a été amplifié par une expansion démographique rapide et un peuplement du territoire, à partir de la vallée du Saint-Laurent, par vagues successives, créant un effet fondateur en cascade au niveau régional [68, 69]. Certaines nouvelles régions peuplées à partir d'un sous-ensemble de la population source ont subi un tri génétique supplémentaire, ce qui a concentré davantage certains allèles.

Cette histoire démographique a conduit à des enrichissements régionaux de variants rares, dont notamment dans la région du Saguenay–Lac-Saint-Jean [70]. Une étude récente suggère que des phénomènes similaires se sont produits dans d'autres régions moins étudiées comme la Beauce [71].

1.8.2 Le Saguenay–Lac-Saint-Jean comme modèle d'étude

La région du Saguenay–Lac-Saint-Jean (SLSJ, Figure 1) représente un modèle exceptionnel pour la génétique des populations en raison de plusieurs caractéristiques uniques.

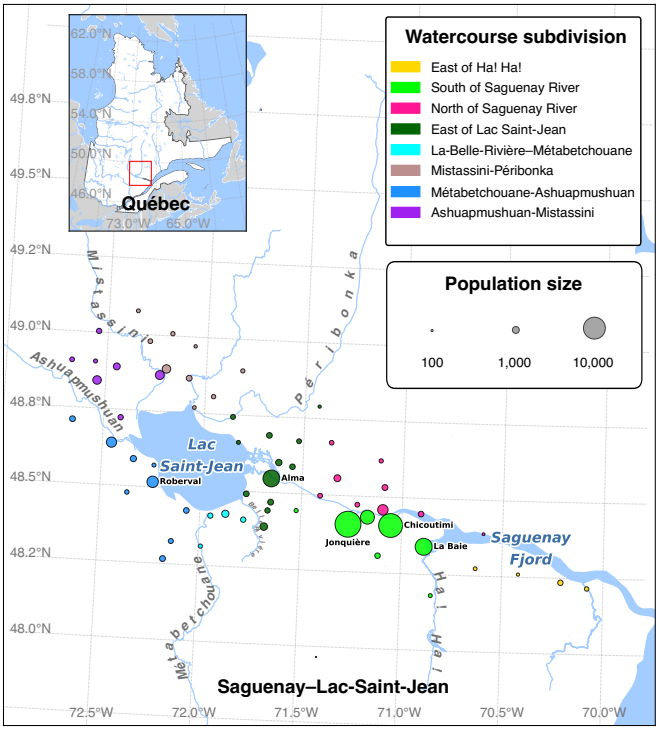


Fig. 1.1 : Carte du Saguenay–Lac-Saint-Jean.

La taille des cercles reflète le nombre de proposants mariés entre 1931 et 1960 dans chaque municipalité de BALSAC (pour un total de 80 348). La carte a été générée par l'auteur en utilisant des données cartographiques sous domaine public de *Natural Earth* et des géométries fluviales de *OpenStreetMap* (© contributeurs OpenStreetMap), disponible sous la licence *Open Database License* (ODbL).

Tout d'abord, le SLSJ a subi un **peuplement d'origine européenne récent qui est bien documenté.** Ce peuplement débute en 1838 [72]. La région est principalement peuplée par des pionnières et pionniers de Charlevoix remontant le fjord du Saguenay [73], permettant de retracer les lignées familiales de façon exceptionnelle.

Ensuite, le peuplement a subi un **effet fondateur en cascade.** La population issue de l'immigration principalement française s'est d'abord établie dans la vallée du Saint-Laurent, puis dans Charlevoix, subissant un premier tri génétique. Certains descendants ont ensuite migré vers le SLSJ, amplifiant l'effet fondateur initial.

Également, la population du SLSJ a été témoin d'une **expansion démographique rapide :** sa population a été multipliée par 25 en un siècle, principalement en raison des taux de natalité élevés, alors que dans l'ensemble de la province, on parle plutôt d'une multiplication par 5 pour la même période [74, 75]. Couplée à l'effet fondateur de Charlevoix-Saguenay, cette expansion a favorisé la transmission de certains allèles rares par dérive génétique.

Enfin, la région possède une **fréquence accrue de certaines maladies héréditaires.** En effet, l'histoire démographique de la région a conduit à une prévalence élevée de plusieurs maladies récessives rares, notamment la tyrosinémie de type 1 [76], l'ataxie récessive spastique de Charlevoix-Saguenay (ARSACS) [77, 78], et plusieurs autres conditions [79-82].

1.8.3 Le paradoxe de l'homogénéité présumée

Malgré ces caractéristiques suggérant une structure génétique complexe, certains chercheurs ont caractérisé la génétique du SLSJ comme relativement **homogène** [81, 83], ce qui laisse croire qu'il pourrait y avoir une absence de structure contrairement à ce

qu'on observe à l'échelle de la province. Toutefois, des études démographiques antérieures indiquent qu'une structure de population existe bel et bien au sein du SLSJ [84, 85], ce qui soulève une contradiction entre ces deux perspectives.

1.8.4 Les sources de données

L'étude de la population québécoise est rendue possible grâce à des ressources exceptionnelles :

Le **fichier de population BALSAC** constitue une base de données généalogique et démographique exceptionnelle par son ampleur et sa profondeur. Il permet de reconstruire les structures familiales du Québec depuis le XVII^{e} siècle à partir des actes de l'état civil, principalement les mariages catholiques [86]. Il est donc possible d'effectuer des reconstructions généalogiques qui s'étendent sur jusqu'à 19 générations.

La **cohorte CARTaGENE** (CaG) est une biobanque québécoise [87] qui comprend 43 032 participantes et participants âgés de 40 à 69 ans (en 2005), recrutés dans les principales zones urbaines de la province, notamment dans la ville de Saguenay [88]. Parmi ces participants, environ 30 000 individus ont été génotypés, et environ 10 000 résidentes et résidents du Québec ont été jumelés aux registres BALSAC, ce qui permet une analyse parallèle des données génomiques et généalogiques.

1.9 Problématique

La stratification de population constitue un facteur de confusion critique dans les études d'association génétique, pouvant persister malgré les corrections par composantes principales de variants communs. Les méthodes traditionnelles, efficaces pour distinguer les grandes divergences ancestrales, échouent souvent à capturer la structure fine récente d'une population régionale [42].

Un défi technique majeur freine l'analyse approfondie de cette structure fine : l'échelle des données. Analyser conjointement des données génomiques massives et des généalogies profondes requiert le calcul de coefficients de parenté pour des milliards de paires d'individus. Pour une génération complète de la région du SLSJ, soit environ 80 000 proposants mariés entre 1931 et 1960, il faut évaluer plus de 3,2 milliards de paires. Ce défi de calcul rend les algorithmes standards inapplicables pour l'analyse de populations entières.

Par ailleurs, la question de la structure génétique intrarégionale du SLSJ demeure non résolue. Cette région, souvent présentée comme génétiquement homogène en raison de son fort effet fondateur, pourrait en réalité receler une structure fine significative qui reflète son histoire de peuplement complexe, une structure qui aurait des retombées importantes pour les études d'association et les programmes de dépistage génétique.

1.10 Hypothèse

Nous posons l'hypothèse que l'histoire démographique de la région du SLSJ, soigneusement documentée dans le fichier BALSAC, a engendré une structure génétique fine détectable non seulement à l'échelle régionale, mais jusqu'au niveau municipal. Nous postulons aussi que la structure attendue à partir de la généalogie concorde fortement avec la structure réalisée par la génétique et mesurée par les segments IBD, ce qui validerait ainsi l'usage de la généalogie comme outil de haute précision pour la génétique des populations.

1.11 Objectifs

L'objectif général de ce projet est de caractériser la structure fine de la population du SLSJ en intégrant la génétique et la généalogie à grande échelle. Il se décline en quatre objectifs spécifiques :

1. **Développer** une méthode algorithmique efficace capable de calculer la parenté généalogique pour l'ensemble d'une population (plusieurs dizaines de milliers d'individus sur 19 générations), surmontant ainsi les limites computationnelles actuelles.

2. **Visualiser** et décrire la structure fine de population de toute une génération du Saguenay–Lac-Saint-Jean à partir de la généalogie.

3. **Comparer** la structure de population attendue (généalogique) et réalisée (génétique) afin de valider la concordance entre les archives historiques et les données génétiques contemporaines, en quantifiant cette relation par des coefficients de corrélation.

4. **Cartographier** la structure fine du SLSJ pour révéler les patrons de peuplement et quantifier l'apport génétique spécifique des fondateurs et fondatrices régionaux provenant de différentes municipalités de Charlevoix.

Le chapitre suivant présente l'article scientifique intitulé « *Fine-scale structure of a whole regional population through genetics and genealogies* », en cours de publication dans la revue *Nature Communications*. Cet article détaille la méthodologie développée, présente les résultats de l'analyse de la structure fine du SLSJ et discute de leurs retombées.

Chapitre 2: Article scientifique

Fine-scale structure of a whole regional population through genetics and genealogies

Gilles-Philippe Morin (1, 2, 3), Claudia Moreau (1, 2), Amadou Barry (2, 3, 4), Simon L. Girard (1, 2, 3, 5, 6, *)

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

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

3. Unité mixte de recherche INRS-UQAC en santé durable, Saguenay, Québec G7H 2B1, Canada.

4. Centre Armand-Frappier Santé Biotechnologie, Institut national de la recherche scientifique (INRS), Laval, Québec H7V 1B7, Canada.

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

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

* Corresponding author: simon2_girard@uqac.ca

2.1 Abstract

Population stratification can confound genetic association studies and often persists despite adjustment using principal components of common variants. Demographic history and rare variants can also contribute to this confounding. But to what extent does demographic history impact fine structure and can be detected in human populations? To address this question, we analysed the Saguenay–Lac-Saint-Jean region of Quebec, a recent founder population long assumed to be homogeneous. Integrating genotype data with genealogical records, we show a strong concordance between realised (genetic) and expected (genealogical) kinship. Using a time-efficient algorithm capable of computing billions of pairwise kinship coefficients for all individuals married in the region between 1931 and 1960, we reveal fine structure at the municipal level, including an east–west genetic gradient shaped by differential founders' contribution, migration patterns and socioeconomic factors. These findings challenge the assumption of regional homogeneity and suggest that similar recent structure likely exists in other populations worldwide. These signals may be obscured by coarse, ancient structure under standard stratification corrections used in genome-wide association studies and polygenic risk score analyses which can lead to biases and false associations when realised allele frequencies correlate with phenotypic variation.

2.2 Introduction

Population stratification occurs when genetic structure (non-random allele distribution) and environmental differences between subpopulations confound genetic association studies. This can make environmental effects appear genetic, leading to misleading re-

sults in genome-wide association studies (GWAS) and polygenic scores[1,2]. Family-based GWAS, which controls for a shared familial environment, reveals that standard population GWAS often overestimates genetic effects[3]. This provides evidence that population stratification persists despite conventional population structure controls, challenging assumptions in large-scale GWAS. When these GWAS are used to construct polygenic scores, even small biases can accumulate into large errors. Critically, it has been shown that the impact of population stratification is also shaped by demographic history which is often overlooked even after adjusting for 100 principal components of common variants[4]. This study suggests that using principal components derived from rare variants, which capture most of the genetic diversity, or from identity-by-descent (IBD) segments may correct the stratification for some types of effects.

IBD analysis has proven effective for uncovering fine-scale population structure in large cohorts. For example, a study using the UK Biobank[5] revealed subtle genetic substructure among British individuals, while research on New York City residents identified 16 ancestry groups, including seven founder populations[6], through the All of Us Research Program and the Mount Sinai BioMe biobank. Fine-scale structure has also been documented in the Quebec founder population[7–9]. These findings raise an important question: to what extent can fine-scale structure be detected in populations? In this study, we address this question by examining a small, recent founder population, the Saguenay–Lac-Saint-Jean (SLSJ) region, leveraging extensive genotype and genealogical data to characterise its genetic architecture.

The SLSJ region of Quebec, Canada (Fig. 2.1), represents a remarkable model for population genetics research due to its well-documented founder effect, recent settlement and unique demographic history. From the seventeenth century, European settlers of mostly French origin have migrated along the Saint Lawrence River, then founded the Charlevoix region, before settling in Saguenay and Lac-Saint-Jean from 1838[10]. This serial founder effect was accentuated by a 25-fold population surge in SLSJ within a century, mainly due to high birth rates[11,12]. The region's demographic characteristics are documented through the BALSAC population register[13], a comprehensive database documenting civil records, primarily Catholic marriages, of French-Canadian individuals from the 1660s (New France) to the 1960s (Quebec), encompassing over three centuries of demographic history. This demographic history is known to have led to several variants being more frequent in the region[14] and to the increase of some rare diseases[15–18]. Some researchers have characterised the genetic background of SLSJ as relatively homogeneous[16,19], suggesting that, unlike the broader provincial context, the region might lack meaningful structure. Yet, demographic studies indicate that population structure does exist within SLSJ[20,21], challenging the assumption of homogeneity.

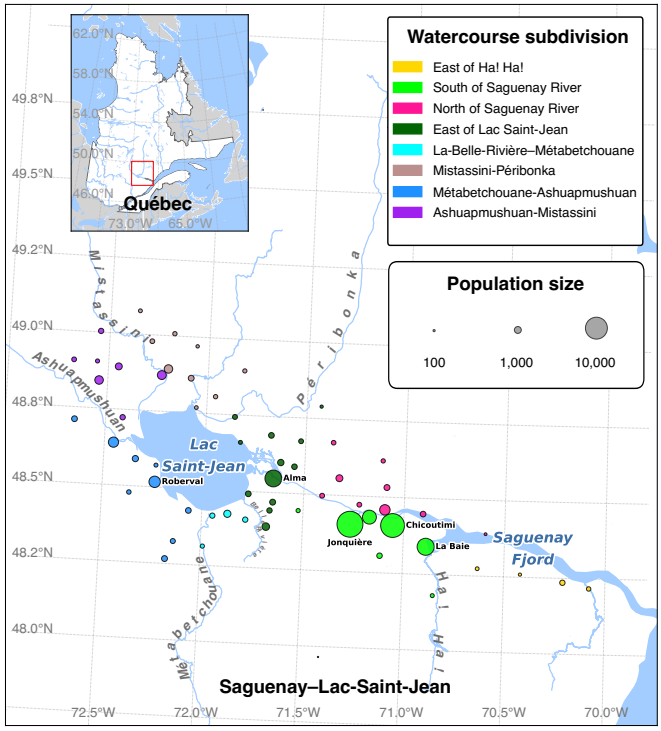


Fig. 2.1: Map of Saguenay–Lac-Saint-Jean municipalities.

Circle size reflects the number of probands married between 1931–1960 in each BALSAC municipality (total 80,348). The map was generated by the authors using public domain cartographic data from Natural Earth and river geometries from OpenStreetMap (© OpenStreetMap contributors), available under the Open Database License (ODbL).

Previous research in the French-Canadian population has shown that both genotype and genealogical data correlate substantially[7–9,22,23]. This shows that genealogy may be used as a proxy to study the genetic structure of populations. Yet, despite the wealth of genealogical and demographic data available for the SLSJ region, significant challenges re-

main in adequately visualising the region's population structure at a large scale. This limitation primarily stems from algorithms that repeatedly compute kinship for redundant pairs of individuals within a genealogy[24], making analysis of complete populations computationally intractable. Although efforts have optimised these computations[25,26], implementing these methods at the scale needed for whole-population analysis, particularly for visualization purposes, has remained challenging, restricting samples to hundreds or a few thousand individuals at a time. The absence of appropriate analytical methods has thus limited comprehensive investigation of fine-scale structure in populations.

This study aims to comprehensively investigate the fine-scale population structure within the SLSJ region by integrating genotype and genealogical datasets. We also retrace the population's origins through the expected genetic contribution of the region's founders. By doing so, this article reveals a fine-scale population structure shaped by geographical boundaries, differential founders' genetic contribution and socioeconomic factors in the recent SLSJ founder population.

2.3 Results

2.3.1 Comparison between genetic and genealogical structure

The CARTaGENE cohort comprises individuals with both genotype and genealogical data, of which 7,970 probands have sufficient genealogical completeness and are not siblings (see methods). Uniform Manifold Approximation and Projection (UMAP) with two dimensions based on realised and expected kinship display strikingly similar popula-

tion structure patterns (Fig. 2.2); a linear regression between realised and expected kinship further supports this consistency with a Pearson correlation coefficient of 0.78 ($p < 0.001$). Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) on a UMAP with ten dimensions identified a subset of 938 individuals most likely originating from the SLSJ region (Supplementary Fig. 1). Within this subset, UMAP projections derived from both realised and expected kinship (Fig. 2.3) reveal partially overlapping gradients of individuals from the diverse subdivisions of SLSJ (Supplementary Fig. 2) with a significant correlation between realised and expected kinship of 0.83 ($p < 0.001$).

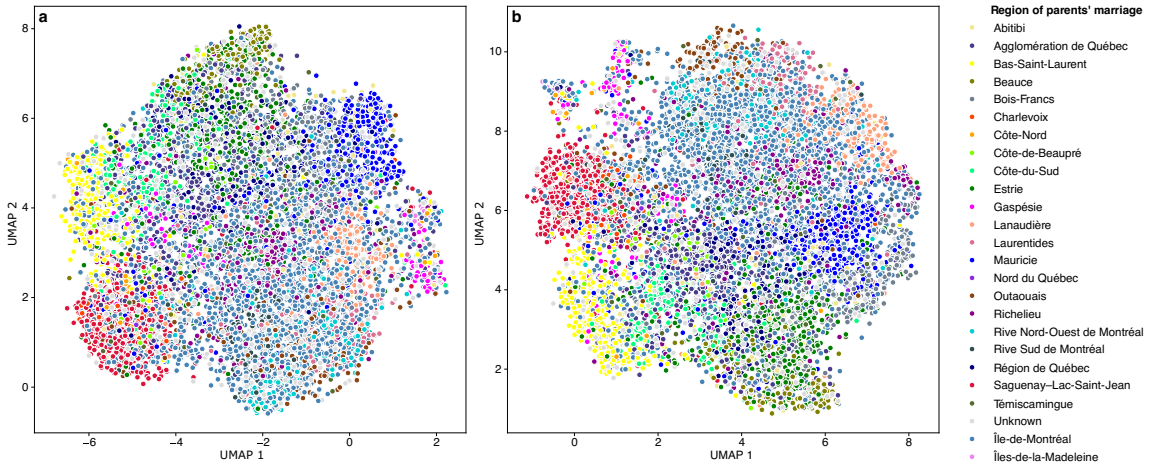


Fig. 2.2: Two-dimensional UMAP of CARTaGENE individuals based on pairwise realised (a) and expected (b) kinship.

Uniform Manifold Approximation and Projection (UMAP) of 7,970 individuals from the CARTaGENE cohort, computed from their (**a**) realised kinship and (**b**) expected kinship transformed as a precomputed distance ($1 - \phi$). The colours represent the region of parents' marriage.

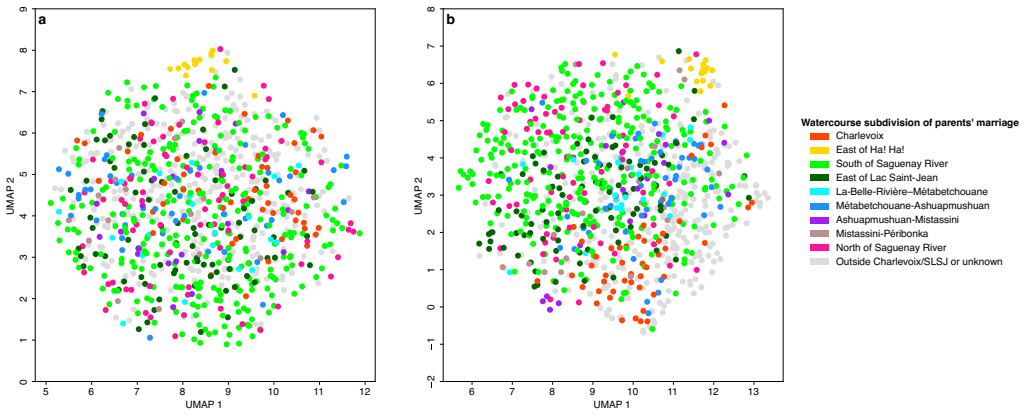


Fig. 2.3: Two-dimensional UMAP of CARTaGENE individuals from SLSJ based on pairwise realised (a) and expected (b) kinship.

Uniform Manifold Approximation and Projection (UMAP) of 938 individuals from the CARTaGENE cohort who are most likely to originate from SLSJ, computed from their (**a**) realised kinship and (**b**) expected kinship transformed as a precomputed distance ($1 - \phi$). The colours represent the SLSJ watercourse subdivision of parents' marriage.

2.3.2 Fine-scale population structure of a whole generation

The great correlation between the population structure observed through genetic and genealogical data[7–9,22,23] allows us to use the genealogy as a proxy for examining the fine-scale structure of the whole SLSJ population of the last generation available in BALSAC. The two-dimensional UMAP projection of non-sibling probands married between 1931 and 1960 in the SLSJ region reveals clear concentrations of individuals whose parents were married in the same subdivision (Fig. 2.4). The observed structure follows an east–

west gradient, extending from the southeast (East of Ha! Ha! subdivision) to the northwest (Ashuapmushuan–Mistassini subdivision) (see Fig. 2.1). This gradient exhibits a circular pattern centred around individuals originating from Charlevoix and concludes with a distinct grouping of individuals whose ancestry lies outside both Charlevoix and SLSJ. More remote, rural municipalities exhibit higher levels of concentration in genealogical relationships, while urban centres display greater dispersion (Supplementary Fig. 3).

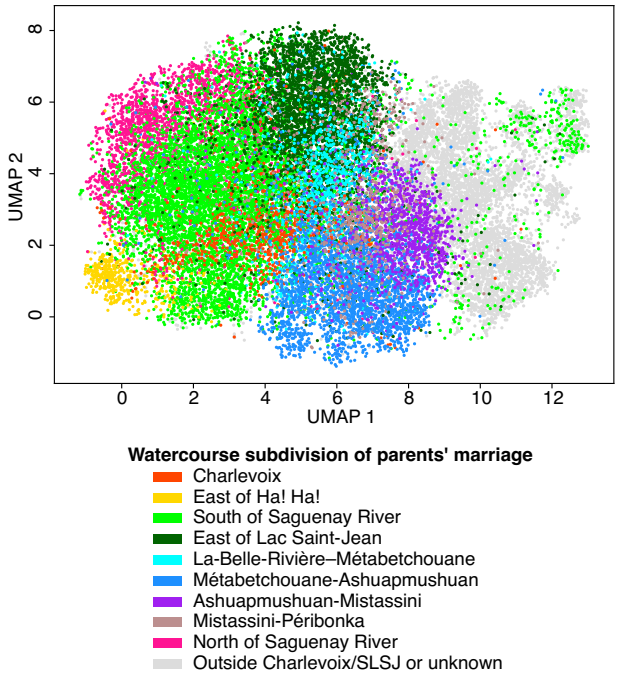


Fig. 2.4: Two-dimensional UMAP of pairwise expected kinship for the last-generation SLSJ population.

Uniform Manifold Approximation and Projection (UMAP) of 26,445 non-siblings who married between 1931 and 1960 in Saguenay–Lac-Saint-Jean, com-

puted from their expected kinship (ϕ) transformed as a precomputed distance ($1 - \phi$). The colours represent the watercourse subdivision of parents' marriage.

2.3.3 Expected genetic contribution of founders

To refine our understanding of the migratory patterns underlying the observed structure, we calculated the expected genetic contribution of the SLSJ founders (ancestors married in SLSJ whose parents married elsewhere, see methods) to the SLSJ probands of the last generation. Mean genetic contribution from the SLSJ founders originating from Charlevoix is the highest in the eastern part of SLSJ and progressively declines along an east–west axis (Fig. 2.5a)[21]. Individuals from the northwesternmost municipalities of Lac-Saint-Jean exhibit more diverse ancestral origins, as well as individuals originating from urban areas, especially south of Saguenay River that exhibit high diversity. A detailed analysis (Fig. 2.5b and 2.5c) reveals that this east–west genetic gradient is primarily driven by founders from La Malbaie and to some extent Les Éboulements, which both contributed most significantly to the Saguenay area (see also Supplementary Fig. 4). In contrast, Baie-Saint-Paul contributed more to the Lac-Saint-Jean region (Fig. 2.5b and 2.5c, see also Supplementary Fig. 4). Notably, Les Éboulements contributed disproportionately to the east of Ha! Ha! which also presents the highest contribution from Charlevoix founders overall (Supplementary Fig. 5). Individuals from the area between Ashuapmushuan and Mistassini largely trace their ancestry to regions outside Charlevoix (Fig. 2.5b and 2.5c), with this diversification becoming prominent around 1885 (Supplementary Fig. 5).

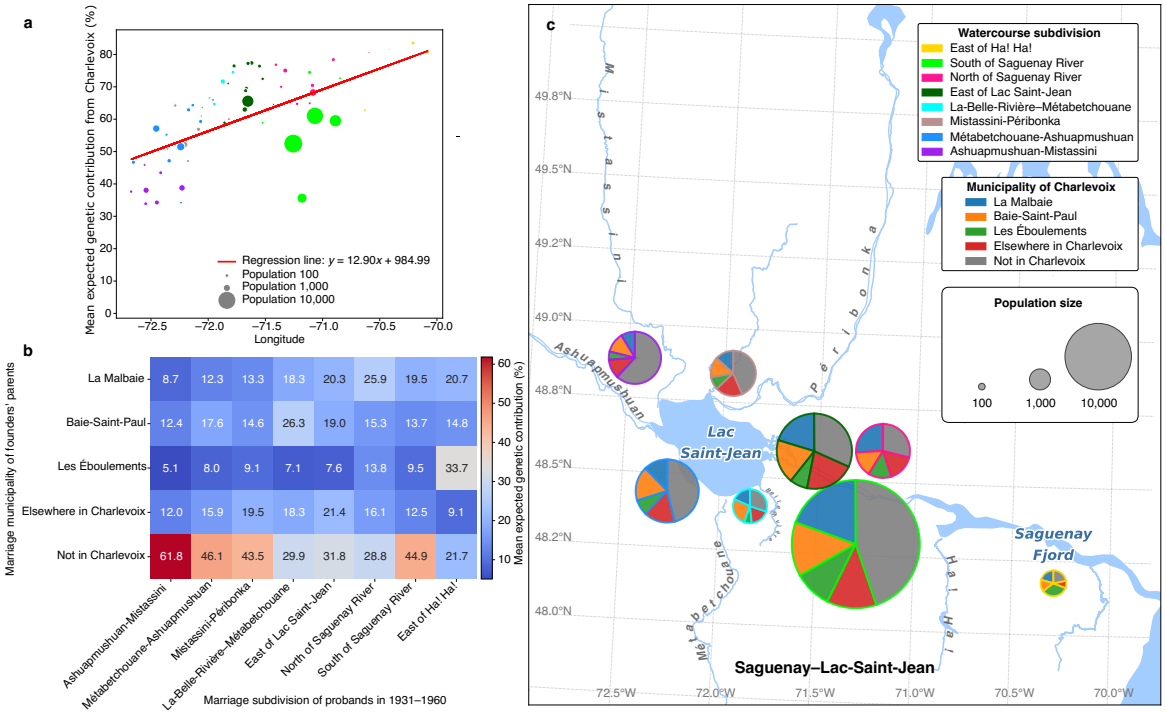


Fig. 2.5: Proportion of the mean expected genetic contribution to probands explained by SLSJ founders depending on geography (a) and on various Charlevoix municipalities (b and c).

Mean expected genetic contribution of Saguenay–Lac-Saint-Jean (SLSJ) founders from Charlevoix to 80,348 probands as a function of the longitude of the probands' subdivisions (**a**). The red line shows the unweighted linear regression with a Pearson correlation coefficient of 0.517 (p-value < 0.001). Mean expected genetic contribution of SLSJ founders from Charlevoix per municipality to probands' in each subdivision (**b** and **c**). The map was generated by the authors using public domain cartographic data from Natural Earth and river geometries from OpenStreetMap (© OpenStreetMap contributors), available under the Open Database License (ODbL).

2.4 Discussion

In this study, we found fine-scale population structure in SLSJ, challenging the notion of homogeneity in this founder population[16,19]. This fine-scale structure was assessed and visualised on the entire SLSJ population, thanks to deep genealogical data and a time-efficient hybrid algorithm for computing kinship coefficients on such a massive scale. It is therefore reasonable to expect that a fine structure due to recent demographic history exists in all human populations.

As already known, the fine-scale population structure may have an impact on rare variants[27,28] and association studies[29,30]. Quebec is known for regional enrichments of rare variants, most notably in SLSJ[14], but evidence points to similar patterns in other less studied regions[28,31–33]. Our study suggests that such enrichments may be more localised than at the regional scale, even within regions with a strong founder effect. Within the population of a single region, the carrier rate may vary from one location to another. If a variant is more frequent in one part of the region, and less in another, considering only the overall carrier rate may mitigate the assessment of enrichment and lead to an under-appreciation of the variant's frequency in the region. This factor is intensified if the carrier tests are performed unevenly, such as in urban centres more than in rural municipalities.

Beyond influencing the distribution of rare variants, fine-scale population structure may pose challenges for GWAS, especially when it results in allele frequencies that correlate with phenotypic variations[34,35]. This could introduce confounding factors that

standard correction methods may not fully address. Previous findings suggesting polygenic adaptation on height across European populations were later shown to be largely attributable to uncontrolled stratification, prompting a reassessment of polygenic score comparisons between populations[1,2]. Subsequent work has demonstrated that geographic structure remains detectable in a large homogeneous cohort even after adjustment for study centre and genetic principal components[36]. Using simulations, we demonstrated that the SLSJ fine-scale population structure, shaped by heterogeneous founder contributions and migration routes, can induce detectable stratification in a GWAS that PCA of common SNPs fails to completely account for (Supplementary Fig. 6 and Supplementary Methods). Detecting such fine-scale structure within a relatively small and homogeneous population underscores the critical need to control for population stratification in GWAS and polygenic analyses beyond the first 20 or 100 principal components generally used[4].

In this study we reaffirm the correlation between genetics and genealogy[7–9,22,23]. Therefore, using genealogy as a proxy for genetic structure is valuable given its strong correlation with realised kinship, which has been shown to capture fine-scale population structure in other populations[5,6]. Indeed, it enabled us to study the fine structure of the entire SLSJ population, which would not be possible with genetic data alone, which are unavailable for all individuals, and are often concentrated in urban areas, highlighting the need for inclusion of smaller and remote communities in future genetic studies[37–40]. This work also showcases some methodological advances, such as the power of parallelised kinship computation via GeneaKit: billions of coefficients inferred efficiently across tens of thousands

of individuals. Our hybrid pipeline, melding Karigl's[24] and Kirkpatrick's[26] algorithms, proved to be time-efficient and is readily extensible to other deep genealogy datasets worldwide.

By applying the new kinship algorithm to the whole SLSJ population, it allowed us to uncover a fine population structure which may be partly explained by the migratory patterns that shaped the SLSJ population. Previous work described a Charlevoix gradient, in which eastern municipalities show a higher genetic contribution from Charlevoix founders than those farther west[21]. However, we show here for the first time at a finer scale that this gradient is less homogeneous than suggested. Indeed, it centres on La Malbaie, the main contributor to Saguenay, while Baie-Saint-Paul's influence predominates in Lac-Saint-Jean. In the eastern part of the region, Les Éboulements shows a strong but previously undocumented genetic contribution to the East of Ha! Ha! subdivision, following an early wave of migration from La Malbaie. Moving westward, individuals show less than half of their genetic input from Charlevoix, instead tracing largely to external Quebec regions other than Charlevoix. In the northwest of Lac-Saint-Jean, the opening of the railroad in 1888[10] brought in diverse founders from outside Charlevoix[21]. By 1912, many Acadians settled south of the Saguenay River attracted by jobs in the paper and aluminium industries[41], diluting Charlevoix's relative genetic contribution in those urban centres. Indeed, the SLSJ fine structure was shaped not only by geography and migration but also by socioeconomic factors.

While our study provides valuable insights into Quebec's genetic structure, it is important to acknowledge certain limitations. The CARTaGENE dataset, although rich in genotype data, primarily comprises individuals recruited in urban areas, which might lead to an underrepresentation of more isolated communities. Similarly, the BALSAC genealogies, while extensive, exhibit regional variations in completeness[8]. This uneven data quality could potentially introduce biases when detecting population structure. Furthermore, our use of parents' marriage location appears to be a better proxy for regional affiliation than the recruitment locations of CARTaGENE. It may still not fully account for individual or family mobility over generations or situations where parents originate from different regions. Despite these limitations, our rigorous quality control, which involved filtering out individuals with more than half null kinship coefficients, allowed us to observe the Quebec-wide genetic structure previously described. This resolution was sufficient to distinguish the peripheries of the SLSJ region. We also found that our UMAP projections, despite the inherent stochastic nature of the algorithm that can make direct comparisons challenging[42], yielded highly similar results with matrices of expected and realised kinship. These findings underscore the robustness of our analysis in capturing the underlying genetic landscape of SLSJ and the province of Quebec.

In conclusion, integrating large-scale genealogical and genotype data reveals fine-scale structure in SLSJ that broadly aligns with geographic gradients. Such genetic heterogeneity suggests uneven distributions of founder variants and, by extension, locally variable prevalence of associated genetic diseases. These findings also have implications

for how we understand human genetic diversity more broadly. The use of self-reported race and ethnicity as proxies for human genetic diversity has long been widespread, and recent applications of UMAP have been employed to support such categorisations[43]. In reality, population structure and genetic diversity are observable not only at the continental scale, but also within regional populations, such as those in the province of Quebec. Our findings further demonstrate that population structure can be detected at even finer geographic resolutions, down to the level of municipalities, where we observe gradients of genetic contribution from different founders in addition to sharply defined clusters for more remote localities. Such regional and local gradients can impact genetic association studies, and should not be overlooked.

2.5 Methods

This study was approved by the ethics board of the Université du Québec à Chicoutimi (UQAC) and complies with all relevant ethical regulations.

2.5.1 Genotype data and cleaning

Genotype data was drawn from the CARTaGENE cohort[44], which comprises 43,032 participants aged 40–69 (in 2005) recruited in the main urban areas of the Quebec Province, notably in the city of Saguenay.

Of those participants, 29,337 individuals were genotyped using a variety of platforms, including Omni 2.5, GSAv1 with a multi-disease panel, GSAv1, GSAv2 with a

multi-disease panel, GSAv3 with a multi-disease panel, GSAv2 with a multi-disease panel and add-on (see https://cartagene.qc.ca/files/documents/other/Info_GeneticData3juillet2023.pdf for details). Each dataset was processed independently using PLINK v1.9[45], retaining individuals with at least 95% genotyping across all SNPs. At the SNP level, we retained autosomal variants with a call rate of at least 95% across individuals and that conformed to Hardy–Weinberg equilibrium ($p > 10^{-6}$, calculated within each dataset). All chips were merged, and the final dataset comprised 148,200 SNPs across 28,358 individuals.

2.5.2 Genealogical data

Genealogical data was obtained from the BALSAC population register[13]. Two different genealogical datasets have been extracted for this study: First, 9,405 Quebec residents from the CARTaGENE cohort have been matched to BALSAC records. Second, to analyse a whole generation, we also selected 80,348 probands (i.e. non-parent individuals) married in SLSJ from 1931 to 1960. For both probands' sets, a multigenerational genealogy extending up to 19 generations was reconstructed. When available, information on the municipality, region, and year of marriage was included; to ensure confidentiality, marriage years are rounded up to the nearest 5. Individual anonymity is rigorously maintained through the use of unique, non-identifying codes.

Traditional geographic subdivisions split SLSJ into Lower Saguenay, Upper Saguenay, and Lac-Saint-Jean[20]. However, it has been demonstrated that geographic features and natural barriers, such as watersheds, can influence patterns of population struc-

ture[7]. We carefully carved eight subdivisions along primary watercourses to consolidate smaller communities and municipalities were grouped based on these watercourse boundaries (Supplementary Table 1 and Fig. 2.1). On average, we observed reduced migration in the last generation within our defined geographical subdivisions than between them (Supplementary Fig. 7).

2.5.3 IBD sharing (realised kinship)

Pairwise IBD segments were inferred on phased genotypes using Refined IBD (version 17Jan20)[46] and Beagle (version 18May20)[47]. A matrix of realised kinship estimation was computed by summing shared segment lengths of at least 2 centiMorgans (cM) using a Python 3 ported version of relatedness_v1.py from Sharon R. Browning (see https://faculty.washington.edu/sguy/ibd_relatedness.html). The diagonal was set to 1.

2.5.4 Genealogical kinship coefficients (expected kinship)

Expected kinship (ϕ)[48] was used to infer and visualise population structure through genealogical data. It corresponds to the probability that, at one locus, one randomly picked allele from individual *i* and one randomly picked allele from individual *j* are IBD or come from the same ancestor. Expected kinship coefficients were computed between all pairs of probands from whole SLSJ and CARTaGENE genealogical datasets. Of the 80,348 SLSJ probands, 26,445 are non-siblings ($\phi < 0.2$) and have a genealogy deemed sufficiently complete, i.e. they are related ($\phi > 0$) with at least half of the probands. Of the 9,405 indi-

viduals in CARTaGENE data, 7,970 are neither parents nor siblings, and have a genealogy deemed sufficiently complete using the same criterion.

Hybrid algorithm for time-efficient computation of billions of kinship coefficients

To efficiently compute kinship coefficients across all pairs of individuals, we implemented a hybrid of two existing algorithms. Karigl's algorithm[24] estimates kinship coefficients on a per-proband-pair basis, enabling parallel processing of the kinship matrix, as in the GENLIB package for R[49]. However, it reprocesses each ancestor every time they appear, either multiple times across genealogies (kinship) or repeatedly within the same genealogy (inbreeding), which makes it computationally costly. Kirkpatrick's algorithm[26] addresses this inefficiency by partitioning the genealogy into successive sub-genealogies (e.g., one generation at a time) and computing kinship coefficients in a top-down manner, with the probands of one sub-genealogy treated as the founders of the next, thereby minimizing redundant ancestor processing. Yet, Kirkpatrick's approach is not designed for parallel execution. Our method combines the strengths of both: we divide the genealogy into generational sub-genealogies, as in Kirkpatrick's approach, but compute each sub-genealogy's kinship matrix in parallel, achieving substantial time gains. For instance, computing kinship for all 6,455,801,104 pairs of 1931–1960 BALSAC SLSJ probands required only two minutes with our algorithm. It was not possible to compare directly with Kirkpatrick's implementation due to the fact that this software is no longer available online, nor with the GENLIB implementation because its execution would require more resources than can be allocated. Following the pre-publication of this work, it was brought to our attention that an

indirect method by Colleau[50] can also be applied to compute kinship coefficients at a large scale[51]. However, this method incorporates all generations in the genealogy, which was not the objective of our analysis.

Our new implemented algorithm is provided as part of GeneaKit 0.1.0, which is freely available on GitHub[52] (https://github.com/Genopop/geneakit) under the MIT licence, using the phi() function for kinship coefficients. A pseudocode of the algorithm is available in the supplementary material (Supplementary Note 1).

2.5.5 Expected genetic contributions

We also computed the expected genetic contribution of the region's founders as a way to measure the region's migratory origins and to better understand the fine structure. Region's founders are defined as individuals who married in the SLSJ whereas their parents married elsewhere. This computation was done using GeneaKit 0.1.0's gc() function, from all regions' founders to all 80,348 genealogical probands who married between 1931 and 1960 in the SLSJ region. The expected genetic contribution of an ancestor corresponds to the probability that an allele is passed down to a given descendant. A founder's region of origin is defined as the region of marriage of their parents.

2.5.6 Clustering and visualisation

CARTaGENE individuals originate from all regions of Quebec, it is thus necessary to identify those who are most likely to originate from SLSJ. UMAP (version 0.5.9)[42]

was performed on both the realised and expected kinship matrices, using precomputed distances of ($1 - \phi$). UMAP was run with ten dimensions, 15 neighbors, a minimum distance of 0.0, and a fixed random seed (random state = 0), to maximise clustering and reproducibility. UMAP was initialised from a classical multidimensional scaling (MDS), also known as principal coordinate analysis (PCoA), of the precomputed distances, implemented as pcoa from scikit-bio 0.6.3[53], using eigendecomposition and a seed of 0. HDBSCAN from scikit-learn 1.8.0[54] was then fitted to the UMAP outputs to identify individuals most likely originating from the SLSJ region, using a min_cluster_size of 25 and a cluster_selection_epsilon of 0.3. Those parameters were chosen as they allowed separation of regional populations in a previous study using the same cohort[44]. From the clustering results for both the realised and expected kinship matrices, the cluster containing the highest number of individuals of known SLSJ origin (i.e. whose four grandparents married in the region) was identified. 585/1141 individuals in the realised kinship cluster and 583/1036 individuals in the expected kinship cluster are of known SLSJ origin. The final set of 938 individuals was then determined as the intersection of these two best clusters (one from the realised kinship matrix clustering and one from the expected kinship matrix clustering). Of those 938 individuals, 566 are of known SLSJ origin.

For visualisation, UMAP (initialised with classical MDS/PCoA) was applied to the kinship matrices using their precomputed distances ($1 - \phi$), two dimensions, and a minimum distance of 0.5.

2.6 Data availability

Due to the informed consent given by study participants, Quebec genotype data is available under restricted access via CARTaGENE's independent Sample and Data Access Committee (SDAC) (https://cartagene.qc.ca/en/researchers/access-request.html). Further information on access to CARTaGENE data is available in the article by McClelland et al.[44]. Access to the BALSAC genealogical data is controlled in accordance with the Policy on Access to BALSAC Data for Research Purposes to protect participant confidentiality and adhere to ethical guidelines. The procedure for requesting access to BALSAC data is described at https://balsac.uqac.ca/acces-donnees/. Access requests must be submitted to the Researchers Service of the BALSAC Project and must include a completed BALSAC Database Access Request Form along with all documents necessary for evaluation, such as a certificate of ethics approval. Applications are reviewed to ensure compliance with the scientific method, feasibility, and BALSAC access policies. Applicants are informed of the decision in writing. Questions regarding data access can be addressed to balsac@uqac.ca. If access is granted, data may be used only within the scope of the approved research project and must be destroyed once the authorized access period expires. The BALSAC data are not publicly available due to ethical restrictions on the publication of genealogical information dating back less than 100 years.

2.7 Code availability

The GeneaKit library is available under the open-source MIT license on GitHub (https://github.com/Genopop/geneakit) and is archived on Zenodo (https://doi.org/10.5281/zenodo.18701153)[52]. The source code used for data processing, analysis, and the production

of figures in this study has been deposited in the following GitHub repository: https://github.com/Genopop/slsj_structure. It is also available on Zenodo: https://doi.org/10.5281/zenodo.18701179[55].

2.8 References

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

2 Berg, J. J. *et al*. Reduced signal for polygenic adaptation of height in UK Biobank. *eLife* **8**, e39725 (2019).

3 Tan, T. *et al*. Family-GWAS reveals effects of environment and mating on genetic associations. *MedRxiv Preprint Server for Health Sciences* https://doi.org/10.1101/2024.10.01.24314703 (2025).

4 Zaidi, A. A. & Mathieson, I. Demographic history mediates the effect of stratification on polygenic scores. *eLife* **9**, e61548 (2020).

5 Nait Saada, J. *et al*. Identity-by-descent detection across 487,409 British samples reveals fine scale population structure and ultra-rare variant associations. *Nat. Commun.* **11**, 6130 (2020).

6 Isshiki, M. *et al*. Genetic disease risks of under-represented founder populations in New York City. *PLoS Genet*. **21**, e1011755 (2025).

7 Anderson-Trocmé, L. *et al*. On the genes, genealogies, and geographies of Quebec. *Science* **380**, 849–855 (2023).

8 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*. **32**, 91–97 (2024).

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

10 Gagnon, S. Sagamiens—Sagamiennes. *Saguenayensia* **30**, 3–6 (1988).

11 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, 1983).

12 Bouchard, G.& De Braekeleer, M. *Histoire d'un génôme: Population et génétique dans l'est du Québec* (Presses de l'Université du Québec, 1991).

13 BALSAC. https://balsac.uqac.ca/.

14 Michel, É. *et al*. Rare diseases load through the study of a regional population. *PLOS Genet*. **21**, e1011876 (2025).

15 Cruz Marino, T. *et al*. Portrait of autosomal recessive diseases in the FRENCH-CANADIAN founder population of SAGUENAY-LAC-SAINT-JEAN. *Am. J. Med. Genet. A* **191**, 1145–1163 (2023).

16 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).

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

18 Scriver, C. R. Human genetics: lessons from Quebec populations. *Annu. Rev. Genom. Hum. Genet* **2**, 69–101 (2001).

19 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).

20 Lavoie, E.-M., Tremblay, M., Houde, L. & Vézina, H. Demogenetic study of three populations within a region with strong founder effects. *Public Health Genom*. **8**, 152–160 (2005).

21 De Braekeleer, M. *L'approche des maladies héréditaires par la démographie génétique: le cas du Saguenay-Lac-Saint-Jean au Québec*. PhD thesis, Univ. Bordeaux 2 (1995).

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

23 Burkett, K. M. *et al*. Correspondence between genomic- and genealogical/coalescent-based inference of homozygosity by descent in large French-Canadian genealogies. *Front. Genet*. **12**, 808829 (2022).

24 Karigl, G. A recursive algorithm for the calculation of identity coefficients. Ann. *Hum. Genet*. **45**, 299–305 (1981).

25 Abney, M. A graphical algorithm for fast computation of identity coefficients and generalized kinship coefficients. *Bioinformatics* **25**, 1561–1563 (2009).

26 Kirkpatrick, B., Ge, S. & Wang, L. Efficient computation of the kinship coefficients. *Bioinformatics* **35**, 1002–1008 (2019).

27 Zhang, B. C., Biddanda, A., Gunnarsson, ÁF., Cooper, F. & Palamara, P. F. Biobank-scale inference of ancestral recombination graphs enables genealogical analysis of complex traits. *Nat. Genet* **55**, 768–776 (2023).

28 Gagnon, L., Moreau, C., Laprise, C. & Girard, S. L. Fine-scale genetic structure and rare variant frequencies. *PLoS ONE* **19**, e0313133 (2024).

29 Henn, B. M., Gravel, S., Moreno-Estrada, A., Acevedo-Acevedo, S. & Bustamante, C. D. Fine-scale population structure and the era of next-generation sequencing. *Hum. Mol. Genet.* **19**, R221–R226 (2010).

30 Hu, S. *et al*. Fine-scale population structure and widespread conservation of genetic effect sizes between human groups across traits. *Nat. Genet.* **57**, 379–389 (2025).

31 Gagnon, M. *et al*. Rare variants and founder effect in the Beauce region of Quebec. *Commun. Biol.* **8**, 1184 (2025).

32 Vézina, H. *et al*. Molecular and genealogical characterization of the R1443X BRCA1 mutation in high-risk French-Canadian breast/ovarian cancer families. *Hum. Genet.* **117**, 119–132 (2005).

33 De Braekeleer, M., Hechtman, P., Andermann, E. & Kaplan, F. The French Canadian Tay-Sachs disease deletion mutation: identification of probable founders. *Hum. Genet.* **89**, 83–87 (1992).

34 Edge, M. D., Gorroochurn, P. & Rosenberg, N. A. Windfalls and pitfalls. *Evol. Med. Public Health* **2013**, 254–272 (2013).

35 Rosenberg, N. A. & Nordborg, M. A general population-genetic model for the production by population structure of spurious genotype–phenotype associations in discrete, admixed or spatially distributed populations. *Genetics* **173**, 1665–1678 (2006).

36 Haworth, S. *et al*. Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis. *Nat. Commun.* **10**, 333 (2019).

37 Brown, J. T., McGonagle, E., Seifert, R., Speedie, M. & Jacobson, P. A. Addressing disparities in pharmacogenomics through rural and underserved workforce education. *Front. Genet.* **13**, 1082985 (2023).

38 Cohen, A. S. A. *et al*. Genomic answers for kids: toward more equitable access to genomic testing for rare diseases in rural populations. *Am. J. Hum. Genet* **111**, 825–832 (2024).

39 Best, S., Vidic, N., An, K., Collins, F. & White, S. M. A systematic review of geographical inequities for accessing clinical genomic and genetic services for non-cancer related rare disease. *Eur. J. Hum. Genet.* **30**, 645–652 (2022).

40 Fatumo, S. *et al*. A roadmap to increase diversity in genomic studies. *Nat. Med* **28**, 243–250 (2022).

41 Gouvernement du Québec. Place des Acadiens—Saguenay (Ville). Comm. Topon. https://toponymie.gouv.qc.ca/ct/ToposWeb/Fiche.aspx?no_seq=446160.

42 McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. https://doi.org/10.48550/ARXIV.1802.03426 (2018).

43 The All of Us Research Program Genomics Investigators et al. Genomic data in the All of Us Research Program. *Nature* **627**, 340–346 (2024).

44 McClelland, P. *et al*. A multi-ancestry genetic reference for the Quebec population. *Nat. Commun.* **17**, 1319 (2026).

45 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).

46 Browning, B. L. & Browning, S. R. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194**, 459–471 (2013).

47 Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet* **81**, 1084–1097 (2007).

48 Wright, S. The interpretation of population structure by F-statistics with special regard to systems of mating. *Evolution* **19**, 395 (1965).

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

50 Colleau, J.-J. An indirect approach to the extensive calculation of relationship coefficients. *Genet Sel. Evol.* **34**, 409 (2002).

51 Lee, H., Craddock, R. F., Gorjanc, G. & Becher, H. randPedPCA: rapid approximation of principal components from large pedigrees. *Genet Sel. Evol.* **57**, 46 (2025).

52 Morin, G. P. *et al*. Genopop/geneakit: geneakit 0.1.0. https://doi.org/10.5281/ZENODO.18701153 (2026).

53 McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **3**, 861 (2018).

54 Aton, M., McDonald, D., Cañardo Alastuey, J. et al. Scikit-bio: a fundamental Python library for biological omic data analysis. *Nat. Methods* **23**, 274–276 (2026).

55 Pedregosa, F. *et al*. Scikit-learn: machine learning in Python. *J. Mach. Learn Res* **12**, 2825–2830 (2011).

56 Morin, G. P. *et al*. *Fine-scale structure of a whole regional population through genetics and genealogies*. https://doi.org/10.5281/zenodo.18701179 (2026).

2.9 Acknowledgements

This work was made possible by the Digital Research Alliance of Canada which provided access to storage and computing resources. We are extremely grateful to all participants in this research.

Funding for S. L. G. was provided by the Canada Research Chair in Genetics and Genealogy CRC-2022-00444, of which he is the chair holder (http://www.chairs.gc.ca). Funding for G.-P. M. was provided by the Unité mixte de recherche INRS-UQAC en santé durable (grant number UIU0002).

2.10 Author contributions

All authors approved the final version of the manuscript. G.-P. M. and C. M. played an important role in interpreting the results. G.-P. M., C. M. and S. L. G. conceived and designed the study and drafted the manuscript. A. B. revised the manuscript.

2.11 Competing interests

The authors declare no competing interests.

Chapitre 3: Discussion et perspectives

3.1 Retour sur les objectifs

3.1.1 Apport méthodologique : l'algorithme hybride de GeneaKit

Le développement de l'algorithme hybride implémenté dans GeneaKit représente une contribution méthodologique significative. En combinant le calcul en parallèle de l'approche de Karigl avec l'approche « diviser pour régner » de Kirkpatrick, nous avons rendu possible l'analyse à l'échelle de populations entières (plus de 3 milliards de paires calculées en deux minutes) là où les méthodes existantes auraient nécessité des ressources inaccessibles.

La mise à disposition du code source de GeneaKit sous licence MIT sur GitHub (https://github.com/Genopop/geneakit), accompagnée du code d'analyse de l'étude (https://github.com/Genopop/slsj_structure), s'inscrit dans une démarche de science ouverte favorisant la reproductibilité et l'extension de nos travaux par la communauté scientifique.

3.1.2 Hétérogénéité régionale

Les résultats de notre étude remettent en question la notion d'homogénéité génétique au sein de la population à effet fondateur du SLSJ. Contrairement aux caractérisations antérieures suggérant que la région pourrait manquer de structure significative [81, 83], nos analyses révèlent une structure intrarégionale détectable jusqu'au niveau municipal.

Cette structure fine suit un gradient est-ouest, s'étendant du sud-est (subdivision Est de Ha! Ha!) au nord-ouest (subdivision Ashuapmushuan–Mistassini). Ce gradient présente un patron circulaire centré sur les individus originaires de Charlevoix et se termine par un regroupement distinct d'individus dont l'ascendance provient de l'extérieur de Charlevoix et du SLSJ. Les municipalités rurales plus éloignées présentent des niveaux de concentration plus élevés dans les relations généalogiques, tandis que les centres urbains affichent une plus grande dispersion.

3.1.3 Concordance entre généalogie et génétique

Notre étude confirme et renforce les observations antérieures sur la corrélation entre les données génétiques et généalogiques dans la population canadienne-française [50-54]. Les projections UMAP basées sur la parenté réalisée (génétique) et la parenté attendue (généalogique) révèlent des patrons de structure de population remarquablement similaires, avec un coefficient de corrélation de Pearson de 0,78 ($p < 0,001$) pour l'ensemble de la cohorte CARTaGENE, qui atteint 0,83 ($p < 0,001$) pour le sous-ensemble d'individus originaires du SLSJ.

Cette forte concordance valide l'utilisation de la généalogie comme indicateur fiable pour la structure génétique. Elle nous a permis d'étudier la structure fine de l'ensemble de la population du SLSJ, soit 26 445 proposants non apparentés avec des généalogies suffisamment complètes, ce qui n'aurait pas été possible avec les seules données génétiques constituées de moins de 4 000 individus originaires du SLSJ majoritairement recrutés dans

les zones urbaines. Notre cohorte généalogique est donc la plus grande cohorte du SLSJ jamais étudiée.

3.1.4 Contribution différentielle des fondateurs de Charlevoix

L'analyse de la contribution génétique attendue des fondateurs régionaux a révélé que le gradient est-ouest observé et précédemment décrit dans la littérature [84, 85] est plus précisément façonné par des contributions différentielles des diverses municipalités de Charlevoix :

- La Malbaie et, dans une moindre mesure, Les Éboulements ont contribué de manière prédominante à la région du Saguenay;

- Baie-Saint-Paul a davantage contribué à la région du Lac-Saint-Jean;

- Les Éboulements présentent une contribution particulièrement forte mais jusqu'alors non documentée à la subdivision à l'est de la baie des Ha! Ha!, qui présente également la contribution globale la plus élevée de fondateurs charlevoisiens;

- Les individus de l'extrême nord-ouest du Lac-Saint-Jean (entre Ashuapmushuan et Mistassini) tracent leur ascendance en grande partie vers des régions extérieures à Charlevoix. Cette diversification devient plus importante autour de 1885 avec l'ouverture du chemin de fer en 1888 [72, 84];

- Les centres urbains au sud de la rivière Saguenay présentent une haute diversité, notamment en raison de migrations provenant d'autres régions, dont l'établissement de nombreux Acadiens attirés par les emplois dans les industries des pâtes et papiers et de l'aluminium à partir de 1912 [89, 90].

3.2 Retombées pour la santé publique et la médecine de précision

3.2.1 Impact sur les scores de risque polygénique

Les gradients géographiques que nous observons, façonnés par les contributions différentielles des fondateurs et les routes migratoires, pourraient introduire une confusion subtile dans les études qui présument l'homogénéité régionale. Nos résultats renforcent l'importance de contrôler la structure fine même au sein de populations apparemment uniformes.

Les scores de risque polygénique (PRS) cumulent les effets de milliers de variants identifiés par GWAS. Même de petits biais par variant peuvent s'accumuler en erreurs considérables lorsqu'ils sont sommés sur l'ensemble du génome (voir la section 1.4.3 de l'introduction). La structure fine du SLSJ pourrait affecter l'exactitude des PRS pour divers traits [91], bien que des études spécifiques soient nécessaires pour quantifier cet impact dans notre population.

Il existe des préoccupations sur la transférabilité des PRS entre populations [92]. Nos résultats suggèrent que ces préoccupations pourraient s'appliquer non

seulement entre groupes continentaux, mais également au sein de populations régionales apparemment homogènes. La médecine de précision devra tenir compte de cette hétérogénéité locale pour éviter des prédictions de risque imprécises ou inéquitables [93, 94]. Bien que des études récentes suggèrent que la stratification intercontinentale pourrait avoir un effet moindre qu'anticipé sur certaines analyses génétiques [95], négliger la structure fine intrapopulationnelle pourrait néanmoins introduire des biais subtils. Un équilibre doit être trouvé entre la maximisation de la puissance statistique et le contrôle des faux positifs.

3.2.2 Retombées pour les maladies rares et le dépistage génétique

La notion de « gradient de Charlevoix » a des répercussions directes pour la prévention des maladies récessives. Nos résultats suggèrent que les enrichissements de variants rares pourraient être plus localisés qu'à l'échelle régionale : au sein de la population d'une seule région, le taux de porteurs pourrait varier d'une localité à une autre.

Si un variant est plus fréquent dans une partie de la région et moins dans une autre, considérer uniquement le taux de porteurs global peut atténuer l'évaluation de l'enrichissement et conduire à une sous-appréciation de la fréquence du variant dans certaines sous-régions. Ce facteur est amplifié si les tests de dépistage sont effectués de manière inégale, par exemple davantage dans les centres urbains que dans les municipalités rurales.

3.2.3 Considérations pour les études d'association futures

Détecter une structure fine aussi prononcée au sein d'une population relativement petite et « homogène » souligne le besoin critique de contrôler la stratification dans les GWAS et les scores polygéniques au-delà des 20 ou 100 premières composantes principales généralement utilisées. Une autre approche, les modèles linéaires mixtes, pallient ce besoin d'un grand nombre de composantes en se basant directement sur la similarité génétique entre les individus [96], de façon similaire au calcul de parenté de ce projet de maitrise. Toutefois, les approches basées sur les variants rares ou les segments IBD, comme suggéré par Zaidi et Mathieson [42], pourraient s'avérer plus appropriées pour capturer la structure fine récente, car elles ne se limitent pas aux variants communs. De plus, l'imputation de segments IBD serait moins couteuse que le séquençage complet pour inclure les variants rares et ultrarares. Une approche pourrait combiner l'utilisation de variants rares ou de segments IBD à l'application d'un modèle linéaire mixte.

3.3 Limites de l'étude

3.3.1 Biais d'échantillonnage de CARTaGENE

La cohorte CARTaGENE, bien que riche en données génotypiques, comprend principalement des individus recrutés dans les zones urbaines. Cette surreprésentation des centres urbains pourrait conduire à une sous-représentation des communautés plus isolées, là où la structure fine pourrait être la plus prononcée. Nos observations sur la plus grande

concentration des relations généalogiques dans les municipalités rurales suggèrent que l'inclusion de ces communautés dans les études génétiques futures serait pertinente.

3.3.2 Complétude variable des généalogies BALSAC

Les généalogies BALSAC, bien qu'exceptionnellement riches, présentent des variations régionales dans leur complétude [54]. Cette disponibilité inégale des données pourrait potentiellement introduire des biais dans la détection de la structure de population. Notre contrôle de qualité, en retirant les individus ayant plus de la moitié des coefficients de parenté nuls, a permis d'observer une structure très similaire (et une forte corrélation) au niveau de la génétique et de la généalogie de CARTaGENE, ce qui suggère que cette approche est suffisante pour être appliquée à la région du SLSJ.

3.3.3 Indicateur géographique et mobilité

L'utilisation du lieu de mariage des parents comme indicateur pour l'affiliation régionale, bien qu'elle soit meilleure que les lieux de recrutement de CARTaGENE, pourrait ne pas tenir pleinement compte de la mobilité individuelle ou familiale à travers les générations, ni des situations où les parents proviennent de régions différentes.

3.3.4 Validité temporelle de la structure observée

La structure observée reflète la population d'une des dernières générations disponibles dans BALSAC (mariages de 1931 à 1960), les données étant complètes jusqu'à environ 1965 [86]. La question se pose de savoir si cette structure demeure valide pour les

générations plus récentes, compte tenu de la mobilité accrue et de l'urbanisation. Le cas de la région de Lanaudière a été étudié à ce sujet [97].

Plusieurs changements sociodémographiques survenus depuis 1960 pourraient avoir atténué la structure fine que nous observons. Par exemple, si on remonte aux parents des individus mariés en 1931, on peut mentionner qu'aucun pont ne reliait les deux rives du Saguenay jusqu'à 1933 [98]. L'urbanisation et l'exode rural ont entrainé une concentration progressive de la population dans les centres urbains de Saguenay et d'Alma, ce qui a probablement augmenté la mixité entre individus d'origines géographiques différentes. Parallèlement, l'amélioration des transports, notamment le développement du réseau routier et l'accessibilité accrue de l'automobile, a réduit l'isolement des communautés périphériques et probablement facilité les unions entre individus de sous-régions différentes. Sur le plan socioéconomique, l'accès élargi à l'éducation et la diversification économique ont peut-être favorisé les déplacements au sein de la région et vers l'extérieur. Enfin, l'arrivée de nouveaux résidents d'autres provinces ou de l'étranger pourrait avoir introduit une diversité génétique additionnelle dans certaines localités.

Quelques approches permettraient de déterminer si la structure fine persiste dans les générations contemporaines. Bien que BALSAC ne couvre pas les mariages après 1965 et que les mariages se sont raréfiés après cette période, l'utilisation de données extraites à partir de publications telles que les nécrologies pourrait informer sur l'évolution

des flux migratoires intrarégionaux. L'extraction et l'analyse de données nécrologiques sont présentement investiguées par notre laboratoire.

3.3.5 Nature stochastique de UMAP

Les projections UMAP, malgré la nature stochastique inhérente de l'algorithme qui peut compliquer les comparaisons directes [62], ont produit des résultats hautement similaires avec les matrices de parenté attendue et réalisée. L'initialisation par PCoA classique a contribué à la stabilité et à la reproductibilité de nos projections.

3.4 Perspectives futures

3.4.1 Étude des variants pathogènes enrichis

L'étude des variants pathogènes enrichis au Québec gagnerait à inclure des informations sur la distribution géographique fine des variants, permettant une meilleure caractérisation de leur origine et de leur propagation. Jumelée à des données démographiques telles que le nombre d'enfants par ancêtre potentiellement porteur, ces informations permettraient également d'évaluer l'impact des variants potentiellement pathogènes sur la fécondité. Ces informations géographiques et historiques pourraient aider à décider si ces variants devraient être inclus dans des tests de porteurs.

3.4.2 Intégration de données de séquençage complet

L'intégration de données de séquençage du génome entier (WGS) permettrait d'examiner l'impact de la structure fine sur la distribution des variants ultrarares, ce qui

pourrait révéler des enrichissements encore plus localisés que ceux détectables avec les données de génotypage par puces. Toutefois, cette approche demeure couteuse à grande échelle.

3.4.3 Extension à d'autres régions du Québec

Les méthodes développées dans cette étude pourraient être appliquées à d'autres régions du Québec pour déterminer si des structures fines similaires existent. Les travaux récents de Gagnon et al. [71] sur la Beauce suggèrent que de telles structures sont probables.

3.4.4 Applicabilité internationale

La détection de structure fine au sein de populations apparemment homogènes n'est pas unique au Québec. Les travaux de Novembre et al. [99] ont démontré que la variation génétique en Europe reflète étroitement la géographie à un point tel qu'elle permet de prédire l'origine géographique des individus à partir de leurs données génétiques. Leslie et al. [100] ont ensuite révélé une structure génétique fine chez les Britanniques, identifiant des groupes distincts qui correspondent à des régions historiques et des vagues de migration anciennes. Ces études démontrent que la structure fine serait une caractéristique universelle des populations humaines, et non une particularité des populations à effet fondateur.

L'approche méthodologique développée dans cette étude est transférable à d'autres populations qui disposent de généalogies profondes, telles que l'Islande [101] et l'Utah aux États-Unis [102].

3.4.5 Retombées pour les études d'association futures

Nos résultats soulèvent des questions méthodologiques importantes pour la conception des études d'association génétique dans les populations à effet fondateur. D'abord, en ce qui concerne la stratification des échantillons, les études cas-témoins devraient idéalement apparier les cas et les témoins selon leur origine géographique fine, et non seulement selon leur appartenance régionale globale. Ensuite, la réplication des associations dans des sous-populations d'origines géographiques différentes au sein de la même région pourrait permettre de distinguer les vrais signaux biologiques des artéfacts de stratification, bien que cette approche séparée peut présenter une moins grande puissance statistique [95].

3.4.6 Intégration avec les données environnementales et de santé

La structure génétique fine du SLSJ pourrait interagir avec des facteurs environnementaux locaux pour influencer les phénotypes de santé. D'une part, les activités industrielles (aluminium, pâtes et papiers) n'ont pas été uniformément distribuées sur le territoire. Une corrélation entre l'origine géographique et l'exposition environnementale pourrait confondre les associations génétiques. D'autre part, les disparités dans l'accès aux services de santé entre les centres urbains et les communautés rurales pourraient interagir avec la

structure génétique pour influencer la prise en charge. L'intégration de données environnementales géolocalisées avec les données génétiques et généalogiques pourrait permettre des études d'interactions gène-environnement à une résolution géographique élevée.

Chapitre 4: Conclusion

Ce mémoire avait pour objectif de caractériser la structure fine de la population du Saguenay–Lac-Saint-Jean en intégrant des données génétiques et généalogiques à une échelle sans précédent. Nos résultats confirment notre hypothèse initiale : l'histoire démographique unique de cette région, documentée dans le fichier BALSAC, a engendré une structure génétique fine détectable jusqu'au niveau municipal.

La forte concordance observée entre la parenté attendue (généalogique) et réalisée (génétique), avec des corrélations de 0,78 à 0,83 selon le sous-ensemble analysé, valide l'utilisation des registres d'état civil comme outil de haute précision pour la génétique des populations. Cette validation a permis l'analyse de l'ensemble de la population du SLSJ, révélant un gradient est-ouest façonné par les contributions différentielles des fondateurs charlevoisiens et par des vagues migratoires successives.

Le développement de l'algorithme hybride implémenté dans GeneaKit a été déterminant pour atteindre ces objectifs. En permettant le calcul de plus de 3 milliards de coefficients de parenté en quelques minutes, cette innovation méthodologique ouvre la voie à l'analyse de populations entières, là où les méthodes antérieures étaient limitées à quelques milliers d'individus.

Ces découvertes ont des retombées importantes pour la santé publique et la médecine de précision. Elles remettent en question l'hypothèse d'homogénéité régionale et

soulignent le besoin de contrôler la stratification fine dans les GWAS et les analyses de scores polygéniques. Pour les programmes de dépistage génétique, la reconnaissance de cette hétérogénéité intrarégionale pourrait conduire à des approches plus ciblées et plus efficaces.

Au-delà du contexte québécois, notre étude démontre que la structure de population peut être détectée à des résolutions géographiques très fines, jusqu'au niveau des municipalités, où nous observons des gradients de contribution génétique de différents fondateurs ainsi que des groupements nettement définis pour les localités plus isolées. Il est raisonnable de penser qu'une telle structure fine, due à l'histoire récente, existe dans de nombreuses populations humaines.

Ce travail illustre que la généalogie s’avère être un outil puissant en génétique des populations. La structure fine des populations est une réalité universelle qui doit être intégrée dans les modèles statistiques de la génétique contemporaine. Les archives de l'état civil, témoins de siècles d'histoire familiale, sont des ressources inestimables pour comprendre non seulement notre passé démographique, mais aussi les répercussions présentes et futures de notre patrimoine génétique.

Bibliographie

1. Watson JD, Crick FH. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature.* 1953;171(4356):737–8.

2. Tjio JH. The chromosome number of man. *Am J Obstet Gynecol.* 1978;130:723–4.

3. Frazer KA, Murray SS, Schork NJ, Topol EJ. Human genetic variation and its contribution to complex traits. *Nat Rev Genet.* 2009;10:241–51.

4. Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, et al. Rate of de novo mutations and the importance of father's age to disease risk. *Nature.* 2012;488:471–5.

5. Lobo I, Shaw K. Thomas Hunt Morgan, genetic recombination and gene mapping. *Nature Education*. 2008;1(1):205.

6. Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, et al. A high-resolution recombination map of the human genome. *Nat Genet.* 2002;31:241–7.

7. Dawn Teare M, Barrett JH. Genetic linkage studies. *Lancet.* 2005;366(9490):1036–44.

8. Hill WG, Robertson A. Linkage disequilibrium in finite populations. *Theor Appl Genet*. 1968;38(6):226–31.

9. Slatkin M. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nat Rev Genet*. 2008;9:477–85.

10. Marchini J, Howie B. Genotype imputation for genome-wide association studies. *Nat Rev Genet.* 2010;11:499–511.

11. Powell JE, Visscher PM, Goddard ME. Reconciling the analysis of IBD and IBS in complex trait studies. *Nat Rev Genet*. 2010;11:800–5.

12. Thompson EA. Identity by descent: variation in meiosis, across genomes, and in populations. *Genetics*. 2013;194:301–26.

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

14. Okasha S. Population genetics. In: Zalta EN, Nodelman U, editors. *The Stanford Encyclopedia of Philosophy*. Summer 2024 ed. Metaphysics Research Lab, Stanford University; 2024.

15. Fisher RA. *The genetical theory of natural selection*. Oxford: The Clarendon Press; 1930.

16. Hartl DL, Clark AG. *Principles of population genetics*. 4th ed. Sunderland, Mass.: Sinauer Associates; 2007.

17. Kimura M. Evolutionary rate at the molecular level. *Nature*. 1968;217(5129):624–6.

18. Wright S. Evolution in Mendelian Populations. *Genetics*. 1931;16(2):97–159.

19. Slatkin M. Gene flow and the geographic structure of natural populations. *Science*. 1987;236(4803):787–92.

20. Mayr E. *Systematics and the origin of species from the viewpoint of a zoologist*. New York: Columbia University Press; 1942.

21. Risch N, Tang H, Katzenstein H, Ekstein J. Geographic distribution of disease mutations in the Ashkenazi Jewish population supports genetic drift over selection. *Am J Hum Genet*. 2003;72:812–22.

22. Amos W, Hoffman JI. Evidence that two main bottleneck events shaped modern human genetic diversity. *Proc Biol Sci*. 2010;277:131–7.

23. International HapMap Consortium. A haplotype map of the human genome. *Nature*. 2005;437(7063):1299–320.

24. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet*. 2007;39:906–13.

25. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, et al. A global reference for human genetic variation. *Nature*. 2015;526:68–74.

26. Uffelmann E, Huang QQ, Munung NS, de Vries J, Okada Y, Martin AR, et al. Genome-wide association studies. *Nature Reviews Methods Primers*. 2021;1(1).

27. Wright S. The genetical structure of populations. *Ann Eugen*. 1951;15(4):323–54.

28. Weir BS, Cockerham CC. Estimating F-Statistics for the Analysis of Population Structure. *Evolution*. 1984;38(6):1358–70.

29. Holsinger KE, Weir BS. Genetics in geographically structured populations: defining, estimating and interpreting F(ST). *Nat Rev Genet*. 2009;10:639–50.

30. Hellwege JN, Keaton JM, Giri A, Gao X, Velez Edwards DR, Edwards TL. Population Stratification in Genetic Association Studies. *Curr Protoc Hum Genet*. 2017;95:1.22.1–1.22.23.

31. Hamer D, Sirota L. Beware the chopsticks gene. *Mol Psychiatry*. 2000;5(1):11–3.

32. Knowler WC, Williams RC, Pettitt DJ, Steinberg AG. Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. *Am J Hum Genet*. 1988;43(4):520–6.

33. Cardon LR, Palmer LJ. Population stratification and spurious allelic association. *Lancet*. 2003;361:598–604.

34. Chatterjee N, Shi J, Garcia-Closas M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat Rev Genet*. 2016;17:392–406.

35. Gusev S. The Infinitesimal [Internet]. 2025 Mar 28. Available from: https://theinfinitesimal.substack.com/p/how-population-stratification-led

36. Sohail M, Maier RM, Ganna A, Bloemendal A, Martin AR, Turchin MC, et al. Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *Elife*. 2019;8.

37. Berg JJ, Harpak A, Sinnott-Armstrong N, Joergensen AM, Mostafavi H, Field Y, et al. Reduced signal for polygenic adaptation of height in UK Biobank. *Elife*. 2019;8.

38. Haworth S, Mitchell R, Corbin L, Wade KH, Dudding T, Budu-Aggrey A, et al. Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis. *Nat Commun*. 2019;10:333.

39. Devlin B, Roeder K. Genomic control for association studies. *Biometrics*. 1999;55(4):997–1004.

40. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet*. 2006;2:e190.

41. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 2006;38:904–9.

42. Zaidi AA, Mathieson I. Demographic history mediates the effect of stratification on polygenic scores. *Elife*. 2020;9.

43. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet*. 1993;52(3):506–16.

44. Curtin C. Do living people outnumber the dead? *Sci Am*. 2007;297(3):126.

45. Rohde DL, Olson S, Chang JT. Modelling the recent common ancestry of all living humans. *Nature*. 2004;431:562–6.

46. Wright S. Coefficients of Inbreeding and Relationship. *The American Naturalist*. 1922;56(645):330–8.

47. Malécot G. *Les mathématiques de l'hérédité*. Paris: Masson; 1948.

48. Cas index [En ligne]. Office québécois de la langue française. 2020. Disponible : https://vitrinelinguistique.oqlf.gouv.qc.ca/fiche-gdt/fiche/8870404/cas-index

49. Vézina H, Tremblay M, Desjardins B, Houde L. Origines et contributions génétiques des fondatrices et des fondateurs de la population québécoise. *Cahiers québécois de démographie*. 2005;34(2):235–58. Disponible : http://dx.doi.org/10.7202/014011ar

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

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

52. Burkett KM, Rakesh M, Morris P, Vezina H, Laprise C, Freeman EE, et al. Correspondence Between Genomic- and Genealogical/Coalescent-Based Inference of Homozygosity by Descent in Large French-Canadian Genealogies. *Front Genet*. 2021;12:808829.

53. Anderson-Trocme L, Nelson D, Zabad S, Diaz-Papkovich A, Kryukov I, Baya N, et al. On the genes, genealogies, and geographies of Quebec. *Science*. 2023;380(6647):849–55.

54. Gagnon L, Moreau C, Laprise C, Vezina H, Girard SL. Deciphering the genetic structure of the Quebec founder population using genealogies. *Eur J Hum Genet*. 2024;32:91–7.

55. Karigl G. A recursive algorithm for the calculation of identity coefficients. Ann *Hum Genet*. 1981;45(3):299–305.

56. Gauvin H, Lefebvre JF, Moreau C, Lavoie EM, Labuda D, Vezina H, et al. GENLIB: an R package for the analysis of genealogical data. *BMC Bioinformatics*. 2015;16:160.

57. Kirkpatrick B, Ge S, Wang L. Efficient computation of the kinship coefficients. *Bioinformatics*. 2019;35:1002–8.

58. Colleau JJ. An indirect approach to the extensive calculation of relationship coefficients. *Genet Sel Evol*. 2002;34:409–21.

59. Lee H, Craddock RF, Gorjanc G, Becher H. randPedPCA: rapid approximation of principal components from large pedigrees. *Genet Sel Evol*. 2025;57:46.

60. Gower JC. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*. 1966;53(3-4):325–38.

61. Anderson MJ, Willis TJ. Canonical Analysis of Principal Coordinates: A Useful Method of Constrained Ordination for Ecology. *Ecology*. 2003;84(2):511–25.

62. McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction. *arXiv*. 2018.

63. Diaz-Papkovich A, Anderson-Trocme L, Ben-Eghan C, Gravel S. UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. *PLoS Genet*. 2019;15:e1008432.

64. Kobak D, Linderman GC. Initialization is critical for preserving global data structure in both t-SNE and UMAP. *Nat Biotechnol*. 2021;39:156–7.

65. Hartigan JA, Wong MA. Algorithm AS 136: A K-Means Clustering Algorithm. *Applied Statistics*. 1979;28(1).

66. MacQueen J. Some methods for classification and analysis of multivariate observations. Dans : *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1: Statistics. University of California Press; 1967. p. 281–98.

67. Campello RJGB, Moulavi D, Sander J. Density-Based Clustering Based on Hierarchical Density Estimates. Dans : *Advances in Knowledge Discovery and Data Mining. Lecture Notes in Computer Science.* 2013. p. 160–72.

68. Gagnon A, Heyer E. Fragmentation of the Quebec population genetic pool (Canada): evidence from the genetic contribution of founders per region in the 17th and 18th centuries. *Am J Phys Anthropol*. 2001;114:30–41.

69. Moreau C, Bherer C, Vezina H, Jomphe M, Labuda D, Excoffier L. Deep human genealogies reveal a selective advantage to be on an expanding wave front. *Science*. 2011;334:1148–50.

70. Michel E, Moreau C, Gagnon L, Gagnon M, Leblanc J, Tardif J, et al. Rare diseases load through the study of a regional population. *PLoS Genet*. 2025;21:e1011876.

71. Gagnon M, Moreau C, Ricard J, Boisvert MC, Bureau A, Maziade M, et al. Rare variants and founder effect in the Beauce region of Quebec. *Commun Biol*. 2025;8:1184.

72. Gagnon S. Sagamiens – Sagamiennes. *Saguenayensia*. 1988;30(2).

73. Gauvreau D, Bourque M. Mouvements migratoires et familles : le peuplement du Saguenay avant 1911. *Revue de l'histoire de l'Amérique française (RHAF)*. 1988;42(2):167–92.

74. Pouyez C, Lavoie Y. *Les Saguenayens : introduction à l'histoire des populations du Saguenay, XVIe-XXe siècles*. Sillery: Presses de l'Université du Québec; 1983.

75. Bouchard G, De Braekeleer M. *Histoire d'un génôme : population et génétique dans l'est du Québec*. Sillery: Presses de l'Université du Québec; 1990.

76. De Braekeleer M, Larochelle J. Genetic epidemiology of hereditary tyrosinemia in Quebec and in Saguenay-Lac-St-Jean. *Am J Hum Genet*. 1990;47(2):302–7.

77. Bouchard JP, Barbeau A, Bouchard R, Bouchard RW. Autosomal recessive spastic ataxia of Charlevoix-Saguenay. *Can J Neurol Sci*. 1978;5(1):61–9.

78. Engert JC, Berube P, Mercier J, Dore C, Lepage P, Ge B, et al. ARSACS, a spastic ataxia common in northeastern Quebec, is caused by mutations in a new gene encoding an 11.5-kb ORF. *Nat Genet*. 2000;24(2):120–5.

79. Scriver CR. Human genetics: lessons from Quebec populations. *Annu Rev Genomics Hum Genet*. 2001;2:69–101.

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

81. Bchetnia M, Bouchard L, Mathieu J, Campeau PM, Morin C, Brisson D, et al. Genetic burden linked to founder effects in Saguenay-Lac-Saint-Jean illustrates the importance of genetic screening test availability. *J Med Genet*. 2021;58:653–65.

82. Cruz Marino T, Leblanc J, Pratte A, Tardif J, Thomas MJ, Fortin CA, et al. Portrait of autosomal recessive diseases in the French-Canadian founder population of Saguenay-Lac-Saint-Jean. *Am J Med Genet A*. 2023;191(5):1145–63.

83. Cruz Marino T, Tardif J, Leblanc J, Lavoie J, Morin P, Harvey M, 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*. 2022;141:607–22.

84. De Braekeleer M. L'approche des maladies héréditaires par la démographie génétique : le cas du Saguenay-Lac-Saint-Jean au Québec [thèse]. Université Bordeaux-II; 1995.

85. Lavoie EM, Tremblay M, Houde L, Vezina H. Demogenetic study of three populations within a region with strong founder effects. *Public Health Genomics*. 2005;8:152–60.

86. Vézina H, Bournival JS. An Overview of the BALSAC Population Database. Past Developments, Current State and Future Prospects. *Historical Life Course Studies*. 2020;9:114–29.

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

88. McClelland P, Femerling G, Laflamme R, Mejia-Garcia A, Dehkordi MS, Xiao H, et al. A multi-ancestry genetic reference for the Quebec population. *medRxiv*. 2025.

89. St-Hilaire M. La formation des populations urbaines au Québec : le cas du Saguenay aux XIXe et XXe siècles. *Cahiers québécois de démographie*. 2004;20(1):1–36.

90. Place des Acadiens - Saguenay (Ville) [En ligne]. Commission de toponymie du Québec. 2022. Disponible : https://toponymie.gouv.qc.ca/ct/ToposWeb/Fiche.aspx?no_seq=446160

91. Mostafavi H, Harpak A, Agarwal I, Conley D, Pritchard JK, Przeworski M. Variable prediction accuracy of polygenic scores within an ancestry group. *Elife*. 2020;9.

92. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet*. 2019;51:584–91.

93. Ding Y, Hou K, Xu Z, Pimplaskar A, Petter E, Boulier K, et al. Polygenic scoring accuracy varies across the genetic ancestry continuum. *Nature*. 2023;618:774–81.

94. Kachuri L, Chatterjee N, Hirbo J, Schaid DJ, Martin I, Kullo IJ, et al. Principles and methods for transferring polygenic risk scores across global populations. *Nat Rev Genet*. 2024;25:8–25.

95. Dias JA, Chen T, Xing H, Wang X, Rodriguez AA, Madduri RK, et al. Evaluating multi-ancestry genome-wide association methods: Statistical power, population structure, and practical implications. *Am J Hum Genet*. 2025;112(10):2493–508.

96. St-Pierre J, Oualkacha K, Bhatnagar SR, Schwartz R. Efficient penalized generalized linear mixed models for variable selection and genetic risk prediction in high-dimensional data. *Bioinformatics*. 2023;39(2).

97. Bherer C, Brais B, Vézina H. Impact des récentes transformations démographiques liées à l'urbanisation sur le bassin génétique de la région de Lanaudière. *Cahiers québécois de démographie*. 2009;37(2):211–35.

98. Pont de Sainte-Anne [En ligne]. Répertoire du patrimoine culturel du Québec. 2020. Disponible : https://www.patrimoine-culturel.gouv.qc.ca/detail.do?methode=consulter&id=232586&type=bien

99. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, et al. Genes mirror geography within Europe. *Nature*. 2008;456:98–101.

100. Leslie S, Winney B, Hellenthal G, Davison D, Boumertit A, Day T, et al. The fine-scale genetic structure of the British population. *Nature*. 2015;519(7543):309–14.

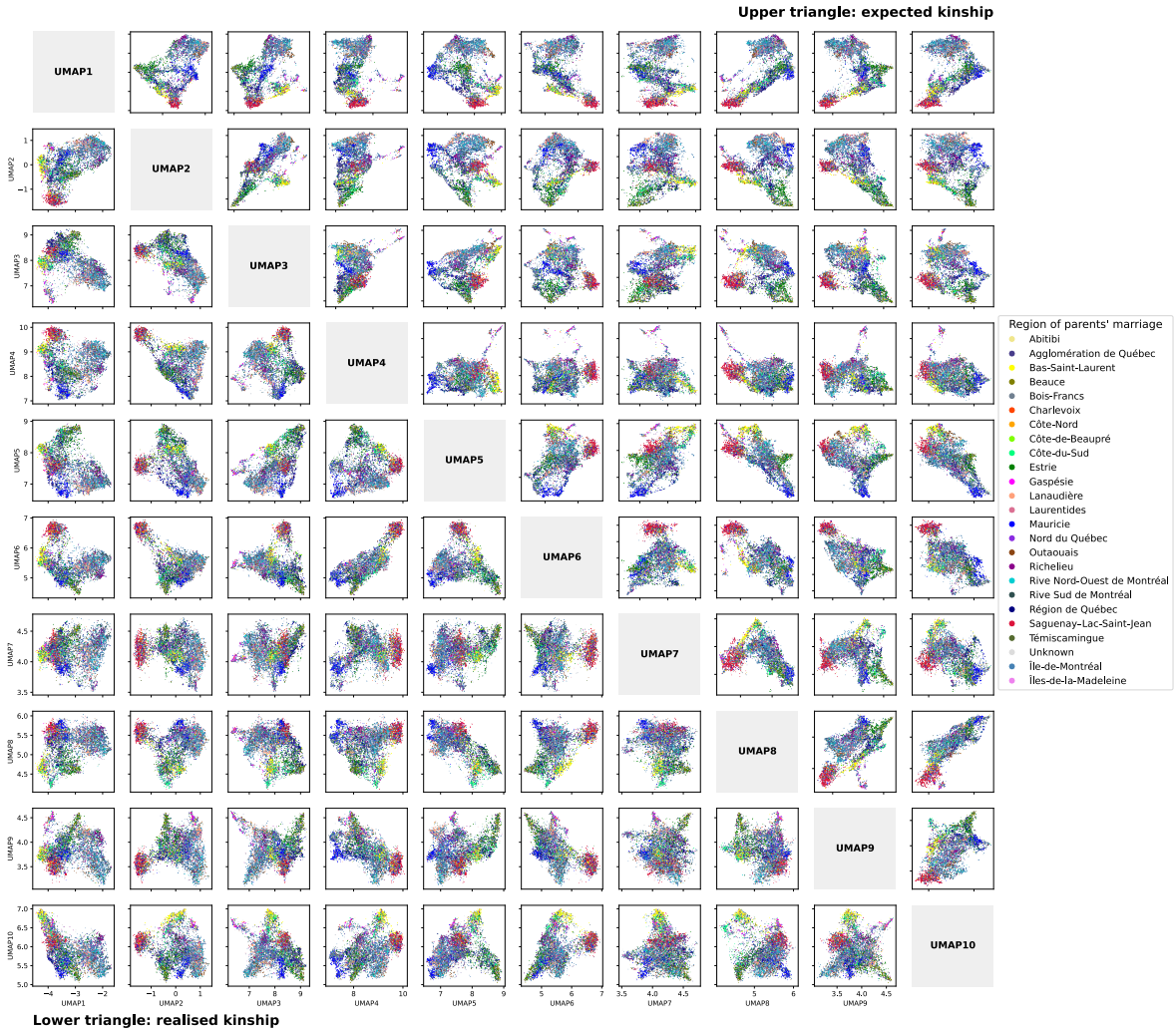
101. Gudmundsson H, Gudbjartsson DF, Frigge M, Gulcher JR, Stefansson K. Inheritance of human longevity in Iceland. *Eur J Hum Genet*. 2000;8(10):743–9.

102. Cannon-Albright LA. Utah family-based analysis: past, present and future. *Hum Hered*. 2008;65:209–20.

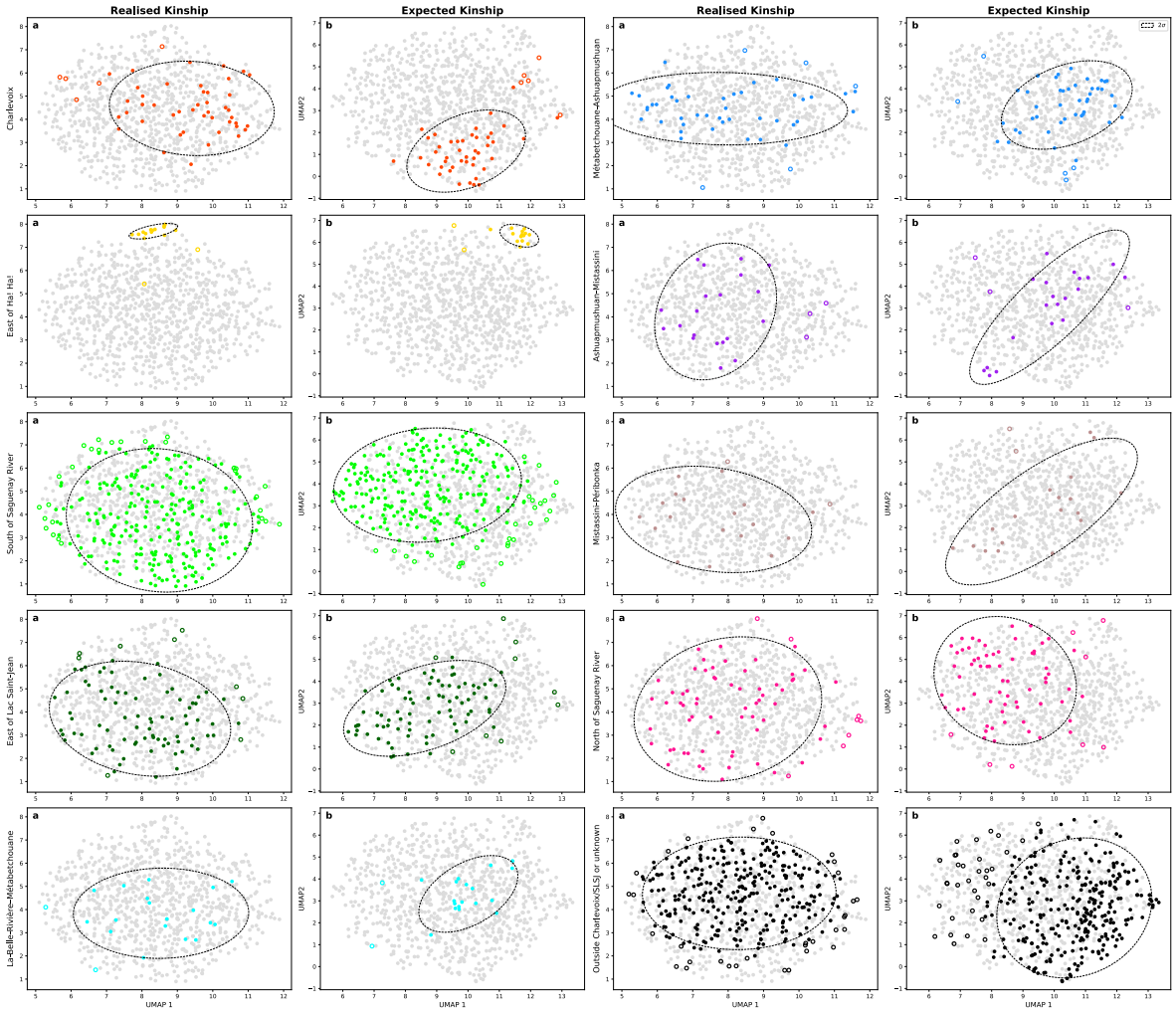
Certification éthique

Ce mémoire a fait l’objet d’une certification éthique auprès du Comité d'éthique de la recherche de l'Université du Québec à Chicoutimi (CER-UQAC). Le numéro du certificat est 2021-560.

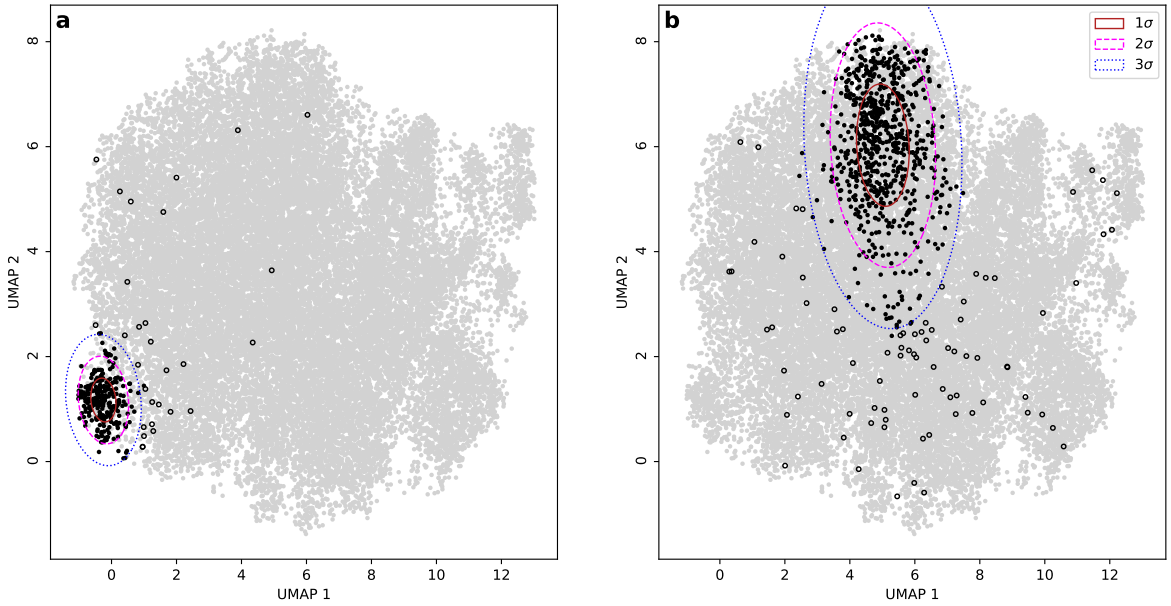
Annexe I
Fine-scale structure of a whole regional population through genetics and genealogies
Supplementary Material



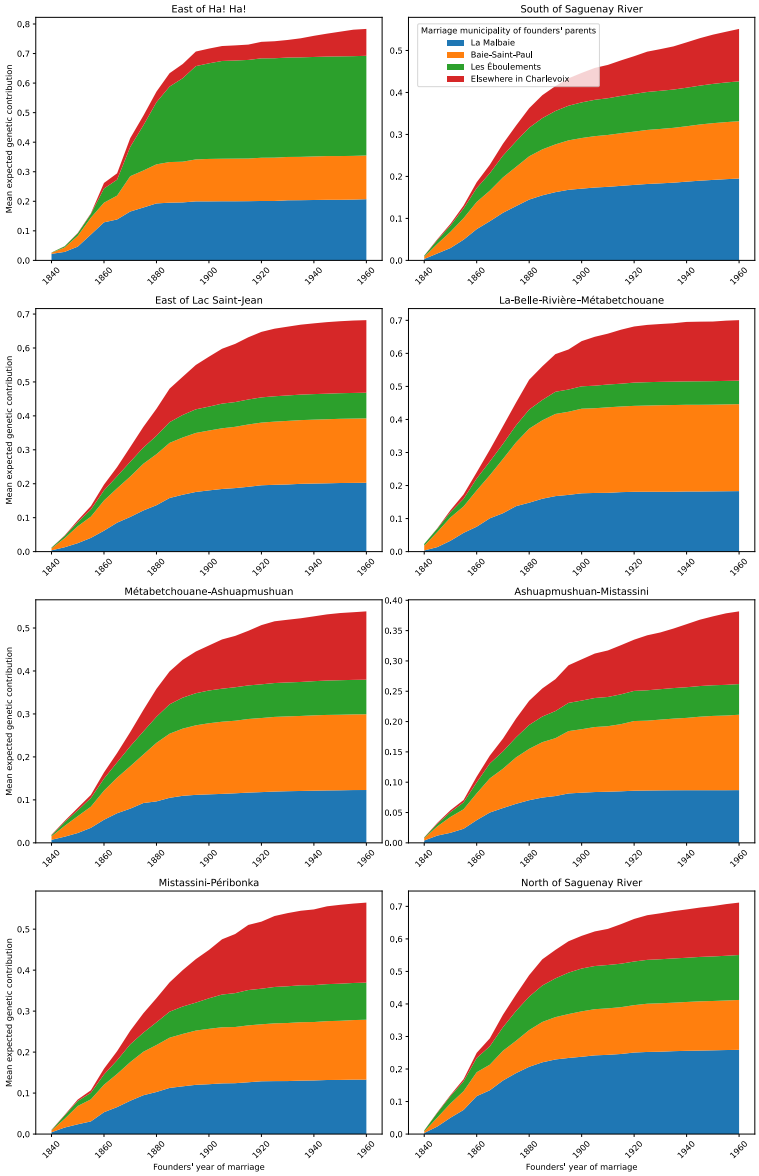
Supplementary Figure 1. **Ten-dimensional UMAP of CARTaGENE individuals based on pairwise realised (lower triangle) and expected (upper triangle) kinship.** Uniform Manifold Approximation and Projection (UMAP) projection of 7,970 individuals from the CARTaGENE cohort, computed from their (lower triangle) realised kinship and (upper triangle) expected kinship transformed as a precomputed distance ($1 - \phi$). The colours represent the region of parents' marriage.



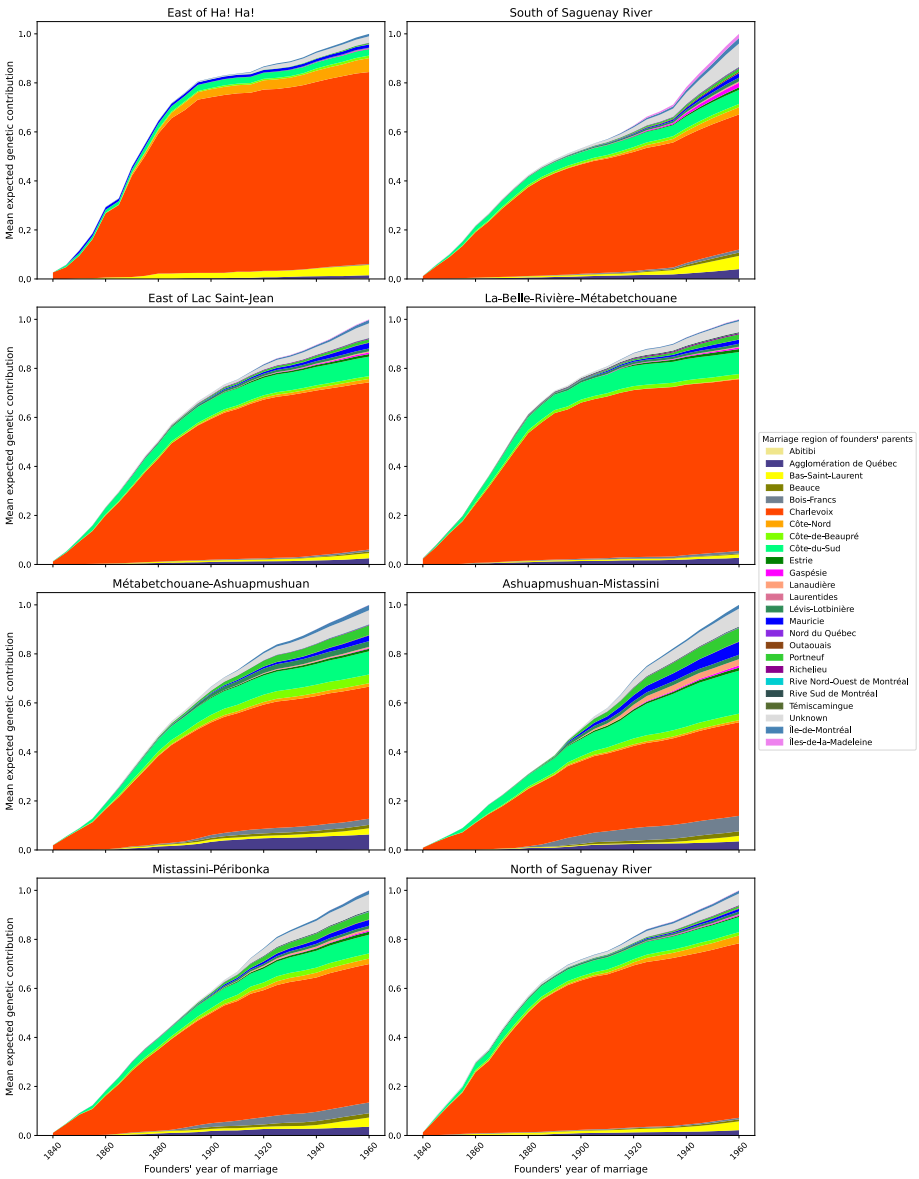
Supplementary Figure 2. **Distribution of CARTaGENE SLSJ individuals whose parents married in each watercourse subdivision.** Uniform Manifold Approximation and Projection (UMAP) projection of 938 individuals from the CARTaGENE cohort most likely originating from Saguenay–Lac-Saint-Jean (SLSJ), computed from their (**a**) realised kinship and (**b**) expected kinship transformed as a precomputed distance ($1 - \phi$). The coloured dots represent individuals whose parents married in the indicated subdivision of SLSJ. Open circles are individuals who were identified as outliers using an ellipse learned from their Gaussian distribution. The confidence ellipses of the remaining inliers have a radius of two standard deviations.



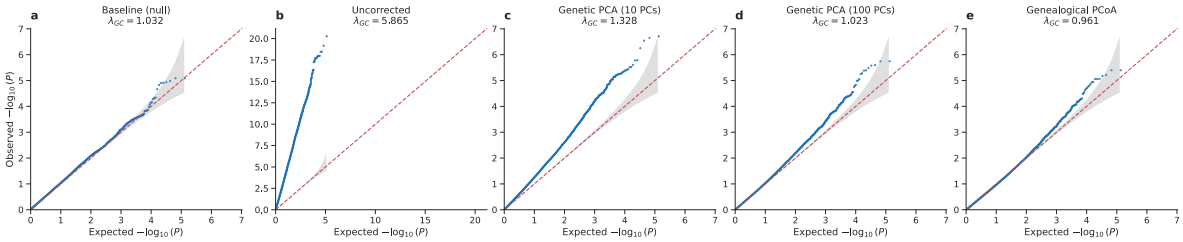
Supplementary Figure 3. **Difference in the spatial dispersion of individuals between a rural (a) and an urban (b) municipality.** Uniform Manifold Approximation and Projection (UMAP) projection of 26,445 non-siblings who married between 1931 and 1960 in Saguenay–Lac-Saint-Jean, computed from their expected kinship (ϕ) transformed as a precomputed distance ($1 - \phi$). Black dots represent individuals whose parents married in **(a)** L'Anse-Saint-Jean (n = 284); and **(b)** Alma (n = 805). Black open circles are individuals who were identified as outliers, whereas the confidence ellipses of the remaining inliers have a radius of one, two, and three standard deviations (σ).



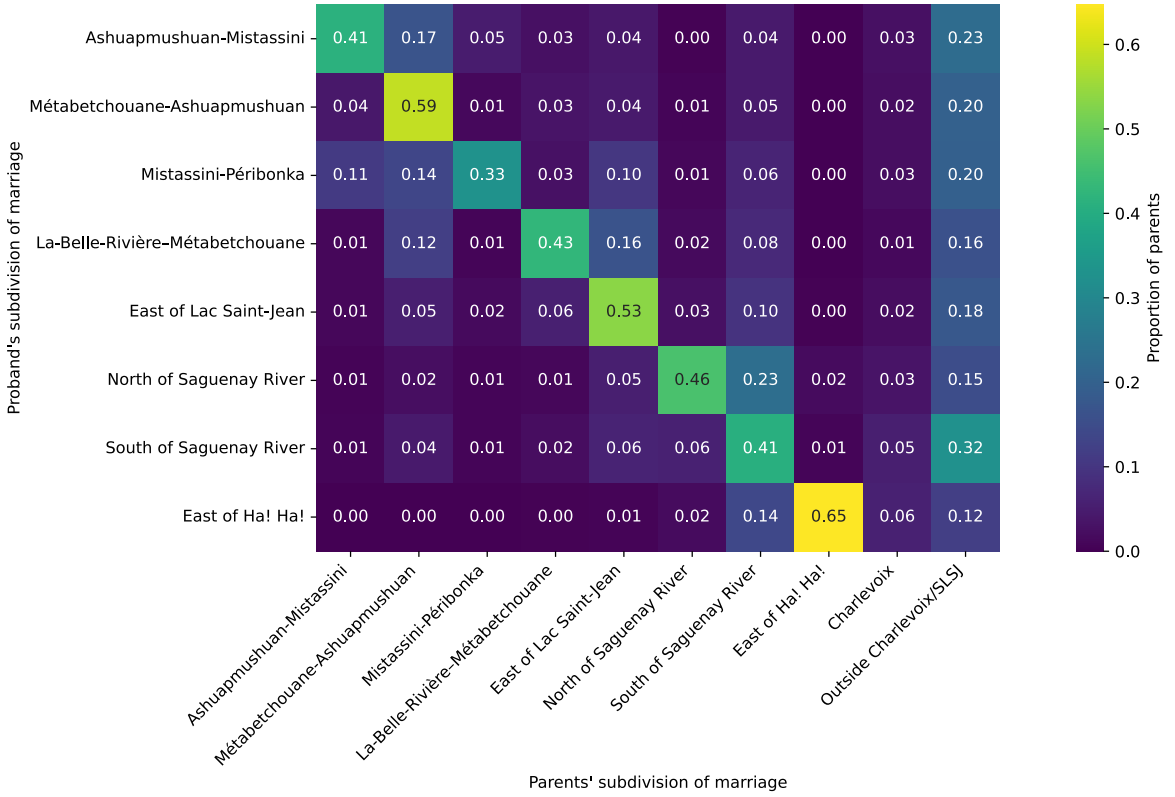
Supplementary Figure 4. **Cumulative genetic contribution of SLSJ founders originating from Charlevoix to probands in each watercourse subdivision per period of arrival.** Cumulative mean expected genetic contribution of Saguenay–Lac-Saint-Jean (SLSJ) founders per municipality of Charlevoix and period of marriage to probands (total 80,348) married in various subdivisions.



Supplementary Figure 5. **Cumulative genetic contribution of SLSJ founders originating from Quebec to probands in each watercourse subdivision per period of arrival.** Cumulative mean expected genetic contribution of Saguenay–Lac-Saint-Jean (SLSJ) founders per Quebec regions and period of marriage to probands (total 80,348) married in various SLSJ subdivisions.



Supplementary Figure 6. **PCA does not adequately control for inflation of GWAS on a simulated phenotype correlated with SLSJ fine-scale structure.** Quantile-quantile plots for simulated data (N = 10,000) across 100 Mb of neutral genomic sequence using `msprime` conditioned on the genealogy. The phenotype was simulated to be partially correlated with the expected genetic contribution (GC, computed on the genealogy) from Charlevoix founders (h^2 = 0.1) since this GC was shown to be a major factor influencing the Saguenay–Lac-Saint-Jean (SLSJ) fine structure. Panels display genome-wide association studies (GWAS) results for: **(a)** a baseline null phenotype (pure noise); **(b)** the correlated phenotype with no correction; **(c–d)** the correlated phenotype corrected using the top 10 and 100 genetic principal components (PCA) of common single nucleotide polymorphisms (minor allele frequency < 0.05); and **(e)** the correlated phenotype corrected using the top 2 coordinates from a principal coordinate analysis (PCoA) derived from the genealogical kinship matrix. The genomic inflation factor (λ_{GC}) is reported for each model to quantify test statistic inflation. The red dashed line represents the null expectation ($y = x$), and the gray shaded region indicates the 95% confidence interval. See Supplementary Methods for details.



Supplementary Figure 7. **Migration within and between subdivisions.** Proportion of parents married in each watercourse subdivision (column) for each SLSJ probands' (total 26,445) subdivision of marriage (row). The sum of each row is equal to one.

Supplementary Table 1. **Municipalities of watercourse-defined subdivisions of Saguenay–Lac-Saint-Jean (SLSJ).**

Subdivision (No of municipalities)	Municipalities (East to West)
East of Ha! Ha! (4)	Petit-Saguenay, L'Anse-Saint-Jean, Rivière-Éternité, Saint-Félix-d'Otis
South of Saguenay River (9)	Ferland-et-Boilleau, La Baie, Chicoutimi, Laterrière, Arvida, Jonquière, Saint-Ambroise, Bégin, Larouche
East of Lac Saint-Jean (13)	Mont-Apica, Notre-Dame-du-Rosaire, Labrecque, Saint-Nazaire, Delisle, Saint-Bruno, Alma, Hébertville-Station, L'Ascension-de-Notre-Seigneur, Hébertville, Saint-Gédéon, Saint-Henri-de-Taillon, Sainte-Monique-de-Honfleur
La-Belle-Rivière–Métabetchouane (4)	Lac-à-la-Croix, Métabetchouan, Desbiens, Saint-André-du-Lac-Saint-Jean
Métabetchouane-Ashuapmushuan (9)	Chambord, Saint-François-de-Sales, Lac-Bouchette, Mashteuiatsh, Roberval, Saint-Prime, Sainte-Hedwidge, Saint-Félicien, La Doré
Ashuapmushuan-Mistassini (7)	Dolbeau, Saint-Méthode, Albanel, Normandin, Girardville, Saint-Edmond, Saint-Thomas-Didyme
Mistassini-Péribonka (10)	Chute-des-Passes, Saint-Ludger-de-Milot, Saint-Augustin-du-Lac-Saint-Jean, Péribonka, Sainte-Élisabeth-de-Proulx, Sainte-Jeanne-d'Arc-du-Lac-Saint-Jean, Saint-Stanislas-du-Lac-Saint-Jean, Mistassini, Saint-Eugène-du-Lac-Saint-Jean, Notre-Dame-de-Lorette
North of Saguenay River (7)	Sainte-Rose-du-Nord, Saint-Fulgence, Chicoutimi-Nord, Saint-Honoré-de-Chicoutimi, Saint-David-de-Falardeau, Shipshaw, Saint-Charles-de-Bourget

Supplementary Methods

Simulating GWAS stratification using the BALSAC SLSJ genealogy. To assess the impact of population stratification driven by a fine structure induced by a regional founder effect on a genome-wide association study (GWAS), we performed a coalescent simulation using 10,000 probands married between 1931–1960 sampled from the BALSAC genealogy of Saguenay–Lac-Saint-Jean (SLSJ). Probands were randomly selected among 16,269 probands related less than first cousins. We generated 100 Mb of neutral genomic sequence using `msprime`[1,2] 1.3.4's `FixedPedigree` model to track ancestry through the known genealogy, recapitated the simulation with the OutOfAfrica_3G09 demographic model[3] via `stdpopsim`[4,5,6] 0.3.0 to provide realistic ancestral diversity, and applied standard human mutation (1.2×10^{-8})[7] and recombination (1.1×10^{-8})[8] rates. A correlated phenotype was constructed based on the expected genetic contribution from founders originating in the Charlevoix region, standardized and mixed with random noise ($h^2 = 0.1$), using the following formula:

$$y \sim \mathrm{GC} \times \sqrt{h^2} + N(0,1) \times \sqrt{1 - h^2}$$

Following filtering for minor allele frequency (MAF > 0.01), we performed association testing using linear regression under four conditions: uncorrected, corrected with 10 or 100 genetic principal components (PCA) of common single nucleotide polymorphisms (MAF > 0.05) pruned for linkage disequilibrium (window = 50 variants, sliding window = 5 variants, $r^2 = 0.2$), and corrected with 2 genealogical principal coordinates (PCoA) derived from the genealogical expected kinship matrix, evaluating model performance via the genomic inflation factor (λ_{GC})[9] and quantile-quantile plots.

Supplementary Note 1

Pseudocode for hybrid kinship algorithm. In the C++ implementation of this algorithm, *genealogy* is a hash map of a data structure named *individual*, referenced through a unique integer ID. The genealogy is sorted so that any ancestor appears before their offspring. Each *individual* contains notably a reference to two other *individual* structures, a father (if present in the genealogy) and a mother (if present). It also possesses an immutable *rank* from 1 to *n* individuals in the genealogy, which indicates the individual's order in the hash map. Finally, its mutable *founder_index* indicates whether the individual is currently a founder (if non null) and, if that's the case, where their kinship values are located in the current founder kinship matrix. The function compute_kinships(*genealogy, proband_IDs*) is called using the hash map and a list of the probands' integer IDs, and returns a kinship matrix with the probands in the same order as in the list. The function compute_kinship_between_probands may run the nested for loop in parallel, for each generation.

```
FUNCTION get_previous_generation(genealogy, current_individuals):
    CREATE a new empty set called 'previous_generation'
    FOR each ID in 'current_individuals':
        GET the person from 'genealogy' using the ID
        IF person has a father:
            ADD father's ID to 'previous_generation'
        IF person has a mother:
            ADD mother's ID to 'previous_generation'
    RETURN 'previous_generation'

FUNCTION get_generations(genealogy, starting_probands):
    CREATE an empty list of sets called 'generations'
    CREATE an empty set called 'current_generation'
    FOR each ID in starting_probands:
        ADD ID to 'current_generation'
    WHILE 'current_generation' is not empty:
        ADD 'current_generation' to 'generations'
        SET 'current_generation' to the result of calling
get_previous_generation(genealogy, current_generation)
    RETURN 'generations'

FUNCTION copy_bottom_up(generations):
    CREATE an empty list of sets called 'bottom_up'
    CREATE an empty set called 'first_generation'
    FOR each ID in the first set of 'generations':
        ADD ID to 'first_generation'
    ADD 'first_generation' to 'bottom_up'
    FOR each index 'i' from the first to the second-to-last set of
'generations':
        CREATE an empty set called 'combined_generation'
        CREATE a set called 'previous_generation' from the individuals in
bottom_up[i]
        CREATE an empty set called 'next_generation'
        FOR each ID in generations[i + 1]:
            ADD ID to 'next_generation'
```

```
        PERFORM a set union of 'previous_generation' and
'next_generation', adding the unique results to 'combined_generation'
        ADD 'combined_generation' to 'bottom_up'
    REVERSE 'bottom_up' to go from oldest to youngest
    RETURN 'bottom_up'

FUNCTION copy_top_down(generations):
    REVERSE the order of 'generations'
    CREATE an empty list of sets called 'top_down'
    CREATE an empty list called 'first_generation'
    FOR each ID in the first set of the (now reversed) 'generations':
        ADD ID to 'first_generation'
    ADD 'first_generation' to 'top_down'
    FOR each index 'i' from the first to the second-to-last set of
'generations':
        CREATE an empty set called 'combined_generation'
        CREATE an empty set called 'previous_generation'
        FOR each ID in top_down[i]:
            ADD ID to 'previous_generation'
        CREATE an empty set called 'next_generation'
        FOR each ID in generations[i + 1]:
            ADD ID to 'next_generation'
        PERFORM a set union of 'previous_generation' and
'next_generation', adding the unique results to 'combined_generation'
        ADD 'combined_generation' to 'top_down'
    RETURN 'top_down'

FUNCTION intersect_both_directions(bottom_up, top_down):
    CREATE an empty list of lists called 'vertex_cuts'
    FOR each index 'i' for every list in 'bottom_up':
        CREATE an empty list called 'current_vertex_cut'
        CREATE a set called 'set_from_bottom_up' from the IDs in
bottom_up[i]
        CREATE a set called 'set_from_top_down' from the IDs in
top_down[i]
        CREATE an empty set called 'common_individuals'
        PERFORM a set intersection of 'set_from_bottom_up' and
'set_from_top_down', adding the unique results to 'common_individuals'
        FOR each ID in 'common_individuals':
            ADD ID to 'current_vertex_cut'
        ADD 'current_vertex_cut' to 'vertex_cuts'
    RETURN 'vertex_cuts'

FUNCTION cut_vertices(genealogy, proband_IDs):
    CREATE an empty list of sets called 'generations'
    CREATE an empty list of lists called 'vertex_cuts'
    SET 'generations' to the result of calling get_generations(genealogy,
proband_IDs)
    CREATE empty lists of sets called 'bottom_up' and 'top_down'
    SET 'bottom_up' to the result of calling copy_bottom_up(generations)
    SET 'top_down' to the result of calling copy_top_down(generations)
    SET 'vertex_cuts' to the result of calling
intersect_both_directions(bottom_up, top_down)
    SET the last list in 'vertex_cuts' to 'proband_IDs'
    RETURN 'vertex_cuts'

FUNCTION compute_kinship(ind1, ind2, founder_matrix):
    SET 'kinship' to 0.0
```

```
    SET 'founder_index1' to individual ind1's founder_index
    SET 'founder_index2' to individual ind2's founder_index
    IF 'founder_index1' is not null AND 'founder_index2' is not null:
        SET 'kinship' to founder_matrix[founder_index1][founder_index2]
    ELSE IF 'founder_index1' is not null:
        IF ind2 has a father:
            ADD 0.5 * compute_kinship(ind1, ind2's father,
founder_matrix) to 'kinship'
        IF ind2 has a mother:
            ADD 0.5 * compute_kinship(ind1, ind2's mother,
founder_matrix) to 'kinship'
    ELSE IF 'founder_index2' is not null:
        IF ind1 has a father:
            ADD 0.5 * compute_kinship(ind1's father, ind2,
founder_matrix) to 'kinship'
        IF ind1 has a mother:
            ADD 0.5 * compute_kinship(ind1's mother, ind2,
founder_matrix) to 'kinship'
    ELSE IF ind1's rank is equal to ind2's rank:
        SET 'kinship' to 0.5
        IF ind1 has a father AND ind2 has a mother:
            ADD 0.5 * compute_kinship(ind1's father, ind2's mother,
founder_matrix) to 'kinship'
    ELSE IF ind1's rank is less than ind2's rank:
        IF ind2 has a father:
            ADD 0.5 * compute_kinship(ind1, ind2's father,
founder_matrix) to 'kinship'
        IF ind2 has a mother:
            ADD 0.5 * compute_kinship(ind1, ind2's mother,
founder_matrix) to 'kinship'
    ELSE:
        IF ind1 has a father:
            ADD 0.5 * compute_kinship(ind1's father, ind2,
founder_matrix) to 'kinship'
        IF ind1 has a mother:
            ADD 0.5 * compute_kinship(ind1's mother, ind2,
founder_matrix) to 'kinship'
    RETURN 'kinship'

FUNCTION compute_kinship_with_oneself(probands, founder_matrix,
proband_matrix):
    FOR each index 'i' for every proband in 'probands':
        SET 'proband' to probands[i]
        SET 'kinship' to 0.5
        IF proband has a father AND proband has a mother:
            ADD 0.5 * compute_kinship(proband's father, proband's mother,
founder_matrix) to 'kinship'
        SET proband_matrix[i][i] to 'kinship'

FUNCTION compute_kinship_between_probands(probands, founder_matrix,
proband_matrix):
    FOR each index 'i' for every proband in 'probands':
        SET 'proband1' to probands[i]
        FOR each index 'j' from the first proband up to (but not
including) the current index 'i':
            SET 'proband2' to probands[j]
            SET 'kinship' to compute_kinship(proband1, proband2,
founder_matrix)
```

```
            SET proband_matrix[i][j] to 'kinship'
            SET proband_matrix[j][i] to 'kinship'

FUNCTION compute_kinships(genealogy, proband_IDs):
    SET 'vertex_cuts' to the result of calling cut_vertices(genealogy,
proband_IDs)
    CREATE a new matrix called 'current_founder_matrix' with dimensions
(size of vertex_cuts[1], size of vertex_cuts[1]) and filled with zeros
    FOR each index 'i' for every individual in the first vertex cut:
        SET current_founder_matrix[i][i] to 0.5
    FOR each index 'i' from the first to the second-to-last vertex cut:
        SET 'founder_index' to 1
        FOR each person ID in vertex_cuts[i]:
            GET the person from the genealogy using their ID
                SET person's founder index (null by default) to
'founder_index'
            INCREMENT 'founder_index'
        CREATE a new matrix called 'current_proband_matrix' with
dimensions (size of vertex_cuts[i + 1], size of vertex_cuts[i + 1])
        CREATE an empty list of persons called 'current_probands'
        FOR each person ID in vertex_cuts[i + 1]:
            ADD 'person' from 'genealogy' using their ID to
'current_probands'
        CALL compute_kinship_with_oneself(current_probands,
current_founder_matrix, current_proband_matrix)
        CALL compute_kinship_between_probands(current_probands,
current_founder_matrix, current_proband_matrix)
        SET 'current_founder_matrix' to 'current_proband_matrix'
    RETURN 'current_founder_matrix'
```

Time complexity of the hybrid kinship algorithm. The preprocessing phase, which identifies vertex cuts through ancestral traversal and set operations, has $O(n)$ time complexity where n is the total number of individuals in the genealogy. The kinship computation phase has $O(\Sigma\, v_i^2)$ time complexity, where v_i represents the number of individuals in vertex cut i and the sum is taken over all g vertex cuts. Since kinship coefficients are computed for all pairs within each vertex cut and memoised in founder matrices, subsequent recursive calls achieve $O(1)$ amortised lookup time. The overall algorithm complexity is dominated by $O(\Sigma\, v_i^2)$, which can be approximated as $O(g \times \bar{v}^2)$ or $O(n \times \bar{v})$ where $\bar{v}$ is the mean vertex cut size.

Supplementary References

1. Baumdicker, F. *et al.* Efficient ancestry and mutation simulation with msprime 1.0. *Genetics* **220**, iyab229 (2022).

2. Kelleher, J., Etheridge, A. M. & McVean, G. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Comput. Biol.* **12**, e1004842 (2016).

3. Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H. & Bustamante, C. D. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* **5**, e1000695 (2009).

4. Adrion, J. R. *et al.* A community-maintained standard library of population genetic models. *eLife* **9**, e54967 (2020).

5. Lauterbur, M. E. *et al.* Expanding the stdpopsim species catalog, and lessons learned for realistic genome simulations. *eLife* **12**, RP84874 (2023).

6. Gower, G. *et al.* Accessible, realistic genome simulation with selection using stdpopsim. *Mol. Biol. Evol.* **42**, msaf236 (2025).

7. Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471–475 (2012).

8. Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nat. Genet.* **31**, 241–247 (2002).

9. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).