Gènes, généalogies, et géographies du Québec

Luke Anderson-Trocmé[1,2], Dominic Nelson[1,2], Shadi Zabad[3], Alex Diaz-Papkovich[1,4], Nikolas Baya[5], Mathilde Touvier[6], Ben Jeffery[5], Ivan Kryukov[1,2], Christian Dina[7], Hélène Vézina[8], Jerome Kelleher[5], et Simon Gravel[1,2,*]

[1]Départment de Génétique Humaine, Université McGill; Montréal, Canada
[2]Centre de Génomique de l'Université Mcgill; Montréal, Canada
[3]École d'informatique, Université McGill; Montréal, Canada
[4]Sciences Quantitatives du Vivant, Université McGill; Montréal, Canada
[5]Big Data Institute, Li Ka Shing Centre for Health Information and Discovery; University of Oxford, Oxford, UK
[6]Université Sorbonne Paris Nord, INSERM U1153, INRAE U1125, CNAM, Équipe de recherche en Épidémiologie Nutritionelle, Centre de Recherche en Épidemiologie et Statistique, Université Paris Cité (CRESS); Bobigny, France
[7]Nantes Université, CNRS, INSERM, l'institut du thorax, F-44000; Nantes, France
[8]Projet BALSAC, Université du Québec à Chicoutimi; Chicoutimi, Canada
[*]Adresser les correspondances à : simon.gravel@mcgill.ca

Résumé:

La génétique des populations se base habituellement sur des modèles très simplifiés. En utilisant une généalogie comprenant des informations tirées de 4 millions d'actes de l'état civil du Québec et des données génotypiques de 2 276 Français et 20 451 Québécois d'ascendance française, nous avons construit un modèle détaillé de l'ascendance génétique de ces derniers au fil du temps et dans l'espace. La disparition de la structure ancestrale française et l'apparition d'une structure spatiale au Québec montrent des traits caractéristiques de plusieurs modèles d'expansion classiques en génétique des populations. La topographie a influencé les migrations sur tout le territoire et nous observons une augmentation des migrations ainsi que de l'apparentement génétique et généalogique à l'intérieur des bassins versants. Finalement, nous rendons accessible un jeu de données incluant des génomes entiers simulés et des métadonnées spatiotemporelles pour 1 426 749 individus reflétant la riche histoire démographique du Québec. Ce type de simulations ouvre la porte à la réalisation d'analyses de génétique des populations à une résolution sans précédent.

En une phrase: Nous présentons un modèle précis et à haute résolution de la variation génétique dans une population fondatrice.

L'histoire génétique humaine est formée de lignées ancestrales entrelacées par des origines communes et la recombinaison (*1*). Elle a été tissée au fil des migrations individuelles dont les grandes tendances peuvent parfois être reconstruites par des analyses génomiques (*2–4*). La relation, complexe, entre les migrations historiques et la variation génétique pose encore des défis considérables (*5, 6*). La dispersion spatiale limitée des individus mène, en règle générale, à un isolement par distance et à une corrélation parfois frappante entre les distances génétique et géographique (*7–9*). Cependant, dans une région donnée, des événements historiques ponctuels ou des barrières géographiques façonnent également la variation de la population (par exemple, (*5, 10, 11*)). L'interprétation d'aspects ponctuels de la variation continue en termes d'unités évolutives distinctes s'est avérée difficile et parfois trompeuse (*12–14*). Bien que de nombreuses études aient envisagé des modèles de migration anisotropes (*15*), et même des modèles détaillés de contrainte ou de 'résistance' géographique (*16, 17*), la comparaison de ces modèles avec les données génétiques est difficile.

Pour étudier les liens entre génétique et géographie, nous tirons profit d'une généalogie de la population québécoise compilée à partir de plus de quatre millions de documents provenant principalement de registres paroissiaux Catholiques (*52*). La numérisation des lieux et dates de mariages permet de retracer l'ascendance génétique dans l'espace et le temps. En liant ces données à des données génotypiques de $20\,451$ individus et à de nouveaux outils de simulation, nous proposons un modèle spatiotemporel de la variation génétique à des échelles allant de quelques dizaines à quelques milliers de kilomètres. En incluant des individus de France et d'Angleterre, nous mesurons le degré auquel la structure présente dans ces deux populations contribue à la différentiation entre les régions du Québec. Nous étudions les profils de variation génétique le long des cours d'eau, puisque les derniers quatre cent ans d'histoire coloniale européenne ont été marqués par une expansion territoriale commençant sur les rives du Saint-Laurent et éventuellement le long de ses affluents. En retraçant l'ascendance généalogique de millions d'individus, nous décrivons une myriade d'effets fondateurs le long de cours d'eau et de formations géologiques qui ont défini le transport et l'activité économique. Cette étude permet

donc de rapprocher, d'une part, les modèles de génétique à l'échelle des familles et à l'échelle des populations et, d'autre part, les modèles théoriques et l'observation empirique de la variation génétique.

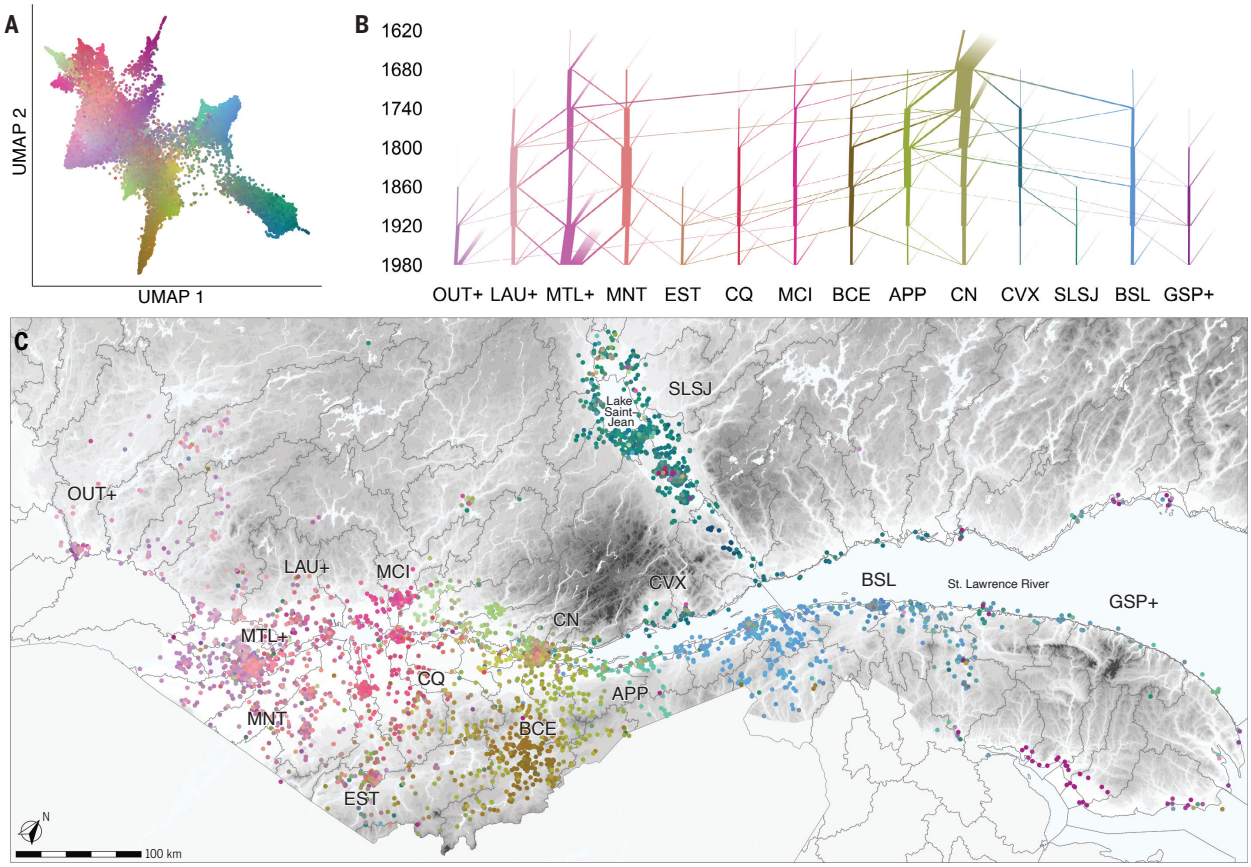

Figure 1: **Les gènes et généalogies des Franco-Québécois reflètent la géographie du Québec.** (**A**) UMAP pour 20 451 individus avec une ascendance franco-canadienne inférée. Une couleur est assignée à chaque individu en fonction de son emplacement dans une projection UMAP en trois dimensions (voir (*24*)). (**B**) Visualisation de l'ascendance des Franco-Québécois à travers les régions (axe x) ordonnée par l'année de mariage (axe y). L'épaisseur de la ligne au temps *t*, allant de l'emplacement *A* à l'emplacement *B*, représente la quantité attendue de matériel génétique ancestral pour la population actuelle, estimée à partir des données généalogiques. Les lignes estompées indiquent les fondateurs généalogiques (c'est-à-dire les individus sans parents documentés). La ligne estompée épaisse en haut indique l'arrivée des colons français à Québec dans la région de la CN. (**C**) Les regroupements dans l'espace génétique coïncident avec des régions du Québec, souvent définies par des caractéristiques géographiques telles que rivières et montagnes. Les limites des bassins versants sont indiquées en noir. Abréviations : OUT – Outaouais+ ; LAU – Laurentides+ ; MTL – Montréal+ ; MNT – Montérégie ; EST – Estrie ; CQ – Centre-du-Québec ; MCI – Mauricie ; BCE – Beauce ; APP – Appalaches ; CN – Capitale-Nationale ; CVX – Charlevoix ; SLSJ– Saguenay–Lac-Saint-Jean; BSL – Bas-Saint-Laurent; GSP – Gaspésie+. Voir aussi tableau S1 et figure S22.

Résultats

Distribution régionale de la variation génétique

Le Québec a une population de $8{,}8$ millions d'individus dont environ $6{,}5$ millions ont le français comme langue maternelle. L'ascendance de la plupart des Québécois remonte à $8\,500$ colons ayant immigré de France aux 17$^{\text{ème}}$ et 18$^{\text{ème}}$ siècles – nous utiliserons le vocable Franco-Québécois pour désigner les individus ayant une telle ascendance. Les premiers $2\,600$ colons français ont contribué les deux tiers du bassin génétique franco-québécois (*18*). Les colons français ont occupé un territoire habité et utilisé par les Premières Nations et les Inuits depuis des milliers d'années (*19*). Malgré des croyances populaires impliquant les origines métissées des Canadiens-Français (*20*), les études génétiques et généalogiques montrent que les Franco-Québécois portent en moyenne moins de 1% d'ascendance génétique autochtone et une majorité d'ascendance française (*21*).

Étant donné la forte corrélation entre ascendance française et religion Catholique, l'arbre généalogique reconstruit par le projet BALSAC à partir de registres paroissiaux est particulièrement complet chez les Franco-Québécois (*22*). Nous avons donc porté notre attention sur $20\,451$ individus dont nous inférons une ascendance franco-québécoise parmi les cohortes CARTaGENE ($12\,064$ (*23*)) et Genizon ($9\,004$) (voir (*24*) pour les détails des cohortes et les méthodes d'inférences de l'ascendance). Nous avons utilisé l'analyse par composante principale (ACP, fig. S1) et l'uniform manifold approximation and projection (UMAP, fig. 1A) (*25, 26*) pour visualiser la variation génétique. Les figures 1B et C montrent la répartition spatiale, à l'échelle régionale, de l'arbre généalogique franco-québécois d'après le lieu de mariage de $4\,882$ individus liés à cet arbre généalogique. Une inspection visuelle montre de fortes corrélations entre similarité génétique et proximité spatiale, et suggère que les gradients de variation génétique coincident avec les barrières géographiques et les voies navigables comme le Saint-Laurent, le Saguenay, et la Chaudière ainsi que les chaînes de montagnes des Laurentiennes et Appalaches.

Une ascendance française déracinée

Des études généalogiques ont montré que la plupart des premiers colons français sont arrivés à Québec, souvent en provenance de l'ouest (Aunis, Poitou) et du nord-ouest (Normadie, Perche) de la France, ainsi que de l'Ile-de-France [voir (*27, 28*) et tableau S2]. Les vagues successives de migration, motivées par des objectifs militaires et de colonisation, avaient des profils démographiques variables (*28*). Pour vérifier si cette variabilité se reflète dans la structure actuelle de la population franco-québécoise, nous avons comparé les génomes des individus vivant dans différentes régions du Québec et de la France. En accord avec les études historiques, les franco-québécois partagent plus d'ascendance génétique, mesurée par l'identité par ascendance (IBD) avec les individus de l'Ouest de la France (fig. S2A). Nous observons aussi une modeste variation entre les régions du Québec dans leur apparentement par IBD avec la France; les régions du Saguenay-Lac-Saint-Jean, de la Beauce, et du Bas-Saint-Laurent ont un apparentement légèrement supérieur, sans doute une conséquence d'effets fondateurs (voir fig S2B et discussions dans (*24*)). À l'aide des statistiques F_4, nous avons vérifié si la différentiation génétique entre régions du Québec était corrélée avec la differentiation entre des régions européennes (sept régions françaises et la Grande-Bretagne). Nous n'avons trouvé aucune corrélation (fig. S3), ce qui suggère que la plupart de la structure présente chez les Franco-Québécois, à grande échelle, est indépendante de la structure ancestrale ou d'une contribution différenciée parmi les régions européennes étudiées.

Génomes simulés en accord avec la généalogie

Pour vérifier à quel point la structure dans la population franco-québécoise peut être modélisée à partir des événements suivant la colonisation, nous avons généralisé le logiciel `msprime` (*29–31*) afin de simuler des génomes par coalescence conditionnelle à l'arbre généalogique disponible pour la population. Le nouveau modèle de simulation `FixedPedigree` de `msprime` version 1.2 prend en compte les événements de coalescence et recombinaison dans la généalogie. Pour prendre en compte l'apparentement entre fondateurs généalogiques, soit les individus dont les parents ne sont pas documentés, nous continuons la simulation de coalescence selon un modèle

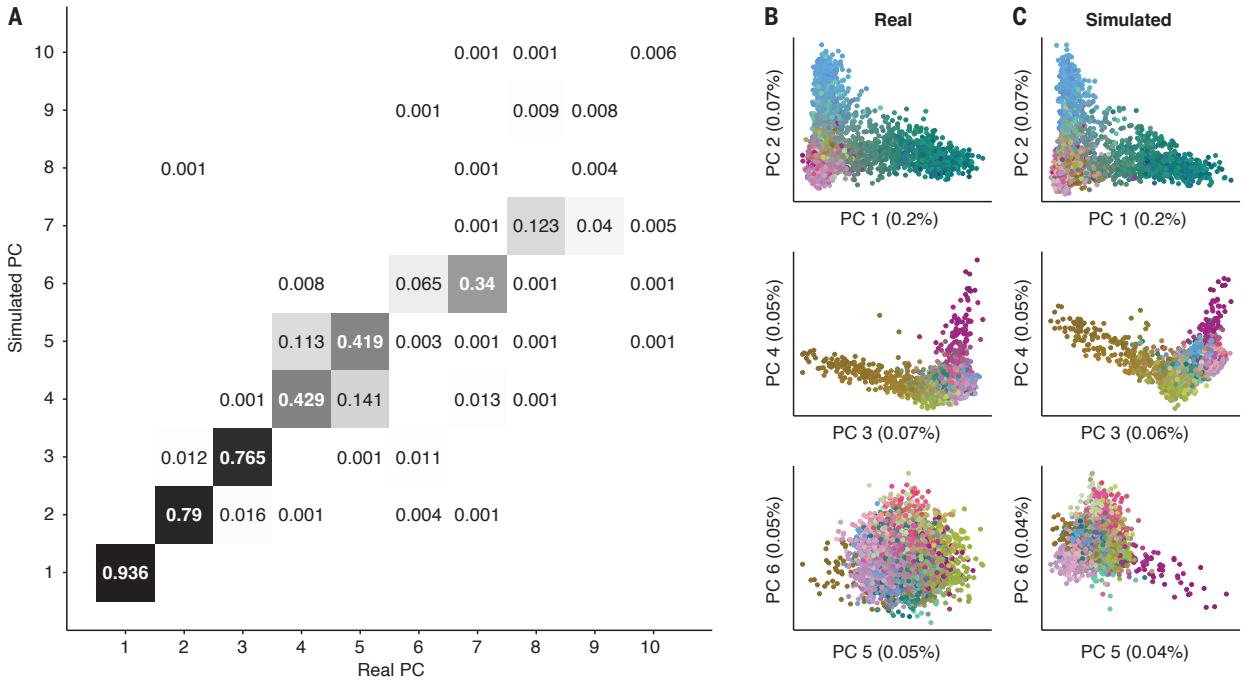

Figure 2: **Les génomes simulés reproduisent la structure observée dans la population.** Comparaison des mêmes 4 882 individus en utilisant des génomes observés et simulés (les couleurs correspondent à la fig. 1, voir le matériel supplémentaire pour plus de détails). (**A**) Corrélation entre les composantes principales observées et simulées. Les valeurs différentes de 0,000 sont affichées. (**B**) Projections ACP des génomes observés. (**C**) Projections ACP des génomes simulés.

démographique pour l'Europe (*32*) (voir fig. S5 et (*24*) pour les détails).

Nous avons comparé les simulations aux données de génotypage de 4,882 individus qui étaient liés à la généalogie. Des études précédentes avaient déjà montré une concordance qualitative entre données génétiques et généalogiques (*33*). Les analyses ACP et UMAP de nos données simulés montrent une excellente concordance avec les données de génotypage, en particulier en ce qui concerne les six premières composantes principales simulées (fig. 2 and S6). Dix simulations indépendantes montrent peu de variance dans leur concordance aux données réelles (fig. S7A). Les dix simulations montrent une excellente corrélation pour les premières huit composantes, puis une moins bonne corrélation pour les composantes plus hautes: l'aspect aléatoire de la transmission mendélienne a peu d'effet sur les premières composantes, mais réordonne les composantes plus

élevées (fig. S7B). En conclusion, les principaux axes de variation génétique parmi les Franco-Québécois reflètent la dérive génique survenue après la colonisation, dont les détails temporels et géographiques sont encodés dans l'arbre généalogique.

Étant donné la précision des données simulées, nous avons simulé de génome entier de 1,4 millions d'individus contemporains dont les 4 grands-parents pouvaient être liés à la généalogie. Ces génomes, codés en format 'tree sequence', contiennent des métadonnées spatiales et temporelles sur chacun des ancêtres communs des 1,4 millions d'individus. La figure S8 montre des analyses UMAP et ACP pour ces données simulées. On y retrouve une distribution continue et triangulaire de points le long des premières deux PCs, puis une distribution en étoile pour les prochaines 16 composantes, avec les branches opposées correspondant à une différentiation fine entre des régions voisines (parfois à 20 kilomètres, figure S9).

Migrations, apparentement, et bassins versants

Les figures 1 et S1 suggèrent que la structure génétique varie de façon continue dans l'espace, ce qui rappelle par exemple la structure de populations à l'échelle européenne (*9*). Par contre, chaque composante principale tend à distinguer une région donnée plutôt qu'un gradient géographique, ce qui rappelle davantage la structure plus fine entre les vallées alpines (*5*). Pour évaluer les effets de la topographie sur les taux de migration, nous visualisons les axes principaux de migration dans l'espace et le temps. La figure 3A montre que l'expansion de la population a été marquée par des migrations successives remontant les affluents du Saint-Laurent à des échelles allant de quelques dizaines à plusieurs centaines de kilomètres.

Le nom du premier lieu d'installation permanente française vient du mot Algonquin *kebec*, référant à une région où la largeur du Saint-Laurent diminue (*35*). La ville de Québec a été fondée à un emplacement stratégique pour les Français qui cherchaient à contrôler l'accès aux Grands Lacs au 17$^{\text{ème}}$ siècle [(*36*), p 49–52]. Faisant face à un vaste territoire forestier utilisé et occupé par les Premières Nations iroquoises et des régions boisées (*19*), les Français ont établi une colonie fluviale formant de minces bandes le long des cours d'eau [(*36*), p 56–57].

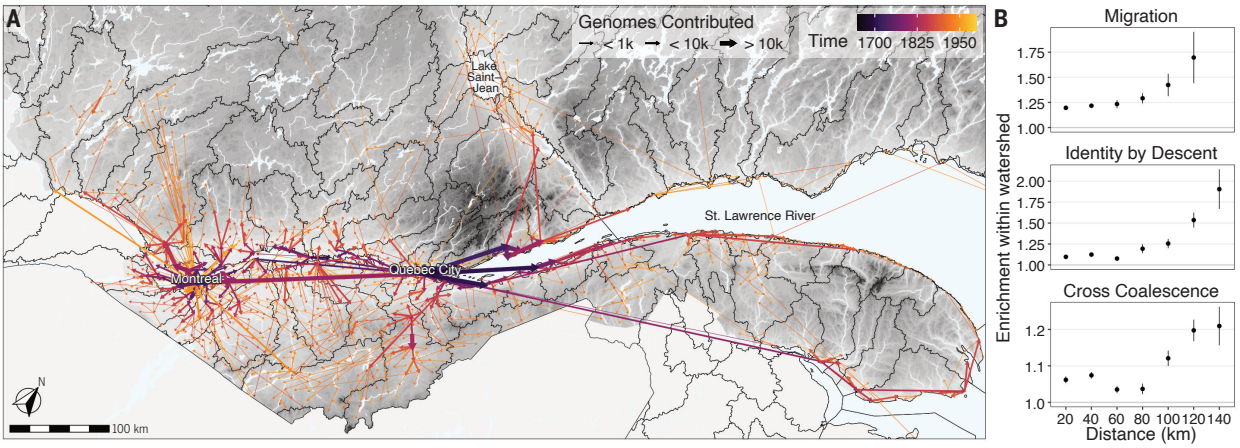

Figure 3: **Les bassins versants influencent les migrations et l'apparentement des Franco-Québécois.** (**A**) Routes principales de l'ascendance génétique estimée. Les segments relient chaque localité à la localité à partir de laquelle les migrants sont estimés avoir contribué le plus de matériel génétique (d'après la généalogie). La largeur des segments indique la contribution génétique totale attendue aux individus contemporains et la couleur indique la date moyenne à laquelle ces contributions attendues se sont produites. Pour éviter une surcharge graphique, nous avons exclu la région de l'Abitibi-Témiscamingue, les migrations de moins de dix kilomètres et les migrations contribuant à moins de dix génomes. (**B**) L'apparentement est plus élevé à l'intérieur des bassins versants. Les points noirs indiquent l'excédent de migrations et d'apparentement pour les localités situées dans le bassin versant par rapport aux localités situées hors du bassin versant à une distance fixe. (voir le matériel supplémentaire).

Pour quantifier l'apparentement entre les individus situés dans 1698 paroisses et les migrations entre ces paroisses, nous avons utilisé trois mesures: l'apparentement génétique (par IBD), l'apparentement généalogique (par coalescence croisée), et les taux de migrations. Les trois mesures indiquent un isolement par distance dans toutes les régions (fig. S10 et S11). Considérant l'importance des cours d'eau dans les modes de colonisation, l'activité économique, et les transports (*37*), nous avons testé l'hypothèse que les patrons d'apparentement et de migrations tendent à suivre les cours d'eau. Ces mesures seraient donc plus élevées pour des paroisses situées au sein d'un même bassin versant, à distance comparable (fig. 3B et (*24*)). Nous observons une telle

augmentations pour les trois mesures et pour toutes les distances. À 120km, où l'augmentation est la plus marquée, il y a 40% plus de migrations, 75% plus d'apparentement IBD, et 20% plus de coalescence croisée. L'impact de la topographie et le degré d'augmentation de ces mesures varient cependant entre les régions (figures 1C et S12).

L'histoire des migrations dans l'espace

Afin d'étudier l'apparition de la structure génétique régionale chez les Franco-Québécois, nous avons défini trois régions correspondant aux bassins versants au sein desquels de nombreux individus définissaient les composantes principales de variation (fig. S13). Ces régions correspondent empiriquement au Saguenay-Lac-Saint-Jean, au Bas-Saint-Laurent, et à la Beauce, et nous utiliserons ces termes pour faire référence aux régions définies par les bassins versants.

Pour chaque région et pour chaque paire d'individus dans la région, nous avons calculé la contribution de chaque ancêtre au coefficient de parenté ou, de façon équivalente, à la coalescence (voir (*24*)). Ceci nous a permis de mesurer le lieu, l'époque, et l'intensité des effets fondateurs (fig. 4 A-C). Le terme 'effet fondateur' fait ici référence à un lieu et une période où le taux de coalescence attendu est élevé. C'est-à-dire que plusieurs paires d'individus ont trouvé un ancêtre commun, indiquant une petite taille efficace de la population mais pas nécessairement une petite taille absolue.

Pour chaque région, l'effet fondateur le plus important est près de la ville de Québec, mais les régions se distinguent pour les effets fondateurs subséquents. Le Saguenay-Lac-Saint-Jean (SLSJ) ($\lambda = 0.005$, fig. 4 A, D, G) a un effet fondateur dominant à Baie-Saint-Paul ($\lambda = 0.001$) et dans les localités voisines de Charlevoix où un astroblème – une formation géologique formée par l'érosion d'un cratère d'impact – a donné naissance à des terres arables dans une région autrement montagneuse ((*38*), fig. S14). La taille restreinte de l'astroblème a mené à des pressions démographiques et une expansion rapide d'abord au Saguenay puis au lac Saint-Jean. [(*36*), p 91]. La majorité de l'apparentement entre individus au SLSJ est donc antérieure à la colonisation du SLSJ. La Beauce ($\lambda = 0.002$, fig. 4 C, F, I), a des effets fondateurs à St-Joseph-De-Beauce ($\lambda = 0.0003$) et le long

de la Chaudière, et des migrations qui rappellent un réseau en étoile. Finalement, le Bas-Saint-Laurent ($\lambda = 0.002$, fig. 4 B, E, H) présente un assortiment d'effets fondateurs, notamment à Rivière-Ouelle ($\lambda = 0.0004$), dispersés sur des centaines de kilomètres de rivage. Cet effet fondateur unidimensionnel marque les premières installations et l'origine de migrations vers l'intérieur des terres (fig. 4 E).

En conséquence de l'expansion de la population, certains des premiers colons (*super-fondateurs*) ont eu une grande contribution génétique à la population contemporaine (*39, 40*). Les dix super-fondateurs les plus importants de chaque région ont contribué 37%, 12% et 14% de l'apparentement au SLSJ, au Bas-Saint-Laurent, et à la Beauce respectivement (fig.S15,S16, S17). Étonnamment, alors que chaque région a son effet fondateur le plus fort près de la ville de Québec, aucune de ces régions ne partage les mêmes super-fondateurs (fig. S18). Pour mesurer le chevauchement entre les effets fondateurs régionaux, nous avons calculé les taux de coalescence croisée (*41*) (voir (*24*), et fig. S19) et trouvons qu'entre 35% et 50% de l'apparentement au Bas-Saint-Laurent et en Beauce peuvent être attribués à un effet fondateur commun (tableau 1). Par contre l'effet fondateur au SLSJ est seulement partagé à 5%, reflétant des effets fondateurs particuliers au SLSJ, à Charlevoix et près de la ville de Québec.

	Proportion de l'effet fondateur partagé avec:		
Région	SLSJ	Bas-Saint-Laurent	Beauce
SLSJ	-	0.055	0.035
Bas-Saint-Laurent	0.456	-	0.358
Beauce	0.412	0.516	-

Table 1: Proportion des effets fondateurs partagés entre les régions, d'après les taux de coalescence croisée divisés par les taux de coalescence des régions.

Ce type d'effet fondateur régional n'est pas présent dans toutes les régions du Québec. La colonisation rapide de l'Abitibi-Témiscamingue, qui peut rappeler celle du SLSJ, n'a pas donné lieu à des effets fondateurs dans les axes principaux de variation génétique. Contrairement à la situation au SLSJ, les colons franco-québécois en Abitibi venaient de plusieurs localités distribuées à travers la province (fig. S20). Alors que plusieurs localités en Abitibi-Témiscamingue ont des

effets fondateurs, ceux-ci se chevauchent peu. Par exemple, les villages de Rémigny et Rollet sont distants de 20 kilomètres le long de la rivière des Outaouais (fig. S20). Les coalescences croisées entre ces villages montrent un chevauchement de 11%. À titre de comparaison, La Baie et Roberval au SLSJ ont 70% de chevauchement malgré une distance de plus de 100 kilomètres. L'Abitibi-Témiscamingue comporte donc peu d'effet fondateur partagé et presque pas d'isolement par distance (fig. S10), alors que des effets fondateurs parallèles ont tout de même créé une importante sous-structure (fig. S20). Ce type de sous-structure cachée pourrait être fréquente dans l'histoire des populations humaines (*42*).

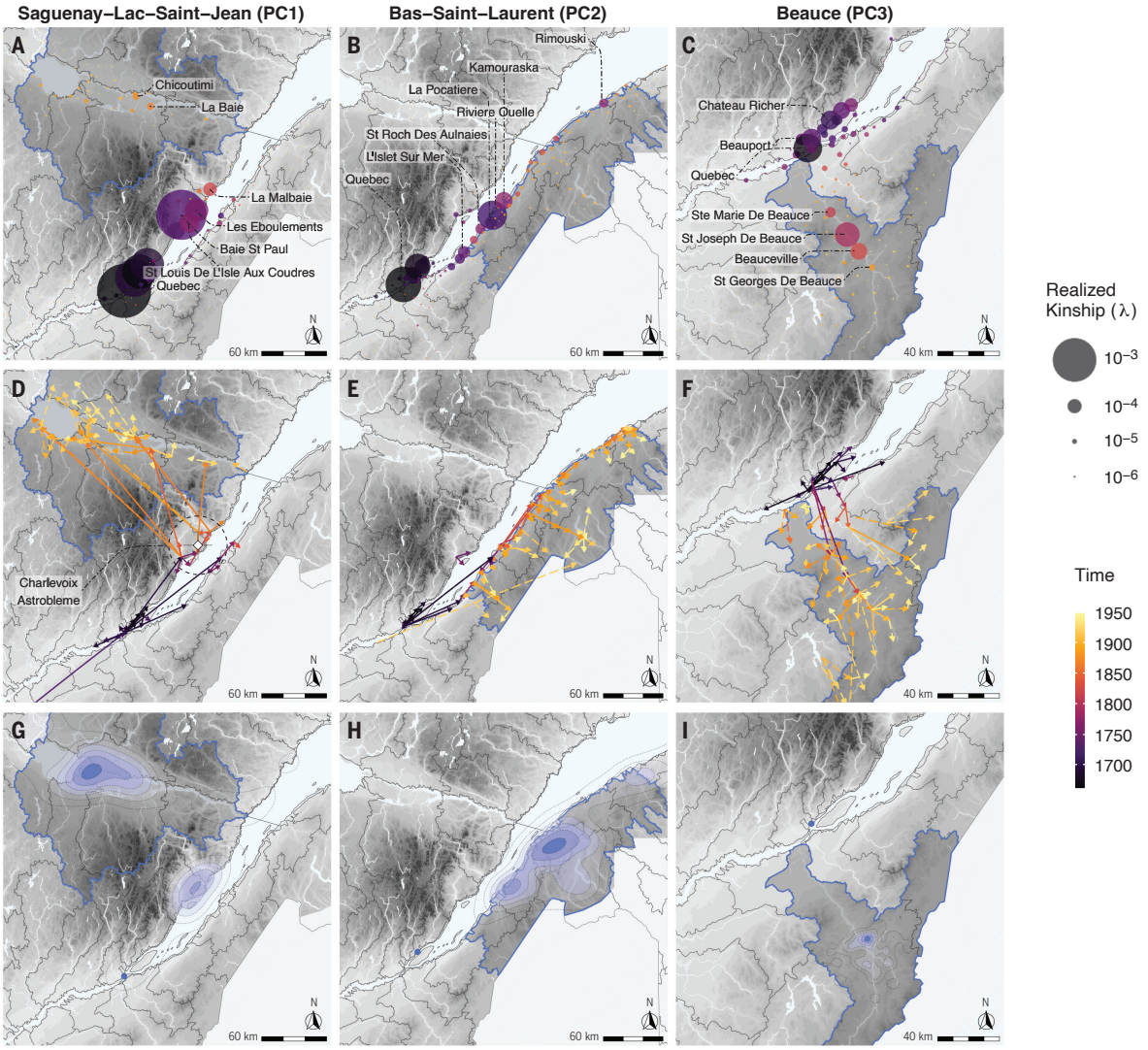

Figure 4: **Les migrations historiques et les événements fondateurs définissent la structure de la population** (**A-C**) Période, emplacement et intensité des événements fondateurs dans trois régions du Québec définies par les trois premières composantes principales (fig. S13). La taille des points est proportionnelle à l'apparentement réalisé dans une région donnée pour une population franco-québécoise contemporaine, tel qu'estimé à partir de la généalogie. La couleur des points représente la date moyenne de mariage des ancêtres communs. (**D-F**) Axes majeurs de migration mesurés par les contributions génétiques aux individus contemporains vivant dans les régions sélectionnées, estimés d'après la généalogie. Les flèches pointillées indiquent les localités ayant la plus grande contribution génétique aux localités des régions étudiées, et les flèches pleines mettent en évidence le cinquième percentile des itinéraires migratoires classés par contribution génétique estimée. Les couleurs des flèches représentent la date moyenne de la contribution génétique en

Discussion

Les modèles classiques en génétique des populations considèrent souvent la reproduction et le choix de partenaire comme des phénomènes aléatoires uniformes. Ces modèles simplifiés sont utilisés pour surmonter la complexité des choix et des objectifs individuels, qui ne sont pas accessibles aux chercheurs, afin d'expliquer des tendances telles que la dérive génétique et la sélection. Des études récentes ont montré les avantages d'une modélisation plus détaillée (par ex. (*4, 43*)). Cependant, abandonner l'hypothèse de l'appariement aléatoire augmente de façon considérable le nombre de paramètres démographiques devant être pris en compte.

La base de données généalogiques de BALSAC, formant un arbre généalogique particulièrement complet, a été utilisées pour identifier des effets multigénérationnels comme l'avantage reproductif d'être sur le front d'expansion (*39*), la transmission des tailles des familles (*44*) et des tendances migratoires (*45*). D'autres études ont considéré les corrélations entre variation génétique et spatiale (*33, 46*) et leurs déterminants historiques, comme l'apparentement des premiers colons de Charlevoix (*37*) ou la colonisation tardive du Saguenay suite à la levée de contraintes liées à la traite de fourrure [(*36*), p. 91].

Nous avons cherché ici à développer un modèle de variation génétique prenant en compte ces effets et bien d'autres. Séparer l'apparentement (ou la coalescence) en contributions individuelles et régionales nous a permis de décomposer les effets fondateurs en multiples contributions historiques. Ce modèle appliqué à la population fondatrice bien connue du SLSJ, a mis en valeur le rôle de la géographie, résultant notamment d'un événement cosmique ayant eu lieu il y a 400 millions d'années à Charlevoix. En fournissant la première description détaillée d'autres effets fondateurs au Québec, nous avons illustré une grande diversité dans la dynamique des effets fondateurs. Les données génétiques et généalogiques appuient l'hypothèse que la géographie, et en particulier les rivières et montagnes, ont joué un rôle majeur dans la mise en place des principaux axes de migration et de variation génétique. Finalement, la forte corrélation entre données empiriques et simulées montre que la plupart de la structure présente chez les Franco-Québécois peut être attribuée à des événements ayant eu lieu en Amérique du Nord, bien que la population

ait conservé des signatures génétiques propres aux régions de France d'où venaient les premiers colons.

Nous avons décrit comment des événements idiosyncratiques ainsi que des effets géologiques, sociaux et historiques se sont traduits en modes de variation génétique à différentes échelles géographiques et temporelles. Bien que nos simulations soient basées sur une véritable généalogie, elles ne contiennent pas d'informations sur les génomes individuels et nous pouvons partager librement cet ensemble de données à l'échelle du génome ainsi que des métadonnées spatio-temporelles pour plus de 1,4 million de personnes. Les généticiens se basent souvent sur des simulations pour évaluer l'exactitude et la robustesse des méthodes de prédiction du risque génétique, d'inférence historique et d'inférence généalogique. Pour cette raison, plusieurs études ont cherché à développer des modèles de simulation qui captent plus précisément la structure généalogique (par exemple, (*30, 47, 48*)). Cependant, ces modèles reposent sur des hypothèses démographiques contraignantes telles que l'appariement aléatoire. Nous pensons qu'un ensemble de données simulées publiquement disponible, avec une structure de population réaliste à l'échelle d'une population, aidera à concevoir des méthodes d'inférence plus robustes.

Évidemment, ces simulations sont loin d'être parfaites. Elles ne tiennent pas compte de la sélection naturelle (*39*), ou de l'apparentement non aléatoire entre les fondateurs généalogiques. La généalogie elle-même contient une certaine quantité d'erreurs liées à des cas de paternité erronée ou à des désaccords entre filiation biologique et filiation déclarée (*49*) (probablement moins de 1% des liens dans cette population). Ces erreurs semblent avoir un effet modéré sur la variation à grande échelle, mais poseront des défis pour les analyses plus fines (*43, 49*). Malgré ces limitations, nous croyons que ce modèle génétique, les outils de simulation, et les données publiquement disponibles découlant de ces simulations offre un nouvel outil pour explorer la variation génétique à une résolution inégalée.

References

1. A. W. Wohns, *et al.*, *Science* **375**, eabi8264 (2022).

2. L. L. Cavalli-Sforza, A. Piazza, P. Menozzi, J. Mountain, *Proceedings of the National Academy of Sciences* **85**, 6002 (1988).

3. B. M. Henn, L. L. Cavalli-Sforza, M. W. Feldman, *Proceedings of the National Academy of Sciences of the United States of America* **109**, 17758 (2012).

4. G. S. Bradburd, P. L. Ralph, *Annual Review of Ecology, Evolution, and Systematics* **50**, 427 (2019).

5. L. L. Cavalli-Sforza, *Scientific American* **221**, 30 (1969).

6. C. Battey, P. L. Ralph, A. D. Kern, *Genetics* **215**, 193 (2020).

7. S. Wright, *Genetics* **28**, 114 (1943).

8. S. Ramachandran, *et al.*, *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15942 (2005).

9. J. Novembre, *et al.*, *Nature* **456**, 98 (2008).

10. C. Bycroft, *et al.*, *Nature communications* **10**, 1 (2019).

11. A. Saint Pierre, *et al.*, *European Journal of Human Genetics* pp. 1–13 (2020).

12. N. A. Rosenberg, *et al.*, *PLoS Genet* **1**, e70 (2005).

13. G. S. Bradburd, G. M. Coop, P. L. Ralph, *Genetics* **210**, 33 (2018).

14. C. Moritz, *Trends in ecology & evolution* **9**, 373 (1994).

15. F. Jay, P. Sjödin, M. Jakobsson, M. G. Blum, *Molecular biology and evolution* **30**, 513 (2013).

16. D. Petkova, J. Novembre, M. Stephens, *Nature genetics* **48**, 94 (2016).

17. B. H. McRae, *Evolution* **60**, 1551 (2006).

18. H. Charbonneau, B. Desjardins, M. Boleda, R. Bates, *The First French Canadians: Pioneers in the St. Lawrence Valley* (University of Delaware Press, 1993).

19. Aboriginal and Northern Affairs Canada, *First Nations in Canada* (2013).

20. Y. Beauregard, *Mythe ou réalité. Les origines amérindiennes des Québécois: Entrevue avec Hubert Charbonneau.* (Cap-aux-diamants, 2010).

21. C. Moreau, *et al.*, *PLoS ONE* **8**, e65507 (2013).

22. H. Vézina, J.-S. Bournival, *Historical Life Course Studies* (2020).

23. P. Awadalla, *et al.*, *International journal of epidemiology* **42**, 1285 (2013).

24. Materials and methods are available as supplementary materials online.

25. L. McInnes, J. Healy, *arXiv preprint arXiv:1802.03426* (2018).

26. A. Diaz-Papkovich, L. Anderson-Trocmé, C. Ben-Eghan, S. Gravel, *PLoS genetics* **15**, e1008432 (2019).

27. L. Stanislas-Alfred, *Bulletin du parler Français au Canada* **2**, 17-18 (1903).

28. H. Vézina, M. Tremblay, B. Desjardins, L. Houde, *Cahiers québécois de démographie* **34**, 235 (2005).

29. J. Kelleher, A. M. Etheridge, G. McVean, *PLoS computational biology* **12** (2016).

30. D. Nelson, *et al.*, *PLoS genetics* **16**, e1008619 (2020).

31. F. Baumdicker, *et al.*, *Genetics* **220** (2021).

32. J. A. Tennessen, *et al.*, *Science* **337**, 64 (2012).

33. M.-H. Roy-Gagnon, *et al.*, *Human Genetics* **129**, 521 (2011).

34. L. Anderson-Trocmé, Simulated genomes from manuscript "On the Genes, Genealogies and Geographies of Quebec", v1.0, Zenodo (2023); `https://doi.org/10.5281/zenodo.7702392`.

35. Y. Hébert, *Les Ponts de glace sur le Saint-Laurent* (GID, 2012).

36. S. Courville, *Quebec: A Historical Geography* (UBC Press, 2009).

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

38. J. Rondot, *Canadian Journal of Earth Sciences* **5**, 1305 (1968).

39. C. Moreau, *et al.*, *Science* **334**, 1148 (2011).

40. D. Labuda, T. Harding, E. Milot, H. Vézina, *PloS one* **17** (2022).

41. S. Schiffels, R. Durbin, *Nature genetics* **46**, 919 (2014).

42. E. M. Scerri, L. Chikhi, M. G. Thomas, *Nature ecology & evolution* **3**, 1370 (2019).

43. D. Nelson, *et al.*, *The American Journal of Human Genetics* **103**, 893 (2018).

44. A. Gagnon, E. Heyer, *American Journal of Human Biology: The Official Journal of the Human Biology Association* **13**, 645 (2001).

45. A. Gagnon, B. Toupance, M. Tremblay, J. Beise, E. Heyer, *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists* **129**, 630 (2006).

46. K. M. Burkett, *et al.*, *Frontiers in Genetics* **12**, 1 (2022).

47. J. Wakeley, L. King, B. S. Low, S. Ramachandran, *Genetics* **190**, 1433 (2012).

48. A. Bhaskar, A. G. Clark, Y. S. Song, *Proceedings of the National Academy of Sciences* **111**, 2385 (2014).

49. H. Vézina, M. Jomphe, È.-M. Lavoie, C. Moreau, D. Labuda, *Cahiers québécois de démographie* **41**, 87 (2012).

50. L. Anderson-Trocmé, Lukeandersontrocme/genes_in_space: v1.0, Zenodo (2023); `https://doi.org/10.5281/zenodo.7703131`.

51. L. Anderson-Trocmé, Lukeandersontrocme/genome_simulations: v1.0, Zenodo (2023); `https://doi.org/10.5281/zenodo.7702392`.

52. Fichier Balsac, Genealogy of Quebec Population 1621-1993 (2022).

53. 1000 Genomes Project Consortium and others, *Nature* **526**, 68 (2015).

54. A. Teucher, K. Russell, *rmapshaper: Client for 'mapshaper' for 'Geospatial' Operations* (2020). R package version 0.4.4.

55. Statistics Canada, *Table 17-10-0009-01 Population estimates, quarterly* (Government of Canada, 2023).

56. Statistics Canada, *Census of Population, 1991 to 2021 (3901)* (Government of Canada, 2021).

57. L. Tang, *Nature Methods* **17**, 876 (2020).

58. A. P. Ragsdale, C. Moreau, S. Gravel, *Current Opinion in Genetics & Development* **53**, 140 (2018).

59. P. F. Palamara, *et al.*, *The American Journal of Human Genetics* **97**, 775 (2015).

60. International HapMap Consortium, *et al.*, *Nature* **449**, 851 (2007).

61. G. Abraham, Y. Qiu, M. Inouye, *Bioinformatics* (2017).

62. J. Melville, *uwot: The Uniform Manifold Approximation and Projection (UMAP) Method for Dimensionality Reduction* (2020). R package version 0.1.10.

63. P. Menozzi, A. Piazza, L. Cavalli-Sforza, *Science* **201**, 786 (1978).

64. N. Patterson, *et al.*, *Genetics* **192**, 1065 (2012).

65. M. Petr, B. Vernot, J. Kelso, *Bioinformatics* **35**, 3194 (2019).

66. O. Delaneau, J.-F. Zagury, M. R. Robinson, J. L. Marchini, E. T. Dermitzakis, *Nature communications* **10**, 1 (2019).

67. Y. Zhou, S. R. Browning, B. L. Browning, *The American Journal of Human Genetics* **106**, 426 (2020).

68. M. De Braekeleer, M. Ross, *Human heredity* pp. 379–384 (1991).

69. R. R. Hudson, *Theoretical Population Biology* **23**, 183 (1983).

70. A. P. Ragsdale, *et al.*, *bioRxiv* (2022).

Remerciements

Nous sommes reconnaissants envers tous les participants qui ont rendu cette étude possible en contribuant leur ADN, ainsi qu'envers les participants qui ont fourni des informations familiales permettant la reconstitution de leur généalogie. Pour les données provenant du Québec, nous remercions l'équipe de BALSAC pour leur gestion et leur curation de la base de données généalogiques, l'équipe de CARTaGENE et Genome Quebec pour leur gestion et leur curation des données de génotype. Pour les données provenant de France, nous remercions l'équipe EREN pour leur gestion et leur curation des données de génotype de SU.VI.MAX. Nous remercions Claude Bhérer, Gil McVean, Claudia Moreau, Aaron Ragsdale, Rob Sladek, Wilder Wohns, Yan Wong pour des conversations utiles.

Sources de financement

Cette étude a été financée en partie par les bourses de la Reine Élizabeth (LAT), le Fonds de recherche du Québec – Nature et technologies (FRQNT: B2X 290358) (LAT); Cette étude a aussi été financée par une subvention projet des Instituts pour la Recherche sur la Santé du Canada (IRSC, bourse 437576), le conseil de recherche en sciences naturelles et en génie (CRSNG), RGPIN-2017-04816, le programme de Chaire de Recherche du Canada (SG) et la fondation Canadienne pour l'Innovation. Les travaux de JK sont financés par la Fondation Robertson.

Contributions

Conceptualisation: LAT, DN, CD, HV, JK, SG

Curation des données: LAT, BALSAC, CARTaGENE, Genome Quebec, EREN, MT

Analyses formelles: LAT

Financement: LAT, JK, SG

Processus de recherche: LAT, SG

Methodologie: LAT, DN, SZ, ADP, JK, SG

Gestion du projet: LAT, JK, SG

Resources: SG

Logiciels: LAT, DN, IK, SZ, NB, BJ, JK, SG

Supervision: JK, HV, SG

Validation: LAT, DN, IK, NB, BJ, JK, SG

Visualisation: LAT

Écriture – première mouture: LAT

Écriture, révision et modifications: LAT, AR, CD, ADP, SZ, JK, HV, SG

Écriture, traduction vers le français: SG

Conflits d'intérêts

Les auteurs ne rapportent aucun conflit d'intérêt.

Disponibilité des données et logiciels

Code en R pour les analyses et visualisations: Zenodo (`https://zenodo.org/record/7703131`). Code en Python pour les simulations: Zenodo (`https://zenodo.org/record/7705869`) Données simulées en format Tree Sequence: Zenodo (`https://doi.org/10.5281/zenodo.6839683`). Les données de génotypage sont disponibles chez les cohortes, via un comité d'accès aux données indépendant chez CARTaGENE (`www.cartagene.qc.ca/en/researchers/access-request.html`) et via Genome Quebec pour Genizon (`https://genomequebec.com/0-genizon-biobank/`). Les données généalogiques sont disponibles sur le portail Dataverse de l'Université du Québec à Chicoutimi (`https://doi.org/10.5683/SP3/BW7DIG`).

Contenu des "Supplementary Materials"

Matériaux and Méthodes

Texte supplémentaire

Figures S1 à S24

Tableaux S1 à S4

Références (51-70)

Science
AAAS

Supplementary Materials for

On the Genes, Genealogies, and Geographies of Quebec

Luke Anderson-Trocmé *et al.*

Corresponding author: Simon Gravel, simon.gravel@mcgill.ca

Science **123**, abc4567 (2023)
DOI: https://doi.org/10.1101/2022.07.20.500680

The PDF file includes:

Materials and Methods
Supplementary Text
Figs. S1 to S24
Tables S1 to S4

Other Supplementary Material for this manuscript includes the following:

MDAR Reproducibility Checklist

Materials and Methods

1 Data **2**
- 1.1 Genetic data . . . 2
- 1.2 Genealogical data . . . 2
- 1.3 Geographical data . . . 3
- 1.4 Demographic data . . . 3

2 Genome simulations **3**
- 2.1 Fixed pedigree simulation model . . . 3
- 2.2 Rescaling demographic models . . . 4
- 2.3 Model specifications . . . 4
- 2.4 Fixed pedigree simulation details . . . 5
- 2.5 Comparing simulations to ascertained genotype data . . . 6

3 Genetic statistics **6**
- 3.1 Genetic ancestry of French Canadians . . . 6
- 3.2 Dimension reduction . . . 7
- 3.3 F statistics . . . 7
- 3.4 Identity by descent . . . 8
  - 3.4.1 Calculation . . . 8

4 Genealogical statistics **8**
- 4.1 Estimated contributions . . . 8
- 4.2 Coalescence rates . . . 9
  - 4.2.1 Within-population coalescence . . . 9
  - 4.2.2 Cross coalescence . . . 11
  - 4.2.3 Overlap of relative cross coalescence . . . 12
  - 4.2.4 Normalization . . . 12
- 4.3 Migration rates . . . 13
  - 4.3.1 Contribution date . . . 13
  - 4.3.2 Relative emigration rate . . . 13
- 4.4 Genealogy flow plot . . . 14
- 4.5 Code availability . . . 14

5 Enrichment analysis **15**
- 5.1 Additional Explanation of Results . . . 18

1 Data

1.1 Genetic data

The genotype data used in this study was compiled from three separate cohorts, each of which was imputed separately using the Michigan Imputation Server then merged (Supplementary Figure S23). The regional distribution of the samples included in the study is summarized in Figure 1 and Figure S22. The Genizon cohort (genomequebec.com/0-genizon-biobank/) is comprised of 9,961 genotyped individuals from Quebec of which 2,431 of them consented to and were successfully linked to genealogical records. The genotype data from this cohort was produced on 4 different chips (HumanHap375, HumanHap550, Illumina1M and Human610-Quad) and due to data being unavailable for chromosome 22 for one of the Genizon chips, we restricted all of our genetic analyses to the first twenty-one chromosomes. The CARTaGENE dataset (cartagene.qc.ca/en/researchers/access-request.html) used here comprises of 12,062 genotyped individuals from Quebec of which 5,733 consented to and were successfully linked to genealogical records. The genotype data from this cohort was also produced on 4 different chips (Omni2.5, Axiom2.0, GSAv1 and GSAv2). The SUVIMAX cohort includes 2,184 genotyped individuals from France, see (*11*) for details. For details about the downsampled sequence data from the 1000 Genomes GBR population, see (*53*).

1.2 Genealogical data

BALSAC is a comprehensive genealogy of the French-Canadian population of Quebec compiled from 4,282,960 marriage records dating back to the 17th Century (*22*). Data are available on the Scholars Portal Dataverse platform from the University of Québec in Chicoutimi (*52*).

1.3 Geographical data

Layers of geographical data including rivers, lakes, watersheds, provincial, and federal boundaries were downloaded from the government of Canada geobase. The National Hydro Network GeoBase Series database can be accessed through : `https://open.canada.ca/`.

Polygons from each layer were simplified using ms_simplify function from the rmapshaper R library (*54*) and projected onto the EPSG:4326 coordinate system. The Canadian Digital Elevation Model GeoTIF files were downloaded from the government of Canada (`https://open.canada.ca/`). The hydrological and altimetry data are licensed under the Open Government Licence allowing for their use, modification and publication.

1.4 Demographic data

Population size estimates of the province of Quebec were obtained from the preliminary results of the 2023 Census (*55*) and estimates of the number of individuals who speak French as a primary language from (*56*).

2 Genome simulations

2.1 Fixed pedigree simulation model

We extended the `msprime` software to include support for simulations conditional on a fixed pedigree. This new extension simulates the effects of recombination and the transfer of ancestral material from children to parents based on the structure of the pedigree. This simulation model can use user-specified recombination rates (or maps), accounts for founders living at different times, and accommodates founders from multiple source populations. Ancestry beyond the fixed pedigree can be simulated using arbitrarily complex demographic models including those specified in the PopSim Consortium (`https://popsim-consortium.github.io/stdpopsim-docs/stable/index.html`) (*57*). Here we used a re-scaled Tennessen

model (*32*).

2.2 Rescaling demographic models

The Tennessen model (*32*) was originally inferred using a high coding sequence mutation rate of $\mu_{cds} = 2.35 \times 10^{-8}$ per base pair per generation. To simulate genomes, we want to use a mutation rate that reflects genome-wide estimates of the mutation rate. For this purpose, we use relative rate of intergenic to coding polymorphism, i.e., $\mu_{int}/\mu_{cds} = 1.53$ as described in (*58*). This would lead us to an unreasonably high genome-wide mutation rate of 3.60×10^{-8} . Recent mutation rate estimates tend to be much lower, such as the estimate of 1.66×10^{-8} from (*59*), corresponding to a synonymous mutation rate of 1.10×10^{-8} . In other words, the Tennessen model was fit using a mutation rate that was $2.35/1.1 = 2.18$ higher than the value suggested by the analyses of (*59*) and (*58*). Fortunately, because the Tennessen model was fit to distributions of allele frequencies, the only role of the mutation rate in the analysis was to provide a scale for the times and population sizes. In other words, we can rescale the Tennessen model to reflect the updated mutation rates by multiplying the times and population sizes by 2.18, and correspondingly dividing the migration rates. To reflect genome-wide mutation rates, we use this rescaled model (fitted to synonymous sites) and a genome-wide mutation rate of 1.66×10^{-8}. We verified empirically that this rescaling only affected the distribution of allele frequencies via a constant scaling factor, thus providing an equivalent fit to the original Tennessen data. We further validated this re-scaled model using two-locus statistics (Figure S5). We also found that using this re-scaled model provided better correlations of principal components between real and simulated data than using the original (unscaled) Tennessen model with the corresponding higher mutation rate.

2.3 Model specifications

Even though our software implementation accommodates multiple source populations for the founders, we considered a single source population of European ancestry as defined by the rescaled two population out-of-Africa model described above. The chromosome length and recombination

rate for each simulated chromosome was defined by the GRCh37 hapmapII genetic map (*60*). The genomic regions belonging to centromeres and telomeres were excluded from our simulations given their lack of documented recombination events. Once the tree sequences were constructed, mutations were added to branches of the tree at a rate of 1.66×10^{-8} per basepair per meiosis (*59*). A summary of model specifications are in Table S4.

2.4 Fixed pedigree simulation details

Genome simulations conditioned on a fixed pedigree use the `sim_ancestry` function in `msprime`. This function takes a fixed pedigree as a parameter (which is passed as a an "initial state" parameter in msprime). To build this initial state, we create an msprime "pedigree" object, and use the `PedigreeBuilder` function to add each individual with their parental information to the `msprime` pedigree object. At this point, we also add additional metadata that can also be used for downstream analyses (such as geographic coordinates). Once the initial state is specified, the `sim_ancestry` function is run twice. In the first run of `sim_ancestry`, the coalescence process is simulated within the pedigree and stops when lineages reach a pedigree founder (i.e., an individual whose parents are not part of the pedigree). At that point, there can be many uncoalesced lineages. In the second run of `sim_ancestry` , we simulate the ancestry beyond the pedigree under an idealised model – in this case the modified Tennessen demographic model described above. This simulation proceeds until all lineages have coalesced and we have generated a fully coalesced tree sequence for the simulation. Once the tree sequence is completed, we add mutations to the tree with the specified mutation rate across both simulation epochs.

A repository with code to run the genome simulation pipeline is available on Zenodo (*50, 51*) and extensive documentation of Msprime (`https://tskit.dev/msprime/docs/latest/api.html#msprime.FixedPedigree`).

2.5 Comparing simulations to ascertained genotype data

The simulation of genomes involves simulating relatedness among individuals, and mutations along the gene genealogy. Thus, simulated polymorphisms do not occur at the same genomic positions as real observed polymorphism. To compare the simulated genomes to ascertained genotype data, we downsampled the simulated polymorphisms to match the density of the genotype data. We did so by removing variants below a 5% minor allele frequency and linkage disequilibrium pruning such that both datasets contained $\sim 60,000$ variants. For this comparison, we restricted our simulations to 4,882 individuals from a total of 5,402 individuals who consented to be linked to the French Canadian pedigree. While we selected individuals for whom genotype data was available, to enable model validation, we did not use the real genotypes themselves in the simulation. We excluded 100 individuals who did not have all four grand-parents present in the pedigree. We also excluded 420 individuals with second cousins or closer relatives to match a quality control step typical in population genetic studies. We performed a principal component analysis (*61*) on both ascertained and simulated data using the same 4,882 individuals, and for visualization purposes, we used the same three-dimensional colours used in Figure 1.

3 Genetic statistics

3.1 Genetic ancestry of French Canadians

The majority of individuals in Quebec derive French Canadian ancestry (*55, 56*), as such we expect the majority of the participants in our cohorts also derive French Canadian ancestry despite only a fraction of them being linked to the genealogy. For the purposes of visualizing the population structure of French Canadian in Figure 1A and C, we sought to leverage the large number of French Canadian participants included in our PCA and UMAP analyses. We defined a threshold based on the genotype data from participants linked to this genealogy and their projections along the first principal component. This threshold kept 21,146 genotyped individuals with presumed French

Canadian ancestry and excluded 617 individuals from our analyses (see Supplementary Figure S24 and Supplementary Table S3).

3.2 Dimension reduction

Flashpca2 was used to perform our principal component analysis (PCA) (*61*) to generate (Supplementary Figure S1) and the R package uwot (*62*) for our uniform manifold and approximation projection (UMAP) (*25*) used to generate Figure 1A. This method takes the first ten principal components of genetic data as input and reduces this high dimensional data to a lower dimension while seeking to preserve local neighbourhoods.

The colours used in Figure 1 were determined by reducing the top ten principal components of genotype data to a three dimensional UMAP and then converted each x, y, z coordinate into an RGB value that is unique to each individual (*63*).

3.3 F statistics

F statistics used in Supplementary Figure S2A were computed using the admixture R package (*64, 65*). We used all 2,184 genotyped individuals from France as they all had regional geospatial information available. We also included 94 British individuals from the 1000 Genomes Project as an additional potential founding population. Together, we refer to the British and French samples as European. For the French Canadian samples, we used the 4,882 individuals linked to the genealogy with geospatial information. The population groupings used for the French Canadian individuals were watershed boundaries, the French groupings used seven French regions (South-East, South-West, West, North-West, Central, Isle-of-France, East) (Table S2), and the British GBR individuals were kept in a single group.

We computed all pairwise F_4 statistics $F_4(qc1, qc2, eu1, eu2)$ between French and Quebec regions. As a positive control we also computed the complement $F_4(qc1, eu1, qc2, eu2)$ statistics for all regions.

3.4 Identity by descent

3.4.1 Calculation

Using the genotype data from samples from Quebec and from France, in the Genizon, CARTaGENE, SUVIMAX, and GBR cohorts, we first phased the data using Shapeit4 (*66*) and downsampled to a set of common SNPs before computing pairwise IBD using Hap-IBD for all samples (*67*) using a minimum segment length of seven centimorgans.

We used all 2,184 genotyped individuals from France as they all had regional geospatial information available. For the French Canadian samples, we used the 4,882 individuals linked to the genealogy with geospatial information. The population groupings used for the French Canadian samples were geographic regions as defined in Figure 4 and the French groupings used the seven French regions (South-East, South-West, West, North-West, Central, Isle-of-France, East).

Using the rates of IBD between all pairs of individuals, we computed the average IBD sharing rates $g(A, B)$ between sets P^A and P^B of individuals in towns A and B respectively:

$$g(A,B) = \frac{1}{|P^A||P^B|} \sum_{i \in P^A, j \in P^B} IBD(i,j) \tag{S1}$$

where $IBD(i,j)$ is the total length in centiMorgans of IBD segments between individuals i and j, $|P^A|$ and $|P^B|$ are the sample sizes in towns A and B.

4 Genealogical statistics

4.1 Estimated contributions

Let us call $\mathrm{K}^P(i)$, the expected genetic contributions of an individual i to a set P of probands, measured in units of diploid genomes,

$$\mathrm{K}^P(i) = \sum_{p \in P} \mathrm{K}^p(i), \tag{S2}$$

where $0 \geq \mathrm{K}^p(i) \geq 1$ is the expected contribution of individual i to proband p. Similarly, the contributions $\mathrm{K}^P(I)$ of a set I of individuals to probands P is simply

$$\mathrm{K}^P(i) = \sum_{i \in I} \mathrm{K}^P(i). \tag{S3}$$

The expected proportion of genomes contributed by ancestor i to a set of probands is simply $\frac{\mathrm{K}^P(i)}{n_p}$.

4.2 Coalescence rates

4.2.1 Within-population coalescence

Founder effects and genetic bottlenecks result in excess kinship among individuals from a population. Given a spatial pedigree, we can identify the specific common ancestors that contribute to kinship between any two individuals, and therefore track a founder effect in space and time. Given a set of probands P, define $\lambda^P(i)$ as the total expected pairwise kinship realized in individual i. In other words, $\lambda^P(i)$ measures how often i is the most recent common ancestor of pairs of individuals in P.

This can be estimated rapidly from the genetic contributions $\mathrm{K}^P(\cdot)$ of the different offspring to individual i: Let's select a random pair of probands p_1 and p_2, and a random haploid copy of their genome (or *ploid*) at a single locus. Consider the probability of finding the most recent common ancestor of these two ploids at that locus in ancestral individual i. For this to happen, each ploid must be inherited from a distinct offspring of i. Thus the probability of coalescence can be written as a sum over unordered pairs of offspring (m,n) of i. The probability that m is ancestral to the ploid in p_1 and n is ancestral to the ploid in p_2 is simply $K^{p_1}(m)K^{p_2}(n)$, and the probability that (m, n) are distinctly ancestral to (p_1, p_2) is $K^{p_1}(m)K^{p_2}(n) + K^{p_2}(m)K^{p_1}(n)$. If both m and n carry an ancestral allele, the probability of coalescence in i is $\frac{1}{8}$, leading to a probability of coalescence

$$\lambda^{p_1,p_2}(i) = \frac{1}{8} \sum_{(m,n)} K^{p_1}(m)K^{p_2}(n) + K^{p_2}(m)K^{p_1}(n).$$

Finally, is we assume that n_p is large such that we can neglect the probability of sampling the same ploid twice, the probability of sampling the ordered pair (p_1, p_2) is $\frac{1}{n_P^2}$, so that the total rate of coalescence for a randomly selected pair of haploid lineages is

$$\lambda^P(I) = \sum_{(p_1,p_2)} \frac{1}{n_P^2} \lambda^{p_1,p_2}(i) \approx \frac{1}{4} \sum_{(m,n)} \frac{K^P(m)}{n_p} \frac{K^P(n)}{n_p}.$$

The realized kinship $\lambda^P(i)$ also measures the instantaneous coalescence rate (ICR) between pairs of ploids among probands P. In this equation we can interpret $(K^P(m))/n_P$ as the proportion of present-day genomes contributed by individual m.

Because each pair of individuals has a single most recent common ancestor at a given genomic coordinate, ICRs are additive across individuals i. We can therefore compute realized kinships over regions, time periods, or the entire genealogy by summing the realized kinships over the corresponding individuals.

To relate the stringency of recent bottleneck experienced in a given region to an effective population size, we can sum the ICRs over all individuals in the genealogy to obtain a coalescence rate, and invert this rate to obtain an estimate of the stringency of the bottleneck as measured by twice the effective population size needed to obtain a corresponding rate of coalescence in a single generation. In coalescent theory, the instantaneous coalescence rate is usually defined as a rate of coalescence among un-coalesced lineages, as this is the quantity that, when inverted, provides an unbiased estimate of the effective population size. Here we do not worry about this distinction, since the proportion of uncoalesced pairs remains very close to 1 throughout the genealogy. As a result, the coalescence rates can be easily added over different times and places. Thus genetic drift in the ancestry of participants can be partitioned into arbitrarily fine "founder events". Whether these founder events occurred sequentially over time or in parralel in different regions, their contributions to the coalescence rate can be added to provide a global estimate of the founder effect in the population.

While rates of coalescence are additive, interpreting N_e for specific regions is challenging

because the effective population sizes are not additive. In the main text, we therefore report realized kinships or ICRs when trying to understand the dynamics of a bottleneck over place and time.

While there is no accepted definition of what amounts of coalescence corresponds to a founder event, events we discussed in the text and displayed correspond to coalescence rates of 1×10^{-3} to 1×10^{-5}, corresponding to the effect on drift of a founder event lasting a single generation with an effective population sizes of 1000 to 100000. However, these effective sizes are a very poor indication of census sizes: small towns often have very low coalescence rates simply because they contribute little overall ancestry to the present population. Small towns thus often have very large "effective population sizes", making the concept rather confusing.

Since coalescence rates are approximately additive over different founder events, whether they occur simultaneously in different regions or sequentially. The overall coalescence rates for Figures 4A-C, summed over all founder events, is 0.005 for panel A (Saguenay-Lac saint Jean), 0.002 for panels B and C (Bas Saint-Laurent). These would correspond to the amount of drift in a population of size 1000 and 2500 over the course of the corresponding time period of 10 generations.

4.2.2 Cross coalescence

In a similar fashion as the within-region realized kinship described above, we can compute the cross coalescence rate between regions to identify the specific common ancestors that contribute to kinship between any two individuals in *different* regions, and therefore determine whether certain founder effects are shared between regions.

Given a set of probands P^A and P^B, define $\lambda^{P^A P^B}(i)$ as the total expected pairwise kinship realized in individual i. In other words, $\lambda^{P^A P^B}(i)$ measures how often i is the most recent common ancestors of pairs of individuals in P^A and P^B. This can be estimated rapidly from the genetic contributions $\mathrm{K}^{P^A}(\cdot)$ and $\mathrm{K}^{P^B}(\cdot)$ of offspring to individual i.

$$\lambda^{P^A P^B}(i) \approx \frac{1}{4} \sum_{(m,n)} \mathrm{K}^{P^A}(m) \mathrm{K}^{P^B}(n), \tag{S4}$$

where (m, n) are the ordered pairs of offspring for individual i.

4.2.3 Overlap of relative cross coalescence

To assess the percent overlap of two bottlenecks, we can contrast the cross-coalescence rates to the within-population coalescence rates. Given a set of probands P^A and P^B, we define $\gamma^{P^A P^B}$ as the ratio of cross coalescence to within-population coalescence P^A and P^B (*41*):

$$\gamma^{P^A P^B} = \frac{2\Lambda^{P^A P^B}}{\Lambda^{P^A} + \Lambda^{P^B}} \tag{S5}$$

where

$$\Lambda^{P^A P^B} = \frac{1}{|P^A||P^B|} \sum_i \lambda^{P^A P^B}(i),$$

$$\Lambda^{P^A} = \frac{2}{|P^A| \times (|P^A| - 1)} \sum_i \lambda^{P^A}(i),$$

and

$$\Lambda^{P^B} = \frac{2}{|P^B| \times (|P^B| - 1)} \sum_i \lambda^{P^B}(i).$$

4.2.4 Normalization

The relative cross-coalescence rate can be interpreted as a measure of the similarity of coalescence history between pairs of individuals within and across population, and is therefore normalized by sample size. This is shown, for example, in Figure 3. By contrast, when trying to compare the amount of kinship realized in historical individuals, we want to account for the fact that individuals who had many descendents contributed a lot to present-day kinship. In Figure 4, we therefore use a non-normalized kinship measure to identify the total contributions of individuals to present-day relatedness.

4.3 Migration rates

For the purposes of studying historical migrations, we assume that individuals are born in the location where their parents married and migrate to the location of their own marriage. We note that this definition is incomplete as it does not account for a cultural practice within French-Canadians where couples would tend to marry in the local church of the female counterpart and then move to the region of origin of the male counterpart. In our analyses, our estimates of migration rates are averaged across both sexes and all generations.

4.3.1 Contribution date

We defined above the genetic contribution for a set of individuals. To study the contributions of migrants specifically, we consider the set $M_{a\to b}$ of individuals born in source-town a and married in sink-town b and their total contribution $\mathrm{K}^P(M_{a\to b})$. To characterize the time period where contributing migrations occurred, we also report the mean contribution date $d(M_{a\to b})$ defined as

$$d(M_{a\to b}) = \sum_{i\in I_{a\to b}} w_i d_i, \tag{S6}$$

where d_i is the marriage date of individual i and w_i is a weight proportional to the total estimated genetic contribution $\mathrm{K}^P(i)$.

4.3.2 Relative emigration rate

In population genetics, a commonly used definition of migration rate is the fraction of immigrants over the total population size. However, in our enrichment analysis, we seek to compare multiple *inbound* migration rates $\delta_{a\to b}$ from multiple choices for source-town a to the same reference sink-town b. For this reason, we use a less common definition of migration rate of :

$$\delta_{a\to b} = \frac{|M_{a\to b}|}{N_a}, \tag{S7}$$

where N_a is the number of people born in source-town a.

The *emigration* rates account for the different population sizes of different source-towns since we normalize over N_a rather than N_b. The sum of migration rates $\delta_{A\to b}$ for a set A of source-towns to b.

$$\delta_{A\to b} = \sum_{a\in A} \delta_{a\to b}. \tag{S8}$$

4.4 Genealogy flow plot

We generated a visual summary of French Canadian ancestry using the riverplot R package. The x axis of the plot was generated by grouping together individuals based on administrative region boundaries in Quebec with some slight modifications highlighted in Supplementary Figure S21. The line thickness was obtained by aggregating the total estimated genetic contributions $\mathrm{K}^P(I_{a\to b,t})$ of individuals $I_{a\to b,t}$ to all probands P in the genealogy based on where they were born a, where they married b, and when they married t. This plot includes missing data as contributions fading to white for each region and time bin. To avoid overplotting, we exclude small migrations contributing less than the top 20 percent of migrant contributions. The y axis of the plot is separated into 60 year time bins starting from 1620 and ending in 1980. A small number of individuals in this dataset married either before or after this time range were added to the first and last time bins respectively.

4.5 Code availability

The R code used to compute cross coalescence and visualize the genealogy are available here `https://github.com/LukeAndersonTrocme/genes_in_space/tree/main/supplementary_code`.

5 Enrichment analysis

Using three metrics measuring the relatedness of individuals in a pair of towns – migrations, identity by descent, and cross coalescence – this enrichment analysis tests the null hypothesis that pairs of towns at a given distance have equivalent relatedness rates regardless of whether they share a watershed. To compare sets of towns of equal distance, we define T to be a set of distal towns within a twenty kilometre wide annulus whose inner radius is d kilometres from a reference town b and S be a set of distal towns within the same watershed as b. The distal towns that are in the intersection of the sets T and S (i.e. they are within the same watershed as b and within the annulus defined by d)

$$T' = T \cap S. \tag{S9}$$

From this, we can define the baseline of our enrichment as the fraction of distal towns – with sampled individuals – sharing a watershed with reference town b within a fixed distance d

$$c(b) = \frac{|T'|}{|T|}. \tag{S10}$$

For each of the three statistics considered for watershed enrichment, we define

$$\omega_m(b, a) \tag{S11}$$

as the value of metric m between reference towns b and distal towns a. The sum of $\omega_m(b, a)$ over sets of distal towns T' and T are

$$\Omega_m(b, T') = \sum_{a \in T'} \omega_m(b, a), \tag{S12}$$

and

$$\Omega_m(b, T) = \sum_{a \in T} \omega_m(b, a), \tag{S13}$$

respectively.

From this, we define the fraction $\eta_m(b)$ of $\Omega_m(b)$ distal towns sharing a watershed with reference town b within a fixed distance d as

$$\eta_m(b) = \frac{\Omega_m(b, T')}{\Omega_m(b, T)}. \tag{S14}$$

Finally, we define the enrichment of metric m for distal towns sharing a watershed with reference town b within a fixed distance d as

$$\epsilon_m(b) = \frac{\eta_m(b)}{c(b)}. \tag{S15}$$

Example

To illustrate our enrichment metric, let us consider a null model where

$$\omega_m(b, t) = 1$$

for all pairs of distal towns t and reference towns b. In this case,

$$\Omega_m(b, T') = T'$$

and

$$\Omega_m(b, T) = T,$$

where

$$\eta_m(b) = \frac{T'}{T} = c(b),$$

which yields

$$\epsilon(b) = \frac{\eta_m(b)}{c(b)} = 1.$$

Enrichment metrics

As mentioned in the section above, the three metrics used in our enrichment analysis are IBD, migrations, and cross coalescence.

The average length of DNA that is IBD between individuals in a given town a and a reference town b similar to 3.4.1:

$$\omega_{IBD}(b,a) \equiv g(a,b) = \frac{1}{|P^a||P^b|} \sum_{i \in P^a, j \in P^b} IBD(i,j), \tag{S16}$$

where $IBD(i,j)$ is the total length in centiMorgans of IBD segments between individuals i and j in towns a and b respectively.

The emigration rate from a given town a to a reference town b defined in 4.3.2 :

$$\omega_{mig}(b,a) \equiv \delta_{a \to b} = \frac{|M_{a \to b}|}{N_a}, \tag{S17}$$

where N_a are the number of people born in distal town a.

The relative cross coalescence between individuals in a given distal town a and a reference town b defined in 4.2.3 :

$$\omega_{coal}(b,a) \equiv \gamma^{ba} = \frac{2\Lambda^{ba}}{\Lambda^b + \Lambda^a}, \tag{S18}$$

where Λ^{ba} is the ratio of relative cross coalescence and Λ^b and Λ^a are the relative within-population coalescences for probands in b and a respectively.

Implementation

Because the Canadian National Hydrographic Network classifies the hundreds of kilometres of shoreline of the St. Lawrence River as one single watershed, we excluded this unusual watershed from the analysis by removing all reference towns within this watershed. We note that distal towns within this watershed remain in the analysis.

Supplementary Text

5.1 Additional Explanation of Results

Supplementary Figure S2B shows that there is some variation in IBD sharing across regions in Quebec. The three regions (SLSJ, BSL, and BCE) with the highest IBD sharing with France share some interesting features: they each have a substantial sample size, they define the top three principal components, and they have experienced founder events within the genealogy (see Figures 2 and 4). A few factors could in principle contribute to this excess IBD.

First, there could be excess recent migration from France into these regions. However, we could not find historical or genealogical evidence for this, and the rest of the genetic information does not particularly support this hypothesis (i.e., we found no evidence that a distinct wave of migration was helpful to describe principal components or f statistics. The fact that Quebec regions with the most IBD with France also have experienced bottlenecks suggests that the bottleneck itself may have contributed to the higher IBD. A second hypothesis is then that the bottlenecks leads to longer shared segments with French individual that are formed by the concatenation of distinct IBD segments. In other words, a past recombination had the effect of bringing together two IBD segments, making this recombination invisible to the IBD calling software. Such invisible recombinations are more common in populations having experienced a bottleneck: recombinations that bring together two lineages that coalesced within the genealogy will be invisible. The proportion of such invisible recombinations in a randomly mating population depends on the rate of pairwise coalescence. The highest rate of regional pairwise coalescence we have observed is 0.005 in SLSJ, meaning that approximately 0.5%. of recombinations would be expected to be invisible in that model, which is likely not enough to explain the discrepancy. Extensive inbreeding could further increase the rate of coalescence among recombining lineages, but this is also low (*68*).

A third and more likely explanation also relates to the bottleneck but in a technical way: our IBD calling pipeline requires phasing the data. Inaccurate phasing can lead to missed IBD segments. The three regions with higher detected IBD share a high coalescence rate and large sample

sizes. As a result, we expect that the phasing algorithm is more likely to correctly phase long segments of the genome in these regions, as the probability that an individual in SLSJ shares an IBD segment with any of the 707 other participants from SLSJ is high – given a genealogical coalescence rate of 0.005, each individual shared a genealogical common ancestor with $707 \times 0.005 = 3.5$ other individuals. Thus long-range phasing is likely more accurate in a founder population with a large sample size.

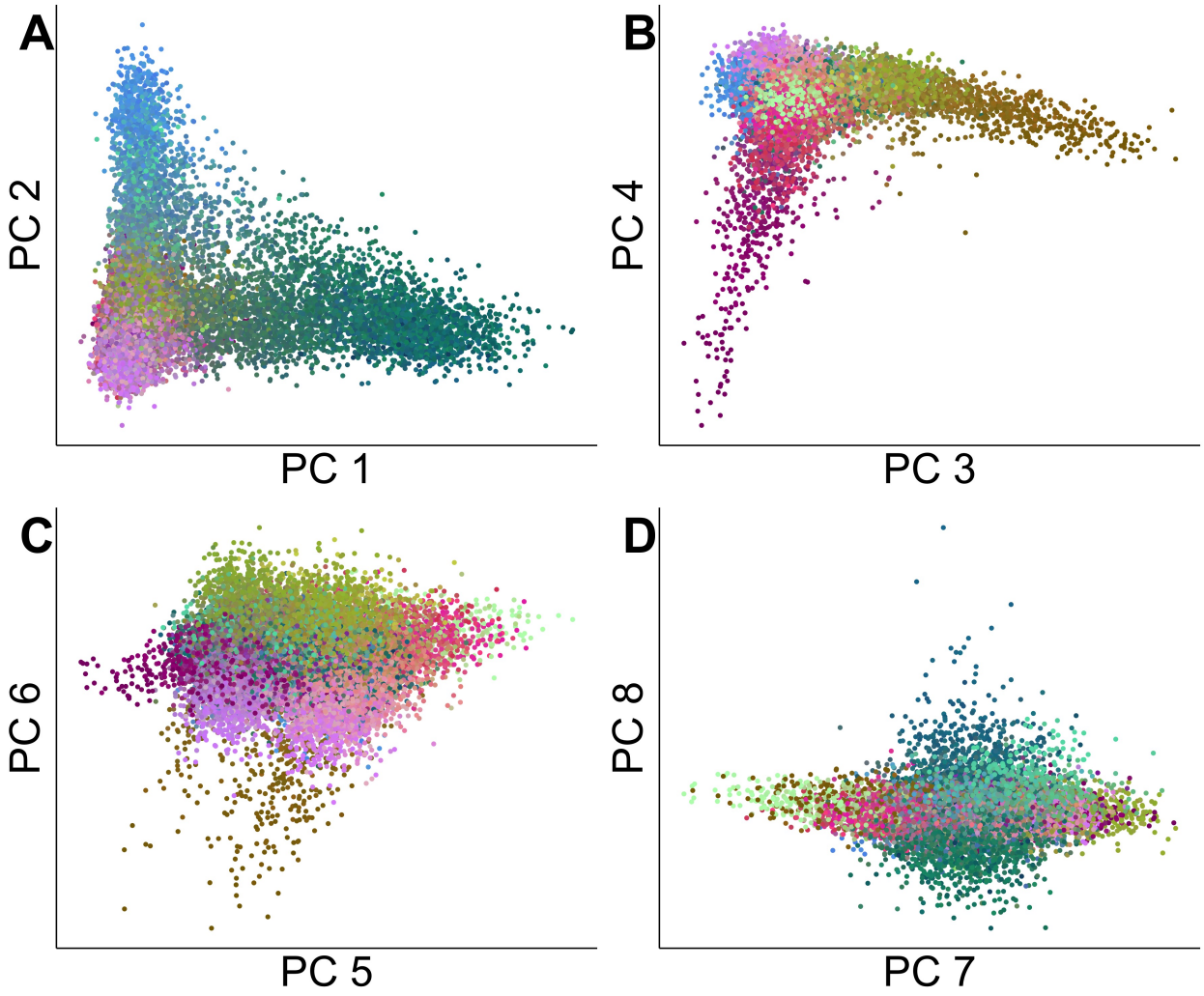

Fig. S1: **Principal Component Analysis of French Canadians.** (**A-D**) The top eight principal components of the genotype data included in the analyses.

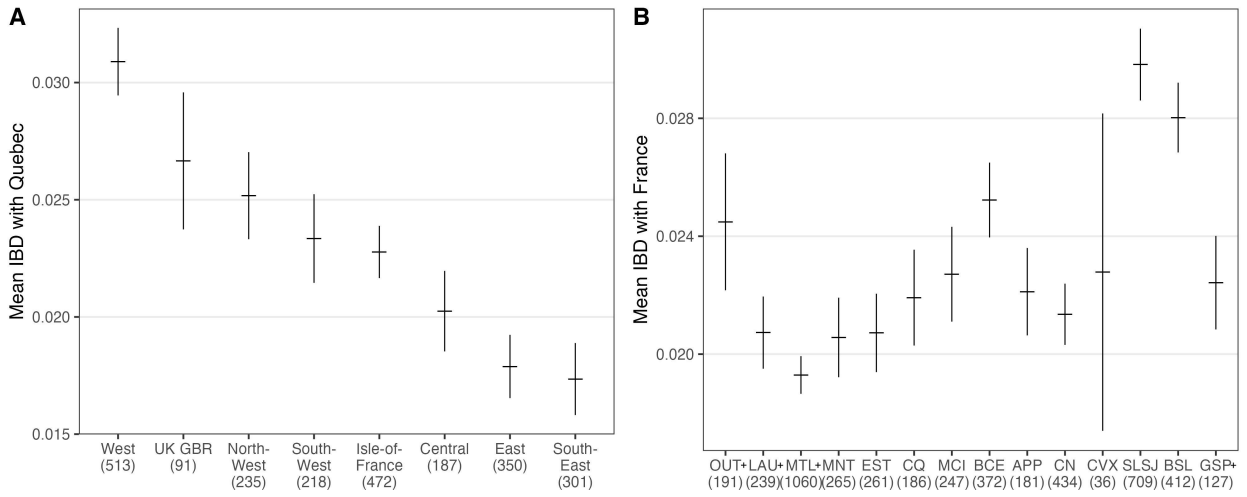

Fig. S2: **Rates of identity by descent in different regions.** The average IBD sharing between individuals living in Quebec, France, and Britain. The error bars in these plots indicate the standard error of the mean length (cM) of IBD sharing across individuals from the region indicated on the x-axis. As expected, regions with low sample sizes (indicated on the x-axis) have larger standard errors. (**A**) Individuals from Quebec have higher rates of IBD with individuals from West and North-West France than with individuals from Central and South-East France. (**B**) Individuals from France have higher rates of IBD with individuals from Saguenay, Beauce, and Bas-Saint-Laurent than with individuals from other regions. See section 5.1 for additional discussion. Abbreviations: OUT – Outaouais; LAU – Laurentides; MTL – Montréal; MNT – Montérégie; EST – Estrie; CQ – Centre-du-Québec; MCI – Mauricie; BCE – Beauce; APP – Appalaches; CN – Capitale-Nationale; CVX – Charlevoix; SLSJ – Saguenay–Lac-Saint-Jean; BSL – Bas-Saint-Laurent; GSP – Gaspésie. See S21 for details.

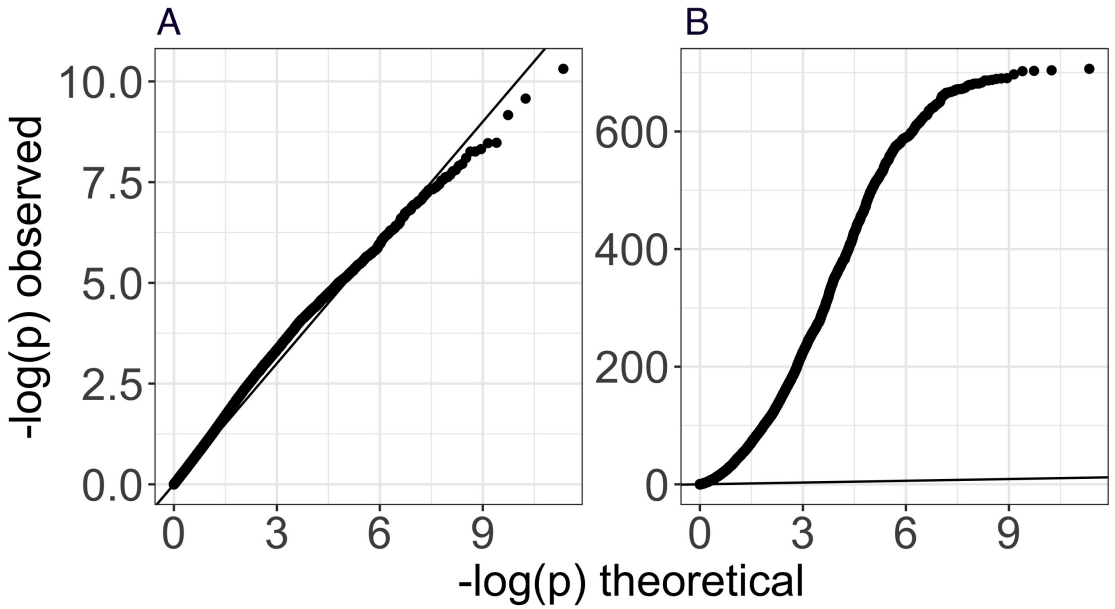

Fig. S3: **Comparison of Quebec, French and British regions using F_4-statistics. (A)** All 504 combinations of regions in Quebec (63 watersheds) and Europe (7 French regions and Britain) were compared using an F_4-statistic: F_4(Qc1, Qc2, Eu1, Eu2). The QQ plot shows no enrichment of significant *p*-values, consistent with the null hypothesis that European population structure is not strongly preserved in Quebec. **(B)** We computed complementary F_4 statistics comparing regions in Europe and regions in Quebec, to other regions in Europe and Quebec (F_4(Qc1, Eu1, Qc2, Eu2)), and find that all are significantly different, confirming that there are enough genetic differences between these populations to identify F_4 statistic signal.

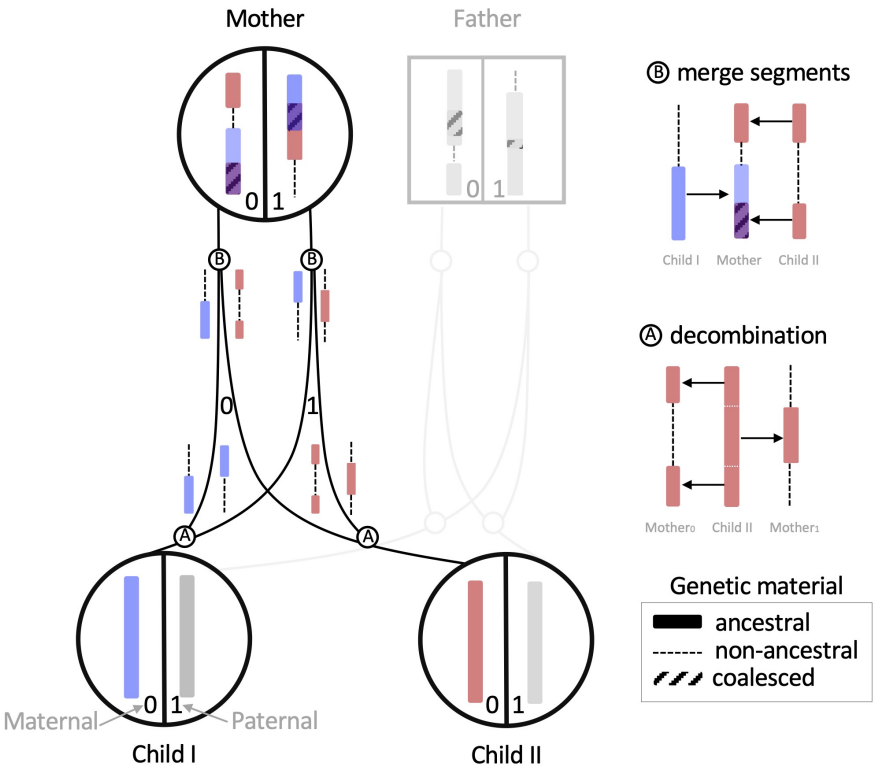

Fig. S4: **Illustrating decombination** (**A**) Ancestry simulations in backwards time *decombine* genetic material from children based on a user specified recombination map. (**B**) Decombined segments are then merged in a common ancestor. Not all of the genetic material contributed to children coalesces with other ancestral material.

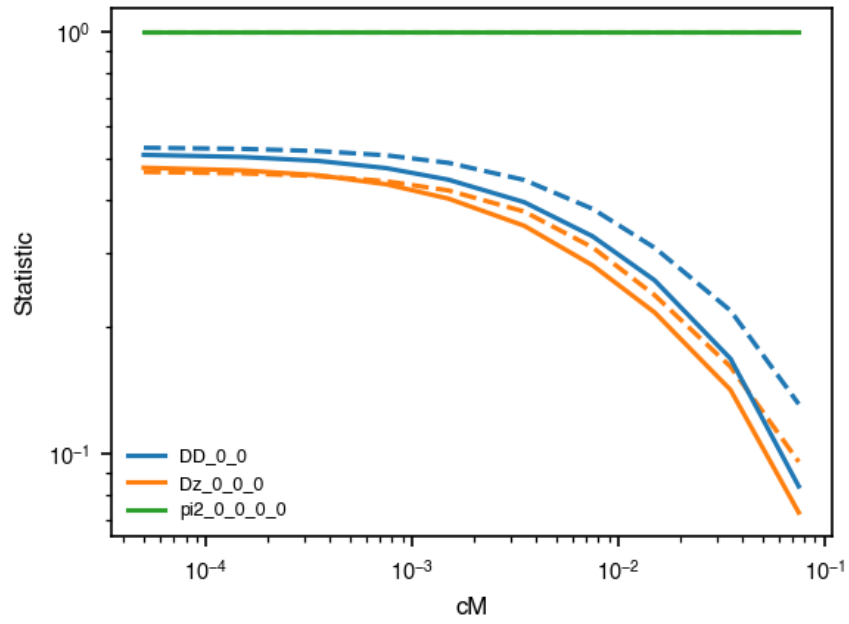

Fig. S5: **Validating the rescaled Tennessen model** After scaling the Tennessen out of Africa model to more realistic parameters, we confirmed that the predicted frequency spectra were identical (data not shown). We also compare predictions from our rescaled model (dashed lines) to predictions for the British population from the merger with migration model described in (*70*) and fit to two-locus statistics (solid lines).

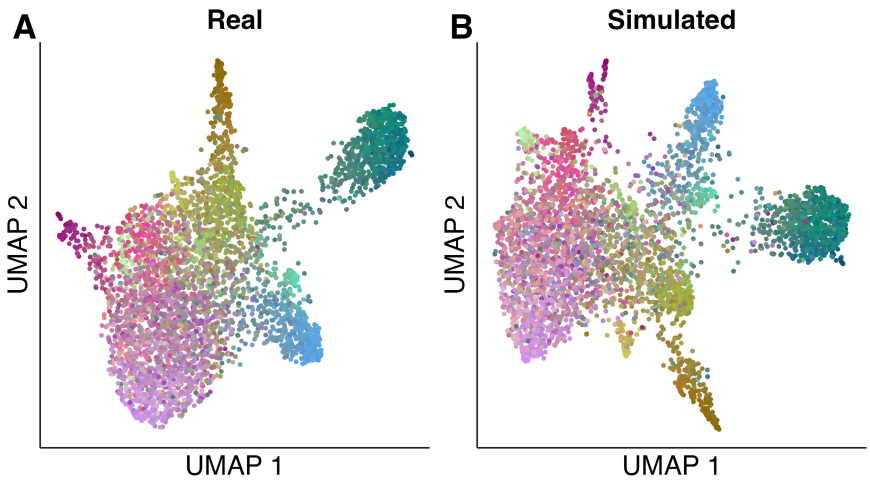

Fig. S6: **UMAP projections of observed and simulated genomes.** UMAP projections comparing the same 4,882 individuals using simulated and observed genomes (coloured as in Fig. 1).

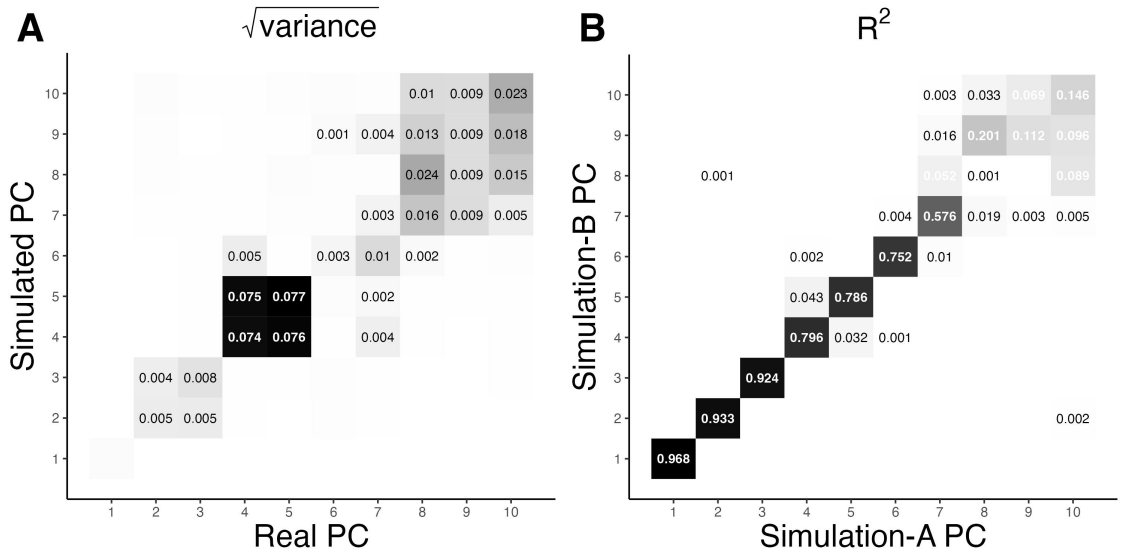

Fig. S7: **Variance across genome simulations.** **(A)** Using the same methods as shown in Figure 2, we compare the variance, across ten replicate simulations, of the R squared values compared to the same whole genome data. While there is some variance in PCs 4 and 5, the first three PCs show negligible amounts of variance across simulations. **(B)** Two replicate simulations compared to each other reveal a strong correlation between the top seven principal components. As noted in (A) there is an increase in variance for PCs 4 and 5. In both heatmaps (A) and (B) we label all values above 0.001, values below 0.001 are still shaded but not labelled.

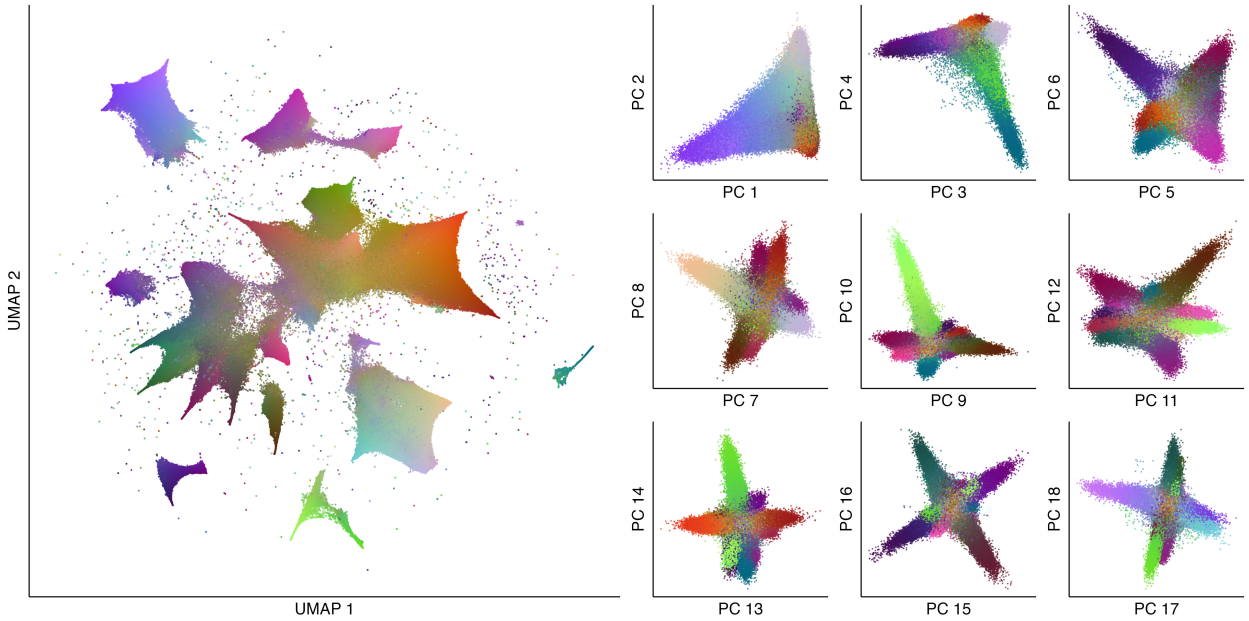

Fig. S8: **Visualizing 1.4 Million simulated French Canadian genomes** (**left**) A UMAP of the simulated genotype data. (**right**) The top eighteen principal components of the simulated genotype data. Colours were generated from a three dimensional UMAP through converting each x, y, z coordinate into an RGB value unique to each individual (see Supplementary Methods 3.2).

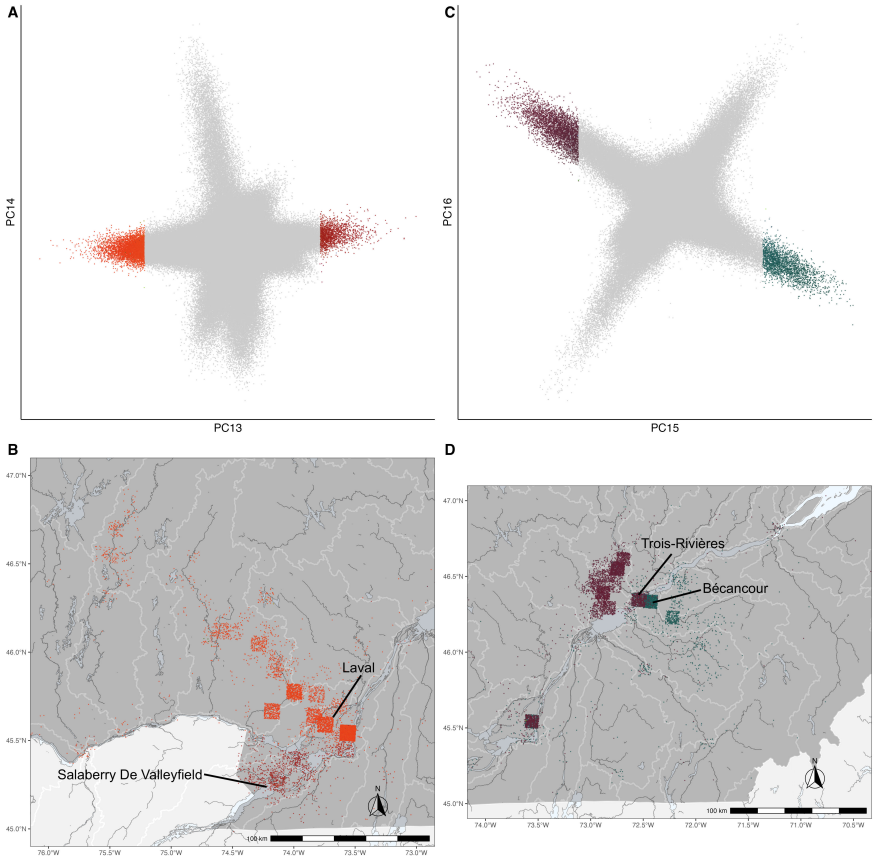

Fig. S9: **Geographic resolution of higher order principal components on 1.4 million simulated individuals** (**A**) Principal component analysis (PCA) of PCs 13 and 14. Light and dark red points are outliers defined by a 99.5% threshold along PC13. The grey points lie within the 99.5% threshold. Colours are consistent with Figure S8. (**B**) The geographic locations of the PC13 outlier points. The squares are the result of added noise to the geographic location of individuals to avoid over plotting. PC13 distinguishes neighbouring groups of people living on the North and South Shore of the St. Lawrence River. (**C**) PCA of PCs 15 and 16. Light and dark purple points are outliers defined by a 99.5% threshold along PC15 while also excluding individuals along the diagonal. The grey points lie within the 99.5% threshold. Colours are consistent with S8. (**D**) The geographic locations of the PC15 outlier points. The squares are the result of added noise to the geographic location of individuals to avoid over-plotting. PC15 distinguishes neighbouring groups of people living on the North and South Shore of the St. Lawrence River.

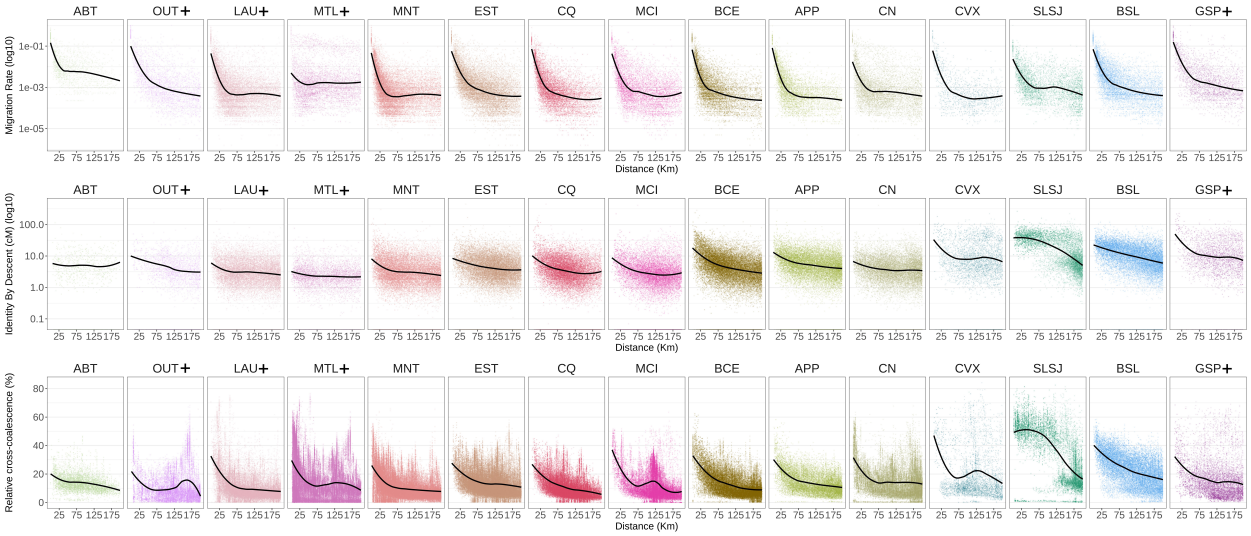

Fig. S10: **Migration rates, Identity by descent rates, and relative cross coalescence rates decay with distance for each region in Quebec.** The y axis is $\omega_m(b, a)$ for all metrics m for all reference towns b and distal towns a. The x axis is the distance in kilometres between reference towns b and distal towns a. Rows of panels show the decay of a single metric across regions in Quebec. Columns correspond to the regions defined in S21, with the exception of Abitibi-Témiscamingue (ABT) plotted separately from OUT+. The solid black line indicates a loess fit line using local polynomial regression fitting for each panel separately.

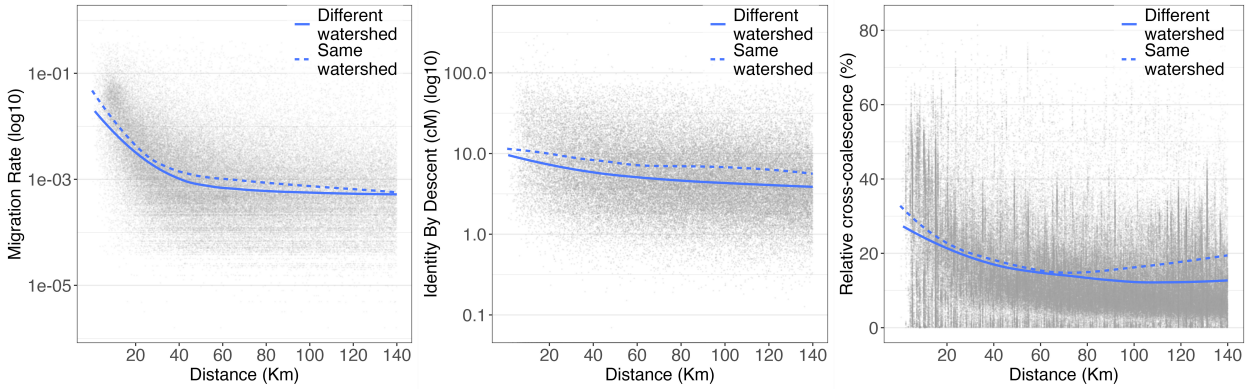

Fig. S11: **Watershed decay rates.** Migration rates, Identity by descent rates, and relative cross coalescence rates decay within and across watershed boundaries. The y axis is $\omega_m(b, a)$ for all metrics and for all reference towns b and distal towns a. The x axis is the distance in kilometres between reference towns b and distal towns a. The dashed and solid blue lines indicate a loess fit line using local polynomial regression fitting for metric m within and across watershed boundaries respectively.

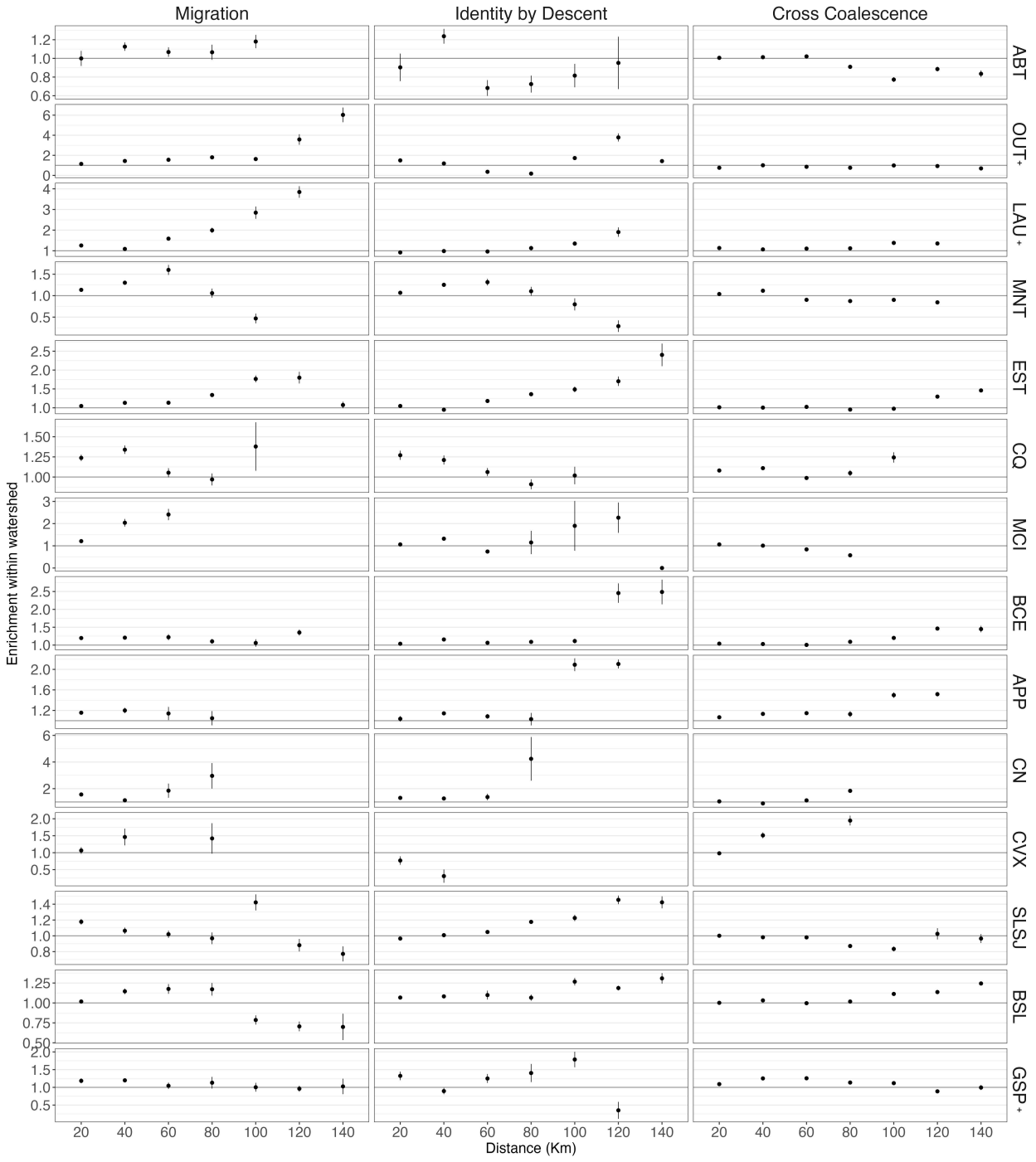

Fig. S12: **Historical migrations that define regional population structure** Watersheds influence migrations, genetic and genealogical relatedness. Rows of panels indicate a separate region's watershed enrichment comparing the migration rates IBD rates and cross coalescence rates between pairs of individuals in towns within and without the same watershed. Note the y scale along rows are consistent, but across rows are scaled separately.

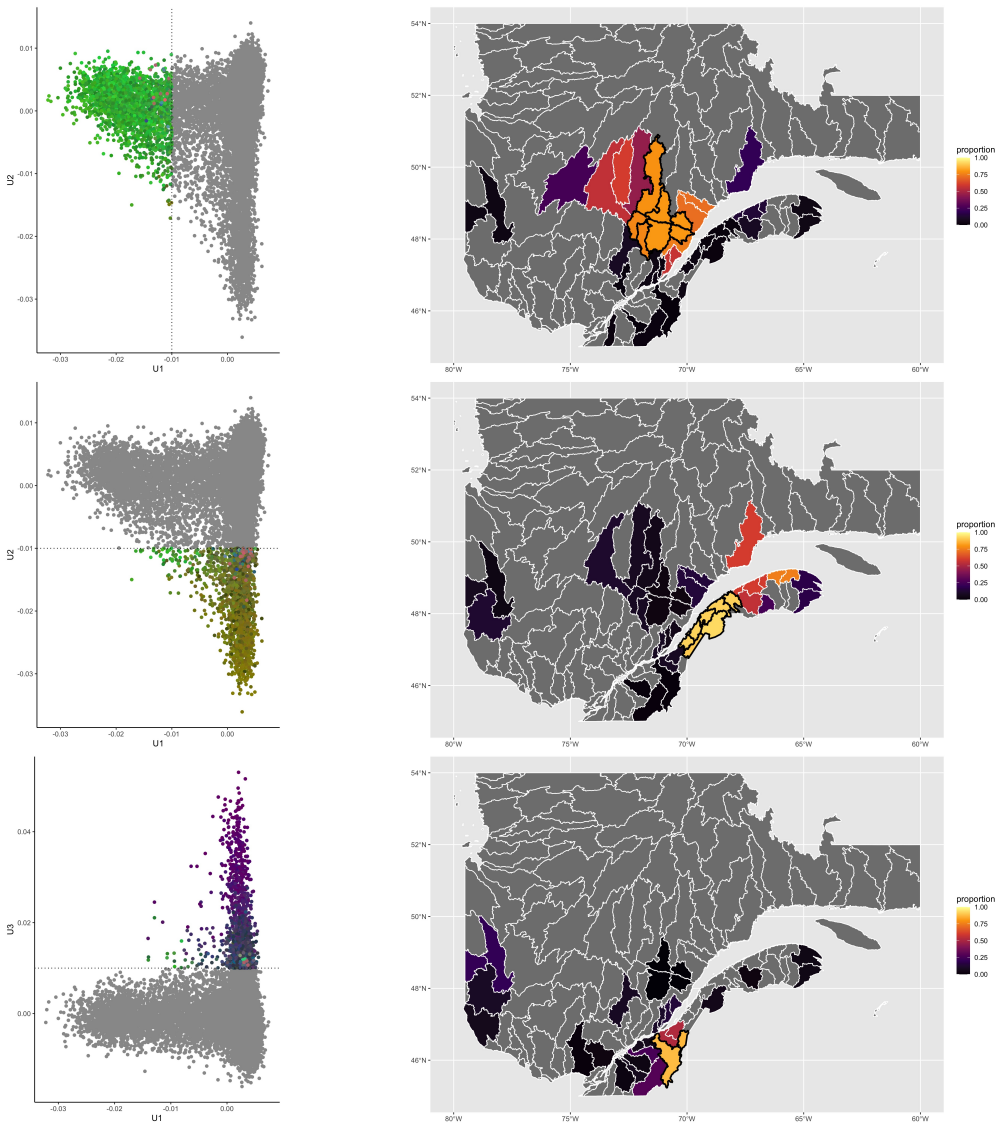

Fig. S13: **Selecting watersheds with individuals driving principal components** We define a threshold of 0.01 along each of the top three principal components and compute the fraction of individuals in each watershed that are beyond this limit. We choose watersheds with more than 75% of individuals beyond the threshold (black contours). These regions broadly correspond to Saguenay-Lac-Saint-Jean (PC1), Bas-Saint-Laurent (PC2), and Beauce (PC3).

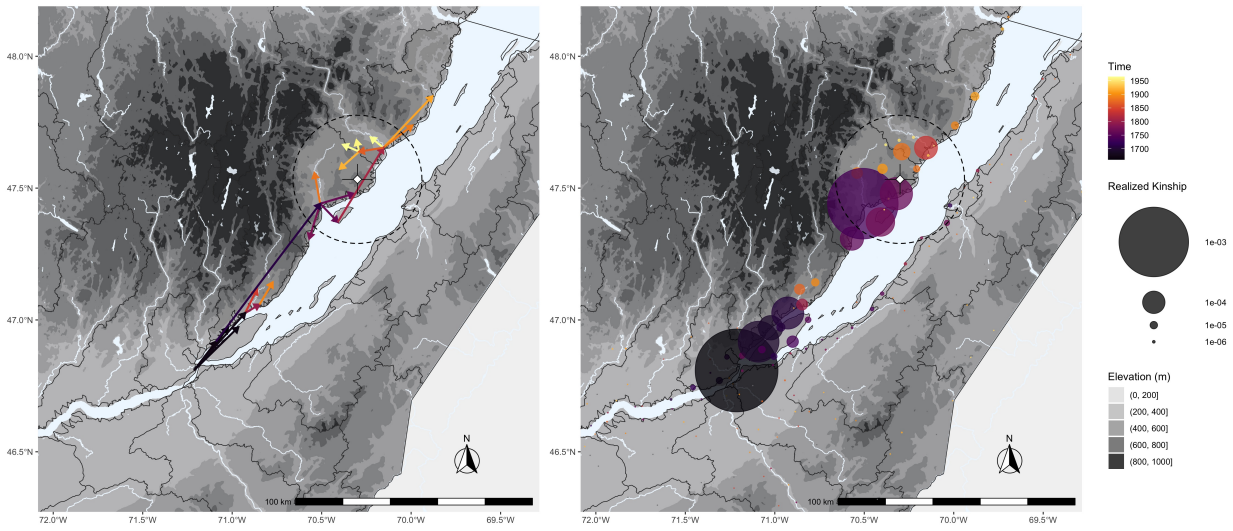

Fig. S14: **Charlevoix astrobleme impact on population structure** For present day individuals living in Charlevoix (highlighted area), we show **(A)** the major migratory axes as well as **(A)** the location, timing and stringency of population bottlenecks measured by realized kinship. The epicentre of the astrobleme is marked with a cross and the radius of the ancient meteor crater is indicated with a dotted line.

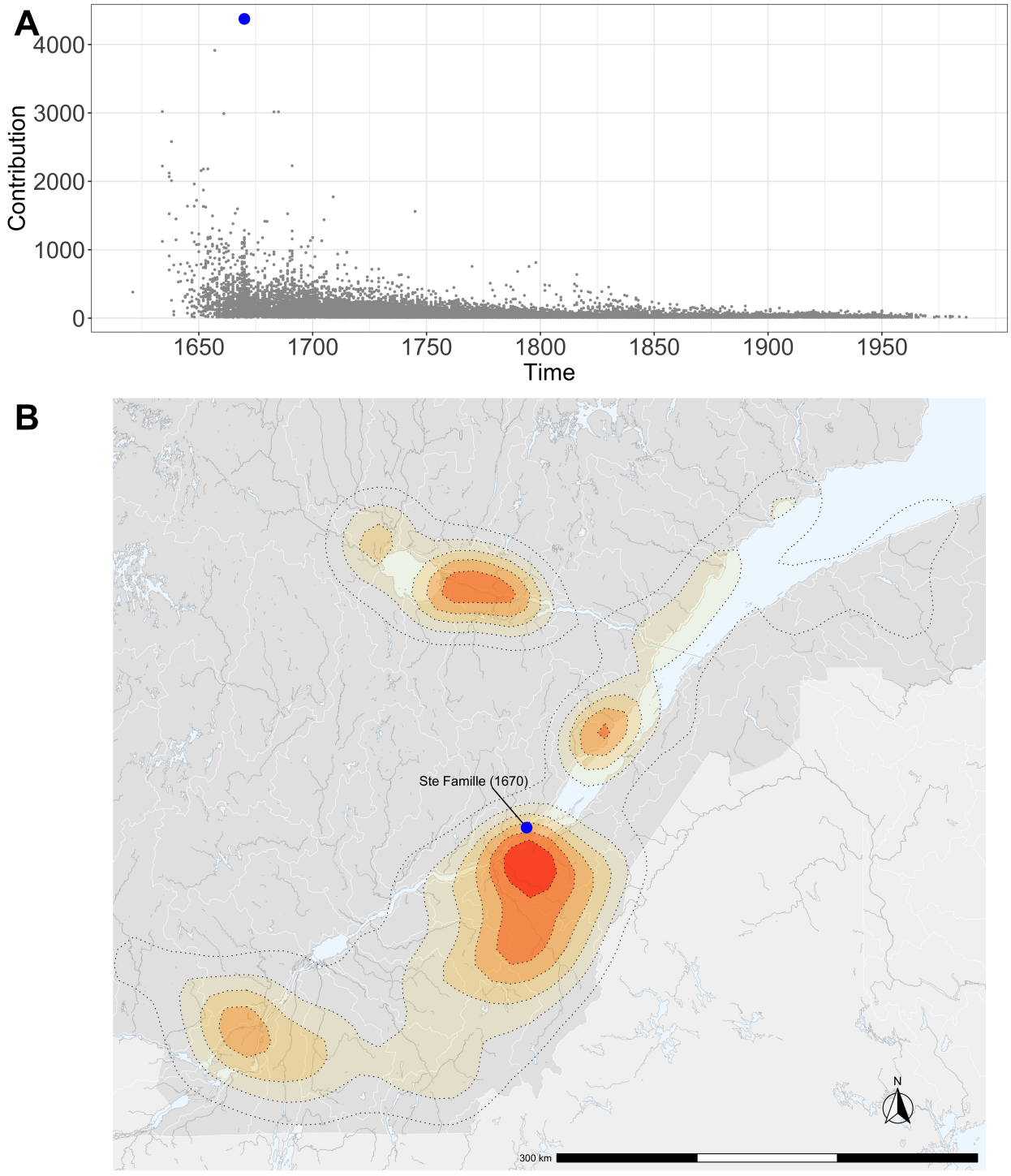

Fig. S15: **Range dispersal for the top contributing ancestor A** The estimated genetic contribution of all ancestors in the pedigree to the 1.4 million probands. A blue dot highlights the ancestor with the greatest contribution to individuals in the pedigree. **B** The range dispersal of the historical individual with the greatest contribution to probands.

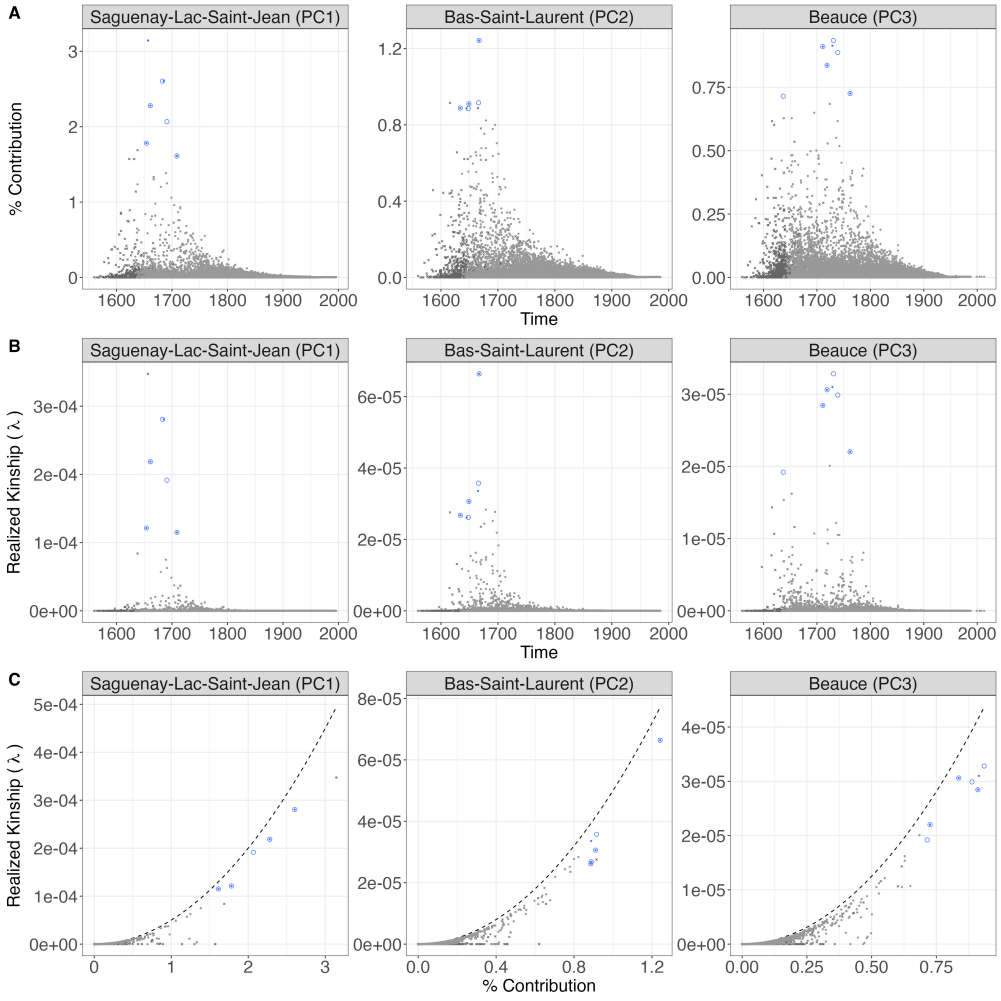

Fig. S16: **Ancestral contributions and realized kinship.** In each row of plots **(A)**, **(B)**, and **(C)**, the panels with PC1, PC2 and PC3 relate to the regions of Saguenay-Lac-Saint-Jean, Bas-Saint-Laurent, and Beauce respectively (see S13). Grey points represent the values of an individual ancestor in Quebec. Dark grey crosses represent the values of an individual ancestor not married in Quebec. Blue points and circles represent the top 10 contributing female and male ancestors respectively. **(A)** The proportion of DNA each ancestor is expected to have contributed to the present-day gene pool of individuals. The x axis places each ancestor in time based on their marriage date. **(B)** The realized kinship $\lambda^P(i)$ each ancestor i contributed to the present-day gene pool of individuals P. The x axis places each ancestor in time based on their marriage date. **(C)** Comparison of the realized kinship $\lambda^P(i)$ to the estimated proportion of DNA each ancestor contributed. The dashed line shows the upper bound $\lambda^P(i)^2/2$.

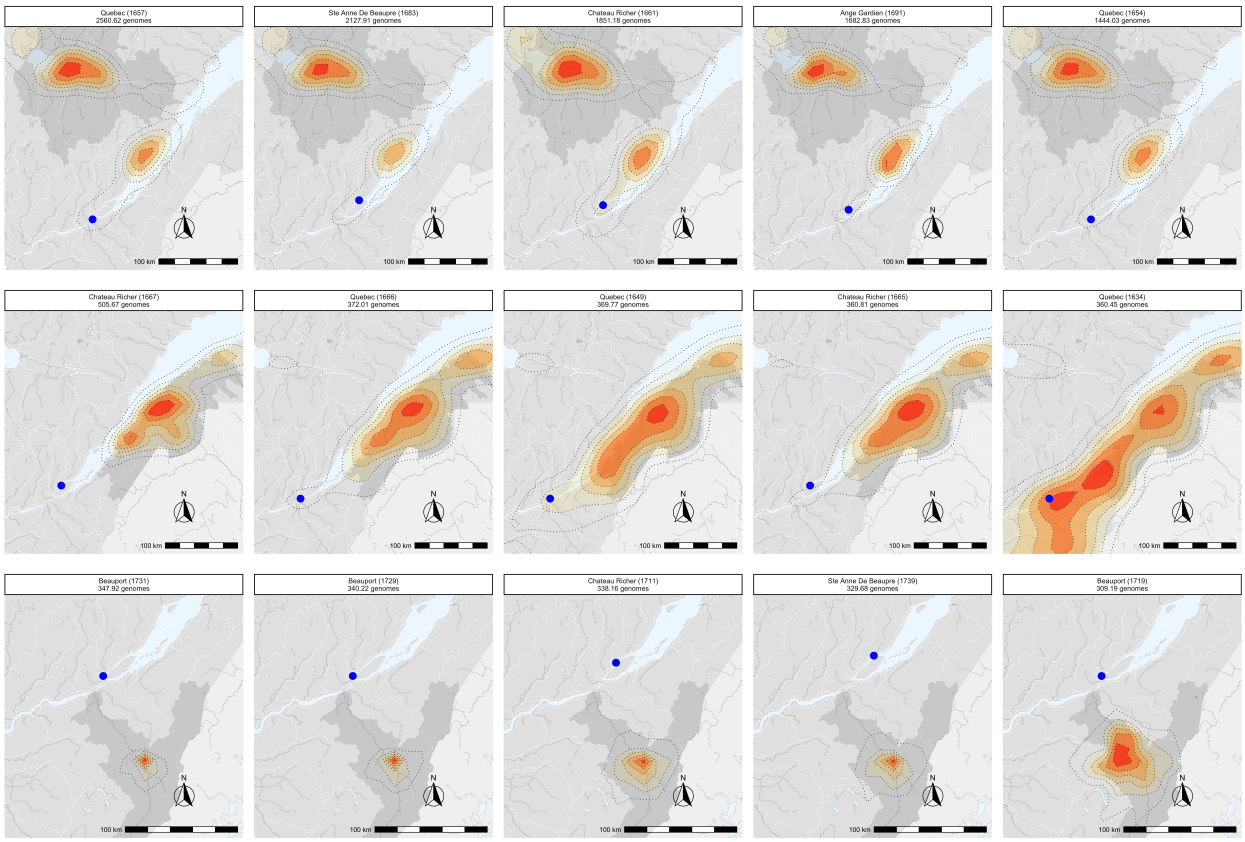

Fig. S17: **Dispersal range of the top five contributors to each region** For each region **(rows)** enriched in individuals driving each of the top three principal components, we show the dispersal range of the top five super-founders **(columns)**. Even though there is some variation between the dispersal ranges of the super-founders of a region, they broadly cover the same geographic regions. We note that when selecting the top super founders in Figure 4 and Figure S18, to avoid selecting both spouses from a couple, we selected the spouse that contributed the most (in instances of remarriage) or defaulted to the female spouse. This ensured that the top ten super-founders of a region would not be married to each other. The scaling of the heat maps are not comparable between plots as they were computed separately.

PC1	PC2	PC3
3.143%	0.296%	0.231%
2.605%	0.226%	0.16%
2.605%	0.226%	0.16%
2.279%	0.201%	0.172%
2.279%	0.201%	0.172%
2.067%	0.221%	0.038%
1.782%	0.18%	0.04%
1.782%	0.18%	0.04%
1.613%	0.192%	0.036%
1.613%	0.192%	0.036%
0.175%	1.242%	0.032%
0.175%	1.242%	0.032%
0.184%	0.916%	0.042%
0.201%	0.911%	0.099%
0.201%	0.911%	0.099%
0.276%	0.888%	0.461%
0.276%	0.888%	0.461%
0.163%	0.888%	0.039%
0.166%	0.886%	0.051%
0.166%	0.886%	0.051%
0.013%	0.009%	0.935%
0.013%	0.009%	0.914%
0.015%	0.01%	0.91%
0.015%	0.01%	0.91%
0.015%	0.009%	0.887%
0.017%	0.01%	0.836%
0.017%	0.01%	0.836%
0.011%	0.006%	0.726%
0.011%	0.006%	0.726%
0.65%	0.212%	0.715%

Fig. S18: **Percent contributions of regional super-founders.** For each region enriched in individuals driving each of the top three principal components, we identify the top ten 'super-founders' based on their estimated genetic contribution to each region. Using the French-Canadian genealogy, we also compute the estimated genetic contributions of these ancestors to individuals living in the other two regions to see how much overlap there between their descendants. We find that each of the super-founders disproportionately contributes to a single region.

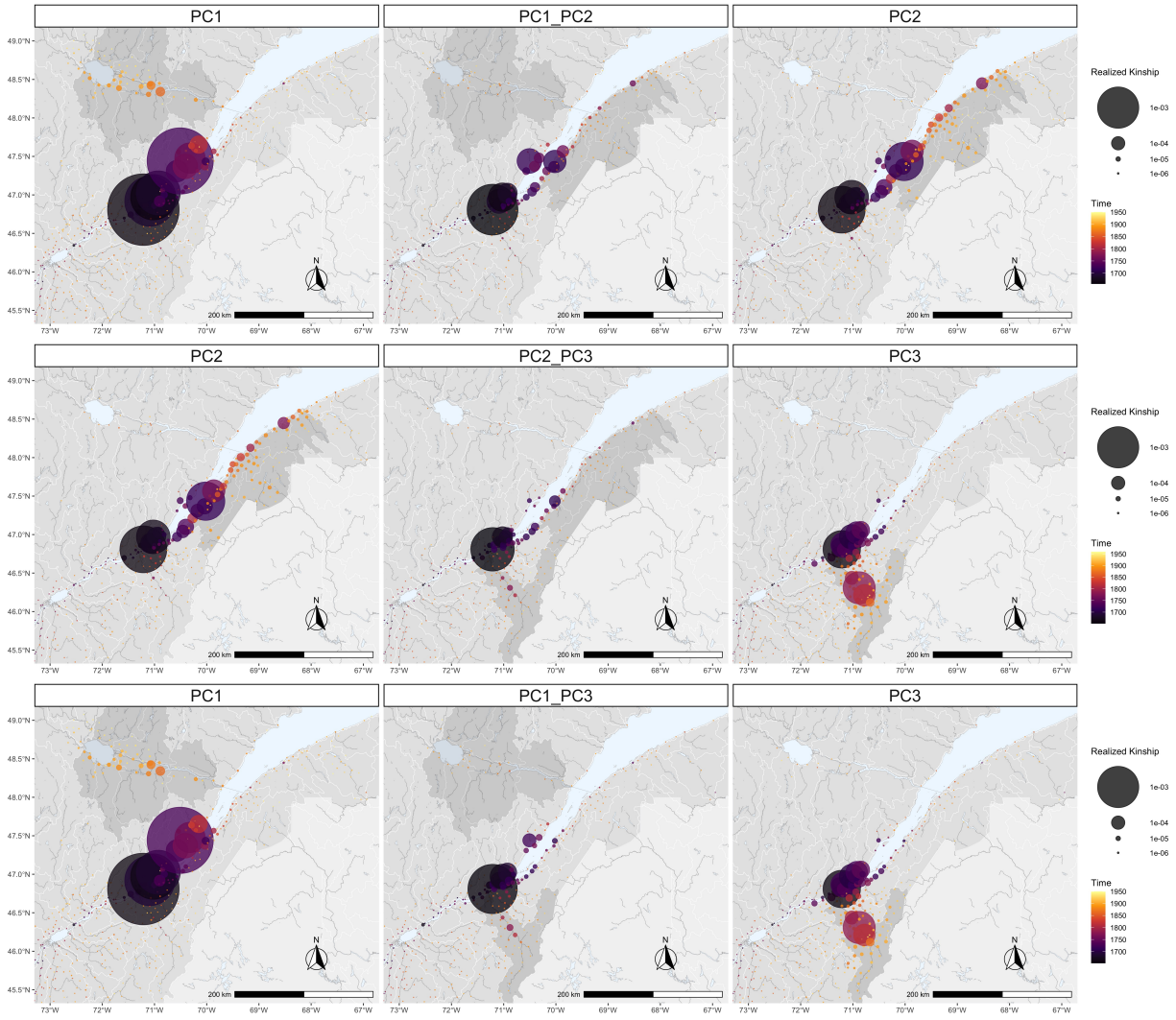

Fig. S19: **Cross coalescence rates** For each pair of regions defined in Figure S13, we compute the cross coalescence rates by ascending the genealogies of individuals in both regions. The central panels represent the between region cross coalescence rate, whereas the first and last column of panels represent the within region coalescence rate. We find that in all three cross coalescence rates, there is a common root in the region around Quebec City.

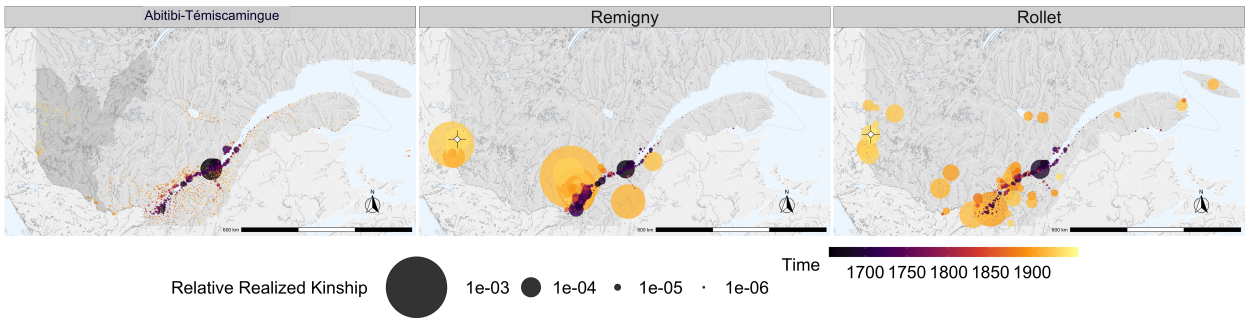

Fig. S20: **Realized kinship for Remigny and Rollet.** Abitibi-Témiscamingue (left panel) shows no major regional founding events other than in Quebec City in the 16th century. When we consider each town separately, bottlenecks become more apparent. The towns of Remigny (centre panel) and Rollet (right panel) have recent founding events from different regions in Quebec despite being twenty kilometres away from each other.

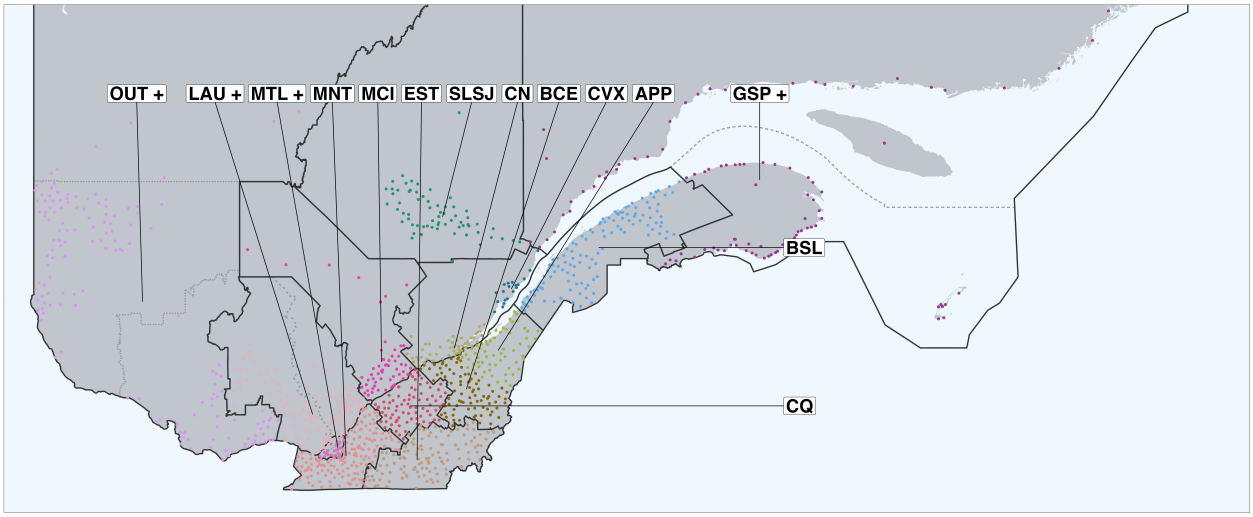

Fig. S21: **Boundaries of the 14 broad regions used to generate genealogy flow plot.** The genealogy flow plot in Figure 1 was generated based on Quebec administrative boundaries. Abbreviations: OUT + – Nord-du-Québec, Abitibi-Témiscamingue, Outaouais; LAU + – Lanaudière, Laurentides; MTL + – Laval, Montréal; MNT – Montérégie; EST – Estrie; CQ – Centre-du-Québec; MCI – Mauricie; BCE – Beauce; APP – Appalaches; CN – Capitale-Nationale; CVX – Charlevoix; SLSJ – Saguenay–Lac-Saint-Jean; BSL – Bas-Saint-Laurent; GSP + – Iles-de-la-Madeleine, Cote-Nord, Gaspésie. We modified the groupings of some regions because they had similar demographic history or small population sizes. The regions of Nord-du-Québec, Abitibi-Témiscamingue, and Outaouais were clumped into a single group. The regions of Laval and Montréal were clumped together. The regions of Lanaudière and Laurentides were clumped together. The regions of Cote-Nord, Gaspésie and Iles-de-la-Madeleine were also clumped together. In addition, we separated two administrative regions with demographic histories that we sought to visualize separately. The region of Charlevoix was separated from Capitale-Nationale and the region of Chaudière-Appalaches was separated into Beauce and Appalaches.

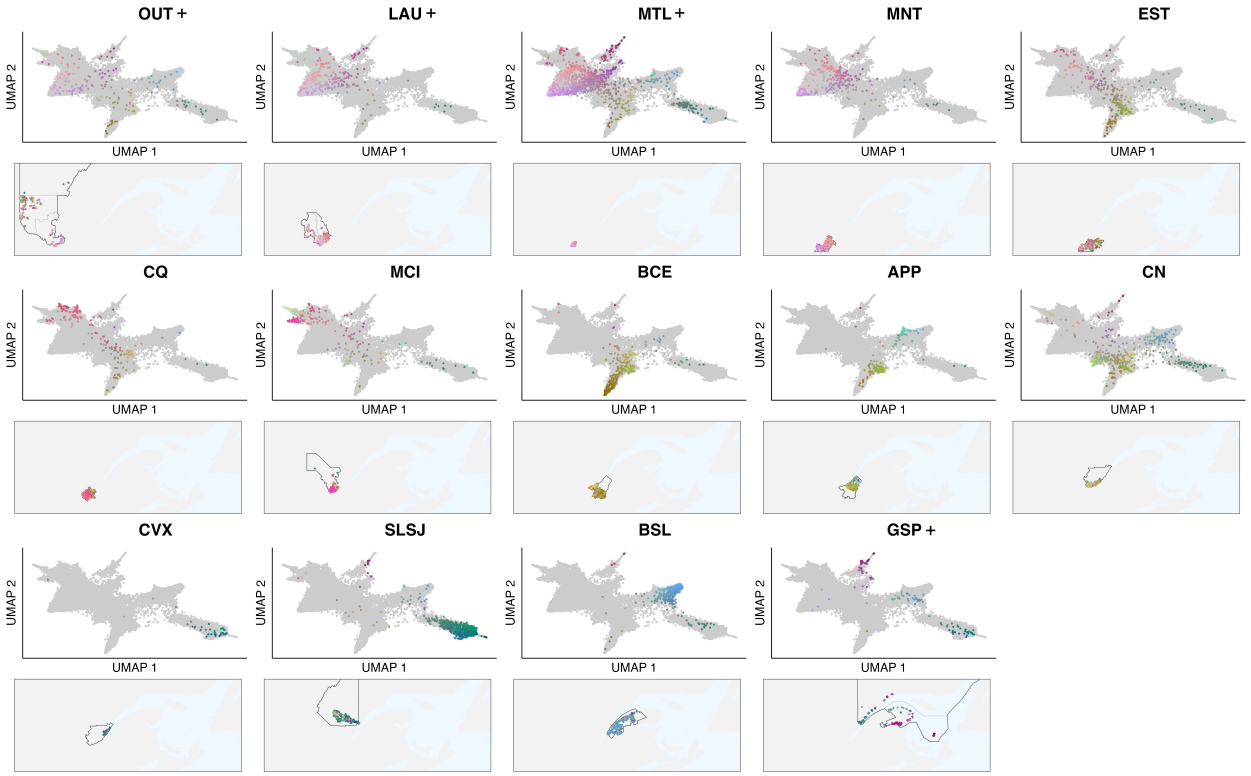

Fig. S22: **UMAP coordinates of individuals from each of the 14 broad regions used in the genealogy flow plot in Figure 1.** For each region used in Figure 1, we mark the outline of this boundary in black. In cases where we merged boundaries (e.g. OUT and GSP), we highlight the subdivisions in gray. We show the geographic location as well as their location in the UMAP projections to show how well these boundaries capture genetic groupings. Abbreviations: OUT – Outaouais; LAU – Laurentides; MTL – Montréal; MNT – Montérégie; EST – Estrie; CQ – Centre-du-Québec; MCI – Mauricie; BCE – Beauce; APP – Appalaches; CN – Capitale-Nationale; CVX – Charlevoix; SLSJ – Saguenay–Lac-Saint-Jean; BSL – Bas-Saint-Laurent; GSP – Gaspésie. See S21 for details.

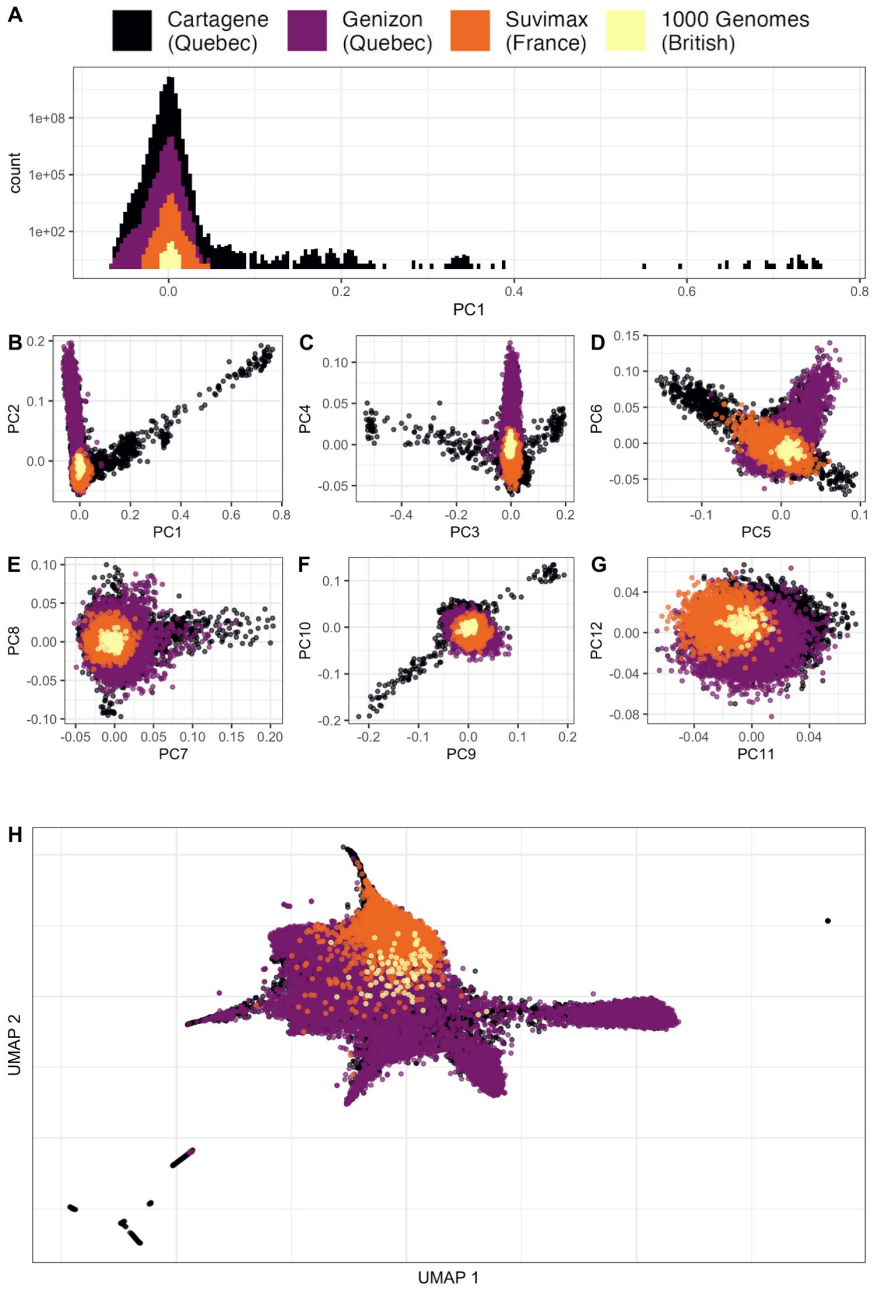

Fig. S23: **Visualization of the PCA and UMAP analyses of the complete genotype dataset.** **A** Is the distribution of samples along the first principal component using all samples form all cohorts. We colour each cohort separately to highlight any possible batch effects or population differentiation. **B-G** Are the top twelve principal components of all samples form all cohorts. **H** Is a UMAP of these samples using the top ten principal components.

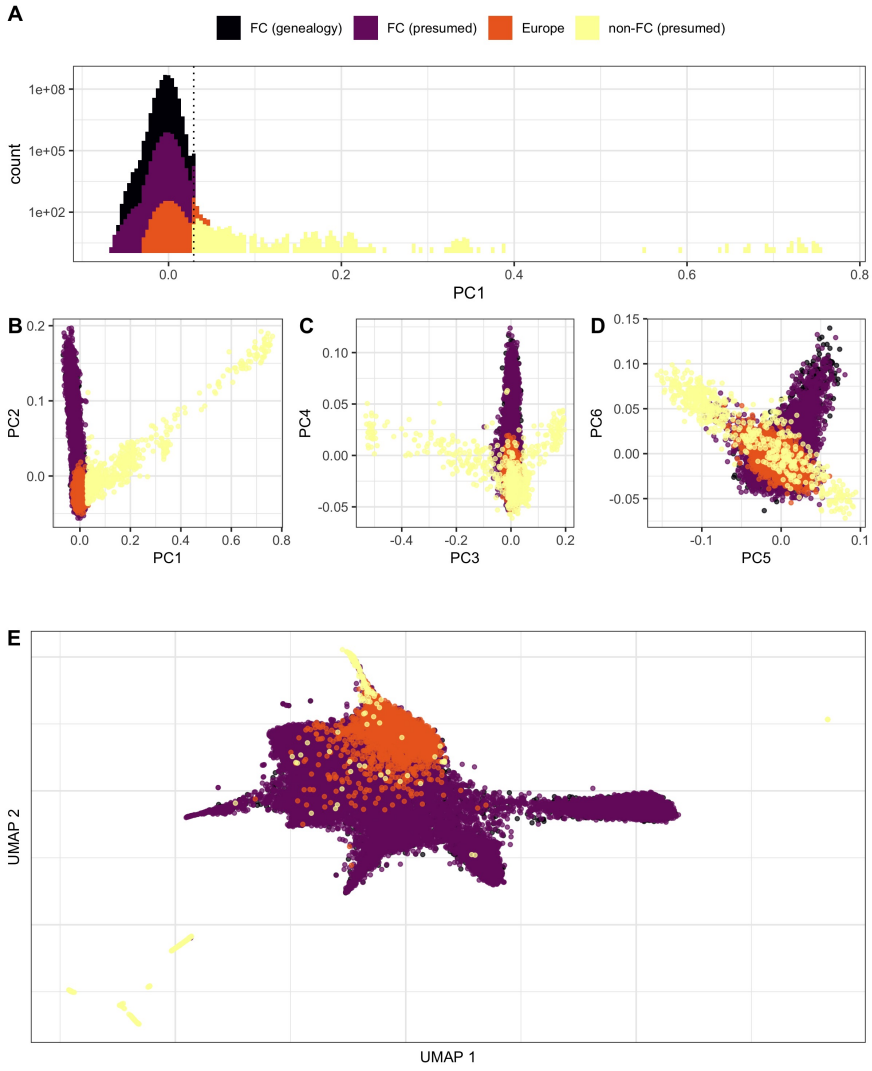

Fig. S24: **Visualization of the thresholds used to define presumed French Canadian ancestry.** **A** Is the distribution of samples along the first principal component using all samples form all cohorts. We colour samples based on their ancestries. In this case, we consider samples from France and from Britain as European, and samples from Quebec as either genealogically confirmed French Canadian ancestry, presumed French Canadian ancestry using a threshold based on the maximum value along the first principal component of genealogically linked individuals, and non-French Canadian ancestry for individuals beyond this threshold. **B-G** Are the top twelve principal components of all samples form all cohorts coloured based on presumed ancestry. **H** Is a UMAP of these samples using the top ten principal components coloured based on presumed ancestry.

Place-Name	Type	Notes
Chaudière River	Geographical feature	Major river streaming from Southeast Quebec across the Beauce region and into the Saint-Lawrence River. Important transportation axis for First Nations, French settlers, and contemporary commerce.
Saguenay River	Geographical feature	Major river streaming from the Lac-Saint-Jean to the Saint-Lawrence river. Important trade route for First Nations and French settlers, and contemporary commerce.
Saint-Lawrence River	Geographical feature	Major river streaming from the Great Lakes into the Atlantic Ocean. Major transportation axis for First Nations, French settlers, and contemporary commerce.
Charlevoix Astrobleme	Geographical feature	Lowland region in an otherwise mountainous area created by an asteroid impact 400Mya years ago. The limited area of arable land led to a founder event relevant to the Charlevoix region and the SLSJ.
Quebec City	Town	Located on the Saint-Lawrence river. First permanent French settlement in Quebec and landing location of most early French settlers; located in Capitale Nationale administrative region; site of shared founder event across multiple Quebec regions.

Table S1: Glossary of Quebec place-names, continued on next page.

Place-Name	Type	Notes
Charlevoix (CVX)	Geographic region	located in the Capitale-Nationale administrative region between Quebec City and the Saguenay River. Place of origin of an important proportion of SLSJ settlers. Characterized by mountainous terrain and the Charlevoix astrobleme.
Saguenay-Lac-Saint-Jean (SLSJ)	Administrative region	Region surrounding the Saguenay River and the Saint-Jean Lake. Open to French Canadian settlement in 1838 attracting a high proportion of settlers from the Charlevoix region. This led to a founder event followed by major population expansion. Extensively studied population. We also use this label to refer to an empirically defined region corresponding to PC1 in our data (Figure 4A).
Bas-Saint-Laurent (BSL)	Administrative region	Located along the south shore of the Saint-Lawrence river, the region is traversed by the Appalachian Mountains. French settlement occurred initially along the littoral, and later inland. We also use this label to refer to an empirically defined region corresponding to PC2 in our data (Figure 4B).
Beauce (BCE)	Geographic Region	Located on the south shore of the Saint-Lawrence river; surrounds the Chaudière River, an affluent of the Saint-Lawrence. Part of the Chaudière-Appalaches administrative region. We also use this label to refer to an empirically defined region corresponding to PC3 in our data (Figure 4C).
OUT+	Geographic region	Broad geographic region in North and Western Quebec combining administrative regions Outaouais, Abitibi-Témiscamingue, and Nord-du-Québec.
MTL+	Geographic region	Geographic region of the Montreal metropolitain area, combining administrative regions of Montreal and Laval
LAU+	Geographic region	Broad geographic region north of Montreal combining administrative regions Lanaudière and Laurentides.
GSP+	Geographic region	Broad geographic region in Eastern Quebec, combining administrative regions Gaspésie-Iles-de-la-Madeleine and Côte-Nord.

Table S1: Glossary of Quebec place-names (Continued)

Place-Name	Type	Notes
Normandy	Historical Region	Historical and administrative region in the North-West France, from which many French settlers originated (*28*). Includes Seine Maritime (76), Eure (27), Calvados (14), Orne (61), Manche (50) departments.
Ile-de-France	Historical Region	Historical and administrative region in France from which many French settlers originated (*28*). Includes the agglomeration of Paris, and has historically been a heavily populated area
Aunis	Historical Region	Historical region in in the West of France from which many French settlers originated (*28*). 18th century boundaries corresponds to part of the Charente-Maritime department (17) in Nouvelle-Aquitaine
Poitou	Historical Region	Historical region in the West of France from which many French settlers originated (*28*). 18th century borders correspond roughly to present-day Vendée (85, Pays de la Loire), Deux-Sèvres (79, Nouvelle Aquitaine), Vienne (86, Nouvelle Aquitaine)
Perche	Historical Region	Former province broadly in the North-West France from which many French settlers originated (*28*). 18th century boundaries correspond to Eure (27, Normandy), Orne (61, Normandy), Eure-Et-Loire (28, Centre, Val-de-Loire)
South-East	Geographic region	Includes departments 1, 4, 5, 6, 7, 13, 26, 38, 42, 69, 73, 74, 83, 84.
South-West	Geographic Region	Includes departments 9, 11, 12, 24, 30, 31, 32, 33, 34, 40, 46, 47, 48, 64, 65, 66, 81, 82.
West	Geographic region	Includes administrative regions and departments: **Nouvelle Aquitaine**: 16 (Charente), 17 (Charente-Maritime), 79 (Deux-Sèvres), 86 (Vienne); **Bretagne**: 22 (Cotes-d'Armor), 29 (Finistère), 35 (Ille et Vilaine), 56 (Morbihan); **Pays de la Loire**: 44 (Loire-Atlantique), 49 (Maine et Loire), 53 (Mayenne) , 72 (Sarthe), 85 (Vendée)
North-West	Geographic region	Includes administrative regions and departments: **Hauts-de-France**: 2 (Aisne), 59 (Nord), 60 (Oise), 62 (Pas-de-Calais), 80 (Somme); **Normandy**: 14 (Calvados), 27 (Eure), 50 (Manche), 61 (Orne), 76 (Seine-Maritime)
Central	Geographic region	Includes departments 3, 15, 18, 19, 23, 28 (Eure-Et-Loire), 36, 37, 41, 43, 45, 63, 87
Ile-de-France	Geographic region	Also used as a historical region. As a geographic region, includes departments 75, 77, 78, 91, 92, 93, 94, 95.
East	Geographic region	Includes departments 8, 10, 21, 25, 39, 51, 52, 54, 55, 57, 58, 67, 68, 70, 71, 88, 89, 90.

Table S2: Glossary of French place-names

Cohorts				
Quebec (Cartagene)	Quebec (Genizon)	France (Suvimax)	GBR (1kGP)	Total
12,064	9,004	2,276	91	23,435
Inferred Ancestry				
FC (genealogy)	FC (inferred)	non-FC (inferred)	Europe	Total
4,882	15,569	617	2,367	23,435

Table S3: Sample sizes of genotyped cohorts and their ancestries. A total of 20,451 individuals genealogically linked or genetically inferred French Canadian (FC) ancestry are used for visualizing population structure using PCA and UMAP.

Parameter	**Specification**	**Citation**
demographic model	European ancestry in two population out-of-Africa model	(*32*), rescaled (Section 2.2)
coalescent model	Hudson	(*69*)
mutation rate	1.66×10^{-8}	(*58*)
recombination map	GRCh37 hapmapII genetic map	(*60*)

Table S4: The model parameters used to simulate the ancestry of individuals in the French-Canadian pedigree.

List of Symbols

Symbol	Definition	Obs.Exp.	Unit
$IBD(i,j)$	length of IBD segments between individuals i and j	O	centiMorgans
$g(A,B)$	mean IBD sharing between pairs of individuals in towns A and B	O	centiMorgans
P^A	probands in town A	O	set
$\|P^A\|$	sample size in town A	O	-
$\mathrm{K}^p(i)$	genetic contribution of individual i to proband p	E	genomes
$\mathrm{K}^P(i)$	total genetic contributions of individual i to proband set P	E	genomes
$\mathrm{K}^P(I)$	total contributions of a set I of individuals to proband set P	E	genomes
$\lambda^P(i)$	kinship realized in individual i given a set of probands P	E	probability for time of first coalescence for all pairs of lineages
$\lambda^{AB}(i)$	kinship realized in individual i given sets of probands A and B	E	probability for time of first coalescence for all pairs of lineages
γ^{AB}	ratio of cross-coalescence to within-population coalescence of sets of probands P^A and P^B	E	probability for time of first coalescence for all pairs of lineages
Λ^{AB}	cross-coalescence rate of the sets of probands P^A and P^B	E	probability for time of first coalescence for all pairs of lineages
Λ^A	within-population coalescence rate of set of probands P^A	E	probability for time of first coalescence for all pairs of lineages
$M_{a\to b}$	set of individuals born in source-town a and married in sink-town b	O	set
$\mathrm{K}^P(M_{a\to b})$	genetic contribution of individuals born in a and married in b	E	genomes
$d(M_{a\to b})$	mean contribution year of individuals born in a and married in b	E	years
$\|M_{a\to b}\|$	number of migrants from a to b	O	individuals
N_a	number of individuals born in a	O	individuals
$\delta_{A\to b}$	inbound migration rate of set A of source-towns to sink-town b	O	proportion
T	towns within a defined annulus centred on town b	O	set
S	towns within the same watershed as b	O	set
T'	towns within the same watershed and annulus centred on town b : $T \cap S$	O	set
$c(b)$	fraction of towns sharing a watershed with b within a defined annulus : $\frac{\|T'\|}{\|T\|}$	O	proportion
m	metric whose enrichment is being considered (i.e. migration rate, IBD sharing rate, cross coalescence rate)	-	-
$\omega_m(b,a)$	value of metric m for reference town b and town a	-	-
$\Omega_m(b,T)$	weighted sum of metric m for reference town b over the set of towns T	-	-
$\eta_m(b)$	$\frac{\Omega_m(b,T')}{\Omega_m(b,T)}$	-	proportion
$\epsilon_m(b)$	watershed enrichment of metric m, that is, $\frac{\eta_m(b)}{c(b)}$	-	proportion
$\Omega_{IBD}(b,a)$	total IBD between individuals in town a and town b	O	centiMorgans
$\Omega_{mig}(b,a)$	number of migrants from town a to town b (equivalent to $\delta_{a\to b}$)	O	individuals
$\Omega_{coal}(b,a)$	overlap of relative cross coalescence between individuals in town a and town b (equivalent to γ^{ba})	E	probability for time of first coalescence for all pairs of lineages