Title: On the Genes, Genealogies, and Geographies of Quebec

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

[1]Department of Human Genetics, McGill University; Montreal, Canada
[2]McGill University Genome Centre; Montreal, Canada
[3]School of Computer Science, McGill University; Montreal, Canada
[4]Big Data Institute, Li Ka Shing Centre for Health Information and Discovery; University of Oxford, Oxford, UK
[5]Sorbonne Paris Nord University, INSERM U1153, INRAE U1125, CNAM, Nutritional Epidemiology Research Team (EREN), Epidemiology and Statistics Research Center, University Paris Cité (CRESS); Bobigny, France
[6]Nantes Université, CNRS, INSERM, l'institut du thorax; Nantes, France
[7]BALSAC Project, Université du Québec à Chicoutimi; Chicoutimi, Canada
[*]Corresponding author. Email : simon.gravel@mcgill.ca

Abstract: Population genetic models only provide coarse representations of real-world ancestry. We use a pedigree compiled from four million parish records and genotype data from 2,276 French and 20,451 French Canadian (FC) individuals, to finely model and trace FC ancestry through space and time. The loss of ancestral French population structure and the appearance of spatial and regional structure highlights a wide range of population expansion models. Geographic features shaped migrations throughout, and we find enrichments for migration, genetic and genealogical relatedness patterns within river networks across Quebec regions. Finally, we provide a freely accessible simulated whole-genome sequence dataset with spatiotemporal metadata for 1,426,749 individuals reflecting intricate FC population structure. Such realistic populations-scale simulations provide new opportunities to investigate population genetics at an unprecedented resolution.

One-Sentence Summary: We present an accurate and high resolution spatiotemporal model of genetic variation in a founder population.

Main Text: The tapestry of human genetic history is formed of ancestral lineages interwoven by generations of coalescence and recombination events (*1*). It was woven across geographic landscapes (*2*) by individual dispersal and historical waves of migrations that can sometimes be reconstructed by genomic analyses (*3, 4*). Yet the complex relationship between spatial migrations and genetic variation still poses formidable challenges (*5, 6*).

As a general trend, the limits of dispersal lead to continuous isolation-by-distance and a sometimes striking correlation between genetic and geographic distances (*7–9*). However, in any region, specific historical events or geographic barriers are often used to explain discrete patterns of population variation (e.g., (*5, 10, 11*)). Reconciling continuous variation into discrete "evolutionarily significant units" has proven to be difficult and sometimes misleading (*12–14*). While many studies have considered anisotropic migration models (*15*)), and even detailed models of geographic constraint or 'resistance' (*16, 17*), comparing these models to genetic data is challenging.

This study takes advantage of a population-scale spatially labelled pedigree (spatial pedigree) compiled from over four million Catholic parish records in the province of Quebec together with genotype data for 20,451 individuals, and new pedigree-aware simulation tools to provide a detailed spatiotemporal model of genetic variation at scales ranging from tens to thousands of kilometres. By including French and British individuals in our analyses we assess how much ancestral population structure has been preserved from these two populations. We highlight the relationship between river networks and genetic similarity as the past four centuries of European colonial history has been marked by rapid frontier expansion beginning along the shores of the St. Lawrence River, and eventually expanding up its tributaries. By tracing the ancestry of millions of individuals over space and time we describe a constellation of distinct founder events arranged along geographic features that defined transportation and economic activity. In doing so, we further bridge the gaps between family pedigrees and continental population structure as well as gaps between theoretical models and empirical demographic histories.

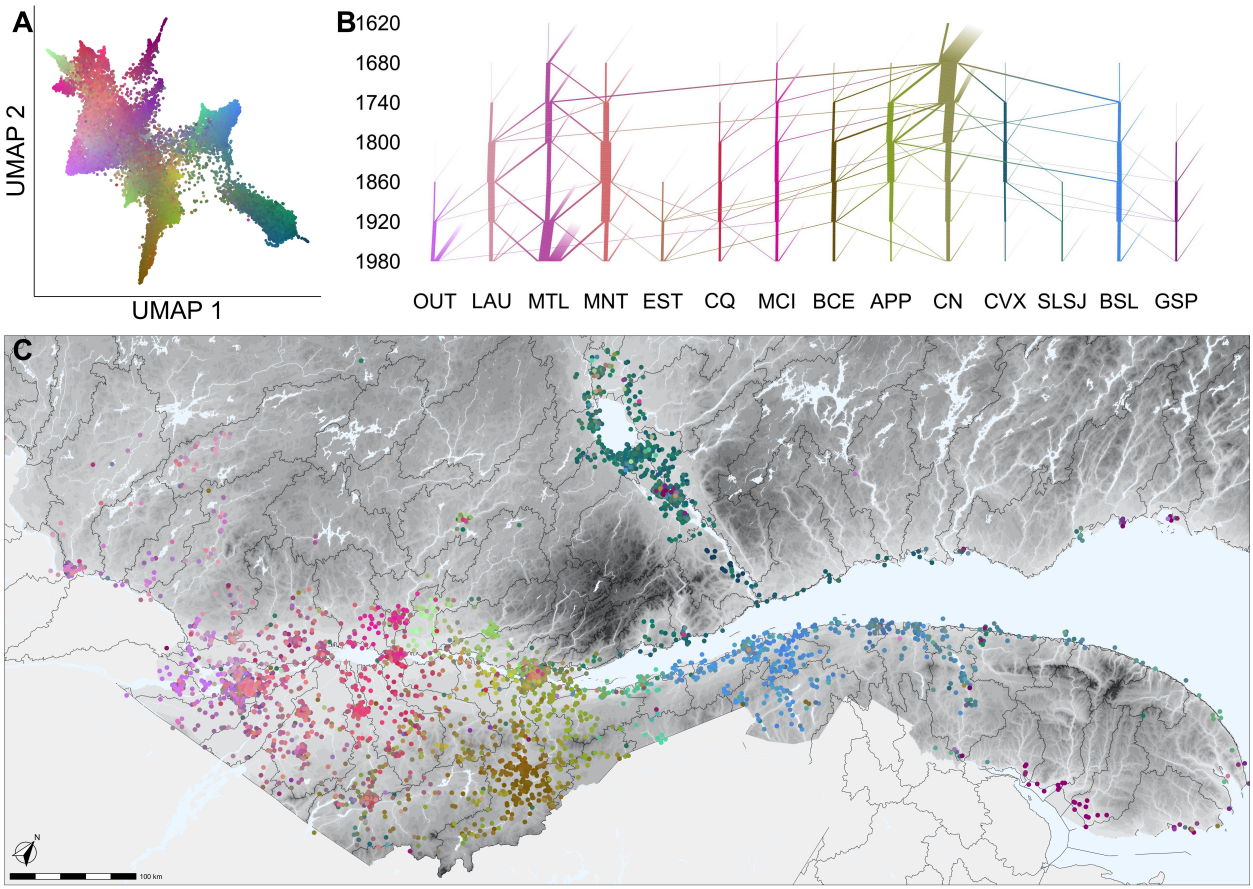


Figure 1: **French Canadian genes and genealogies mirror Quebec's geographies.** (**A**) UMAP for 20,451 individuals with inferred French Canadian ancestry. Each individual is assigned a colour based on their location in a dimensionally reduced genetic projection space (see Supplementary Methods 3.2). (**B**) Visualizing the genetic ancestry of French Canadians across regions (x axis) over time of marriage (y axis). The thickness of the line a time t from location A to location B represents the amount of genetic material ancestral to the present day population. (see Supplementary Methods 4.4 for details). (**C**) Clusters in genetic space coincide with distinct geographic regions of Quebec defined by geographic features like rivers and mountains. Watershed boundaries are shown in black. 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.

Results

Regional distribution of genetic variation

Quebec, a province in Canada, has a population of 8.6 million individuals, of which approximately 7.3 million speak French as a primary language. Because of strong historical correlation between Catholic religion, French ancestry, and French language, the genealogical records are much deeper and more complete for French Canadian (FC) individuals, by which we mean individuals who trace most of their ancestry to early French immigration irrespective of language. The majority of FC ancestry is derived from $\sim$ 8,500 settlers who migrated from France in the 17^{th} and 18^{th} centuries. The first 2,600 settlers contributed to two thirds of this gene pool (*18*). They occupied territory that had been inhabited and used by First Nations and Inuit peoples for thousands of years (*19*). Despite folk histories involving large amounts of indigenous ancestry among French Canadians (*20*), genetic and genealogical studies show that French Canadians born in Quebec carry on average less than 1% of ancestry tracing back to indigenous populations and the rest being mostly attributed to French ancestry (*21*).

Because pedigree data is much deeper and more complete among FC (*22*), we focused our genetic and genealogical analyses to 20,451 individuals inferred to be FC from the Cartagene (12,064 (*23*)) and Genizon cohorts (9,004, first reported here) (see Methods 3.1 for details on cohorts and ancestry inference). The genetic variation of this population was visualized using principal component analysis (PCA) (Fig. S1) and uniform manifold approximation and projection (UMAP) (Fig. 1A) (*24, 25*). A schematic summary of the entire FC spatial pedigree and the geographic location of 4,882 individuals linked to the spatial pedigree are shown in Fig. 1B and C. Visual inspection shows strong correlation of genetic and spatial proximity, and suggests that gradients in genetic variation coincide with geographical barriers and conduits like the St. Lawrence, Saguenay, and Chaudière Rivers or the Laurentian, and Appalachian Mountains.

French ancestry uprooted

Analyses of genealogical records show that a majority of French settlers came from regions in Northern and Northwestern France (Normandy, Ile-de-France, Aunis, Poitou, and Perche) (*26, 27*). However, successive waves of migration related to military and colonization objectives had different origins and demographics (*27*). To assess whether this history is reflected in the Quebec population structure, we compared the genomes of individuals living in different regions of Quebec and France. In agreement with historical records, FC share more recent ancestry with individuals from Northwestern France as measured by DNA that is identical by descent (IBD) (Fig. S2A). Furthermore, individuals living in the Outaouais and South Central regions of Quebec have lower rates of IBD with individuals from France (Fig. S2B). We used F_4-statistics to assess how much French structure has been preserved and find no evidence of specific regions in Quebec being more similar to specific French and British counterparts (Fig. S3).This indicates that most present-day structure among FC is independent of any ancestral structure or to differential contributions by French and British founders.

Simulated genomes with known transmission histories

To assess how much of the FC population structure can be accounted for by events following the arrival of French settlers, we generalized the `msprime` software (*28–30*) to perform genome-wide coalescent simulations conditioning on the known pedigree of the FC population. We defined the `FixedPedigree` ancestry simulation model in `msprime` version 1.2 to trace the ancestry of samples back through the genealogy accounting for coalescence and recombination (Fig. S4). To account for relatedness beyond the founders of the known pedigree, we modelled coalescence under a demographic model for European ancestry (*31*) (see Methods 2 for details).

We compared the simulations to ascertained data with a subset of 4,882 individuals who were both genotyped and linked to the pedigree. The PCA and UMAP of the simulations show clear qualitative agreement with the fine scale structure of ascertained genotype data (Fig. 2 and Fig. S3) as is evident from the first six PCs showing strong correlation (Fig. 2C). Thus, leading axes of

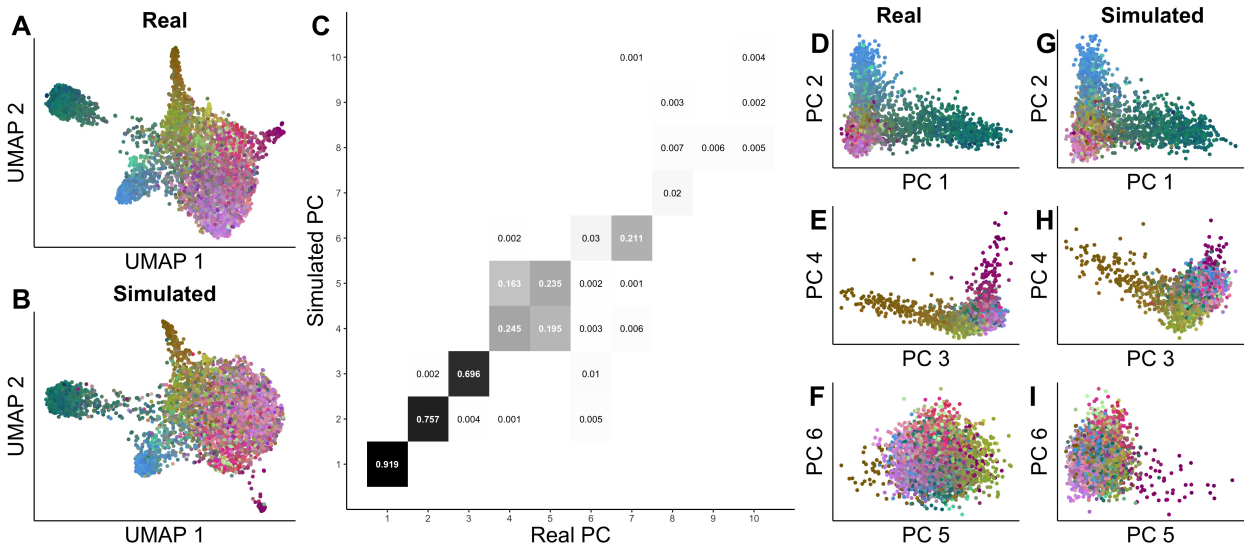


Figure 2: **Simulated genomes capture observed population structure.** Comparison of the same 4,882 individuals using simulated and observed genomes (coloured as in Fig. 1, see Methods 2 for details). (**A-B**) UMAP projections of observed and simulated genomes. (**C**) The correlation between observed and simulated principal components. (**D-F**) PCA projections of observed genomes. (**G-H**) PCA projections of simulated genomes.

genetic variation among FC reflect genetic drift that followed French settlement and is encoded in the spatial pedigree.

Following this, we simulated whole genomes of 1.4M present day individuals with at least four grandparents linked to the pedigree. (see Fig. S5 for UMAP and PCA). Tree sequences of these simulations are freely available on Zenodo (*link*). Although the tree sequences have been censored to remove personal identifying information, we have included temporal (decade) and spatial (latitude and longitude) information for the 1.4M samples and their ∼2M genealogically recoded genetic ancestors.

Gene flow within watershed boundaries

Figure 1 shows axes of genetic variation that are restricted geographically, in a way more reminiscent of differentiation across alpine valleys (*5*) than of the continuous population structure seen in

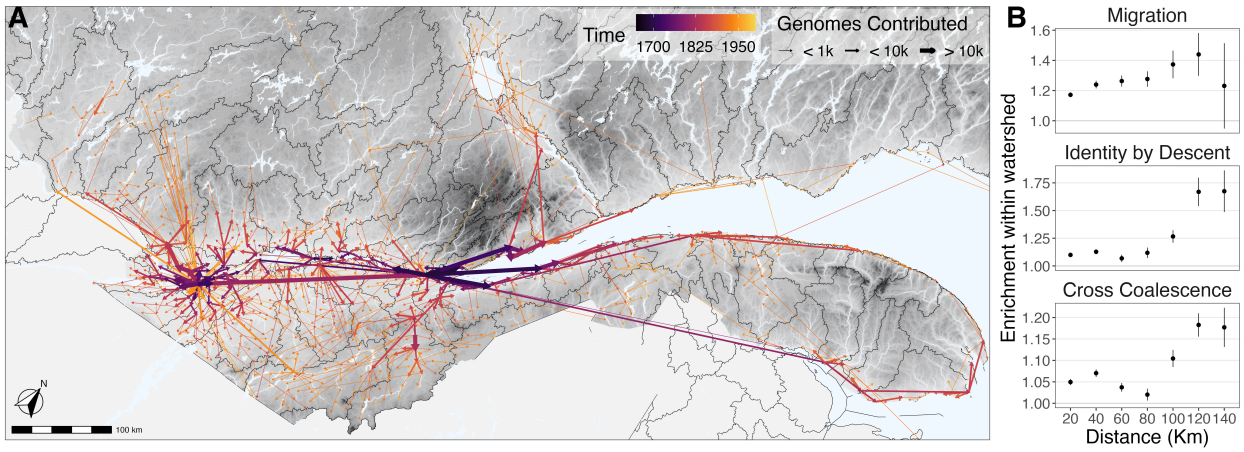


Figure 3: **Watersheds influence French Canadian migrations and relatedness.** (**A**) Primary routes of genetic ancestry. Segments link each town to the town from which migrants contributed the most genetic material. The width of segments indicates the total genetic contribution to present-day individuals and the colour indicates the mean date of when these contributions occurred historically. To avoid over plotting, we excluded the region of Abitibi-Témiscamingue and migrations of less than ten kilometres and migrations contributing less than ten genomes. (**B**) Relatedness is enriched within watershed. Black dots indicate the excess migrations and relatedness for towns within watershed relative to towns across watersheds at a fixed distance. See Methods 5 for details.

Europe (*9*). To assess how migration rates are impacted by topographical features like rivers and mountains, we visualize the major migration events shaping the FC population through space and time. Figure 3A exhibits waves of frontier expansion consisting of a series of sequential migrations up the tributaries of the St. Lawrence River ranging from tens to hundreds of kilometres in scale.

The first permanent French settlement took its name from the Algonquin word *kebec* making reference to the region where the St. Lawrence River becomes narrow (*32*). This was a strategic bridgehead location for the French as they sought to gain control of the main entrance to the Great Lakes in the 17th century [(*33*), p 49–52]. Facing vast forested territory used and occupied by Iroquoian and Woodland First Nations (*19*), the French formed a fluvial colony with thin ribbons of settlements along shorelines using a riverfront land division strategy [(*33*), p 56–57].

To study patterns of relatedness and dispersal between individuals distributed in 1,698 parishes

across the landscape, we considered three quantitative measures of relatedness and migration propensity between distinct geographic locations: migration rates, identity by descent, and cross-coalescence rates. All three show clear patterns of isolation by distance in all regions (Fig. S7). Given the importance of rivers in early settlement strategies, economic activity, and transportation (*34*), we hypothesized that relatedness and migration patterns broadly followed directions defined by local rivers and are therefore enriched for towns within watersheds. For each of the three metrics, we computed this enrichment for a tessellating set of 80 watersheds as a function of distance (Fig. 3B and Methods 5). We find an enrichment for the three metrics for all distances. At 120km, where the enrichment is strongest, we find a 40% enrichment in migration rates, 75% enrichment in genetic relatedness rates, and a 20% enrichment in cross coalescence rates. However, each region separately has a distinct migration pattern with varying degrees of topographical influence (Fig. 1C, Fig. S6).

Historical migrations in space and time

To follow the formation of regional substructure in the FC population, we defined three regions with large proportions of individuals driving the top principal components (Fig. S8). We ascended the spatial pedigree for pairs of individuals living in each region, computed the realized kinship for each common ancestor (Methods 4.2.1), and determined the location, timing and stringency of historical genetic bottlenecks (Fig. 4 A-C).

All regions exhibit substantial bottlenecks in and around Quebec city, but differ in their patterns of subsequent differentiation. The Saguenay-Lac-Saint-Jean (SLSJ) (Fig. 4 A,D) has a dominant early bottleneck in the region of Charlevoix, where an astrobleme (*35*) created a small pocket of fertile land within otherwise mountainous terrain (Fig. S10). Limited carrying capacity within the astrobleme led to demographic pressure and subsequent rapid expansion up the Saguenay River [(*33*), p 91], resulting in a vast majority of kinship predating the colonization of SLSJ. The Beauce region (Fig. 4 C,F) has a handful of bottlenecks in St-Joseph-De-Beauce and along the Chaudière River, with migrations reminiscent of a hub-and-spoke model. Finally, the Bas-Saint-Laurent (Fig.

4 B,E) has an assortment of bottlenecks including a dominant bottleneck in Rivière-Ouelle, but also more minor bottlenecks scattered across hundreds of kilometres of shoreline acting as a one dimensional regional hub for subsequent inland migrations (Fig. 4 E).

As expected in an expanding population, some early settlers (*super-founders*) had a large contribution to the population gene pool (*36, 37*). Here the top ten super-founders in each region contribute 32%, 11% and 12% of the realized kinship in SLSJ, Bas-Saint-Laurent and Beauce respectively (Fig. S11, S12, S13). And while each region has its deepest and most significant bottleneck near Quebec City, no two regions share the same super-founders (Fig. S14). To assess the overlap of each regional bottleneck, we computed cross-coalescence rates (*38*) (Methods 4.2.2 and Fig. S9) and find that 35% to 50% of kinship in Bas-Saint-Laurent and Beauce can be attributed to a shared bottleneck (Table 1), but the bottleneck in SLSJ is only 5% shared, reflecting much more intense founder events in Charlevoix and the region around Quebec City that are unique to SLSJ.

	Proportion of bottleneck shared with:		
Region	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 of founder events shared between regions, as measured by the ration of cross-coalescence rate divided by within-region coalescence rate.

Not all regions of Quebec exhibit such spatially defined bottlenecks. Even though Abitibi-Témiscamingue (AT) was settled by a similar process of rapid frontier expansion as seen in SLSJ, it did not lead to bottlenecks reflected in leading axes of genetic differentiation. In contrast to the events in SLSJ, the settlers to AT came from numerous villages scattered throughout the province. While many of the villages in AT have measurable bottlenecks, at a regional level these bottlenecks seldom overlap. To illustrate this, we consider the villages of Remigny and Rollet separated by twenty kilometres along the Ottawa River (Fig. S15). Cross-coalescence rates of these villages have 11% overlap. For comparison, La Baie and Roberval in SLSJ have 70% overlap despite being over one hundred kilometres apart. Even though towns in AT show limited evidence of a shared

founder effect, the parallel founding events create sub-structure beyond isolation-by-distance (*39*).

Discussion

Classical population genetic models often approximate reproduction and mate selection as a uniform random process. By approximating the effects of a myriad of individual motivations and choices that are unavailable to scientists, classical models provide an explanation for trends such as drift and selection. However, large population samples highlight the limitations of these models (*4, 40*). As we give up the simplified assumption of uniform random mating, the number of demographic parameters relevant to evolution grows rapidly.

The BALSAC pedigree – a particularly complete population scale spatial pedigree – has been instrumental in identifying multigenerational demographic effects like reproductive advantages of being on a wave-front expansion (*36*) or the transmissibility of family size (*41*) and migration propensity (*42*). Others have used it to study variation in runs of homozygosity (*43*) and its historical determinants, such as the kinship of the first settlers of Charlevoix (*34*) or the delay in the settlement of the Saguenay due to restrictions related to the fur trade [(*33*), p 91].

Here we sought to develop a comprehensive genetic model that captured all these effects, and more. We used this model to highlight how one of the best-studied human founder populations in SLSJ was influenced by the unique geography of the region that was shaped by a cosmic event occurring 400Mya in Charlevoix. We also found concordant genetic and genealogical support for the idea that geographic features like rivers and mountains played a systematic role in influencing migration rates and defining major axes of genetic variation. Finally, the strong correlation between empirical and simulated genetic data provides evidence that the structure within the FC population can largely be attributed to events in North America, while the population kept a genetic signature of the regions in France contributing more early French settlers.

We describe how geological, social, historical effects, as well as idiosyncratic events, translated in genetic variation patterns at various geographic scales and over centuries. Although our simu-

lations are based on a real pedigree, they do not contain identifying information and we can freely share this genome-wide dataset along with spatiotemporal metadata for over 1.4M individuals. Of course, these simulations are far from perfect. They do not account for natural selection beyond what is captured in the pedigree (*36*). The pedigree itself contains some amount of recording errors. Mismatches between genealogical ancestry and biological ancestry are not uncommon, and these will be difficult to overcome (*22, 40*). However, we believe that the genetic model, the simulation tools, and the publicly available simulated data we described provide a lens to investigate population genetics at an unprecedented resolution.

Acknowledgements

We acknowledge the contributions of Ivan Krukov who we consider eligible for authorship, but were unable to contact for approval. We are grateful to all of the participants who enabled this study by contributing their DNA, and to the participants who provided family information enabling the reconstruction of their genealogy. For data from Quebec, we thank the BALSAC team for their management and curation of the genealogy database, the CARTaGENE team and the Genome Quebec for their management and curation of genotype data. For data from France, we thank the EREN team for their management and curation of SU.VI.MAX genotype data. We thank Wilder Wohns, Yan Wong, and Gil McVean for enlightening conversations at early stages of the project, and Aaron Ragsdale for his comments on early drafts of this manuscript.

Funding

This work was supported in part by the Canadian Queen Elizabeth II Diamond Jubilee Scholarship (QES) (LAT) and the Fonds de recherche du Québec – Nature et technologies (FRQNT: B2X 290358) (LAT); also supported by NSERC discovery grant (SG), CIHR operating grant (SG), Canada Research Chair program (SG).

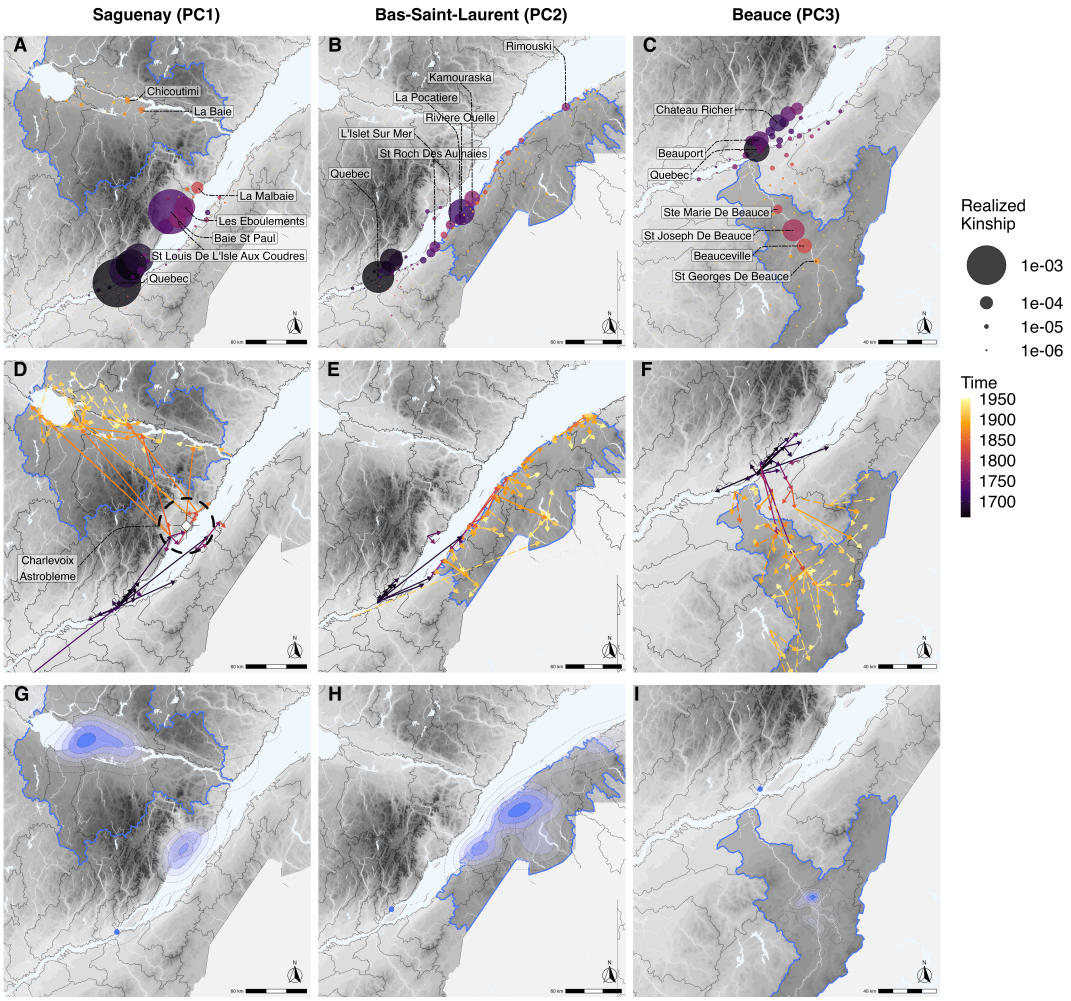


Figure 4: **Historical migrations define population structure** (**A-C**) Time, location and stringency of genetic bottlenecks in three regions of Quebec defined by the top three principal components (Fig. S8). Size of points is proportional to realized kinship to present day FC population.Point color represents mean date of the realized kinship based on marriage dates of ancestors. (**D-F**) Major axes of migration as measured by the estimated genetic contributions to present day individuals living within highlighted regions. Dotted arrows indicate towns having the largest genetic contribution to towns within the highlighted regions, and solid arrows highlight the fifth percentile of migratory routes ranked by estimated genetic contribution. Arrow colours represent the mean time of the genetic contribution based on the marriage dates of ancestors. (**G-I**) The dispersal range for the single ancestor with the highest contribution to each of the three regions. For a given ancestor, we generate a heat map using the location of towns weighed by the average contribution of that ancestor to the probands of each town. The marriage location of each of the major contributors is indicated by a blue dot.

Author contributions

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

Data curation: LAT, BALSAC, CARTaGENE, Genome Quebec, EREN, MT

Formal analysis: LAT

Funding acquisition: LAT, JK, SG

Investigation: LAT, SG

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

Project administration: LAT, JK, SG

Resources: SG

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

Supervision: JK, HV, SG

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

Visualization: LAT

Writing – original draft: LAT

Writing – review & editing: LAT, AR, CD, ADP, SZ, JK, HV, SG

Data and materials availability

R code for analyses and visualizations are available (`https://github.com/LukeAndersonTrocme/genes_in_space/tree/main/supplementary_code`). Tree sequences of simulated genomes are freely available on Zenodo (*link*). Quebec genotype data are available upon request to (`www.cartagene.qc.ca`) and (`https://www.mcgillgenomecentre.ca/`) for CARTaGENE and Genizon cohorts respectively. Genealogical data are available on the Scholars Portal Dataverse platform from the University of Québec in Chicoutimi (`https://doi.org/10.5683/SP3/BW7DIG`.)

References

1. Wohns, A. W. *et al.* A unified genealogy of modern and ancient genomes. *Science* **375**, eabi8264 (2022).

2. Cavalli-Sforza, L. L. *et al.* Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. *Proceedings of the National Academy of Sciences* **85**, 6002–6006 (1988).

3. Henn, B. M. *et al.* The great human expansion. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 17758–64 (2012).

4. Bradburd, G. S. & Ralph, P. L. Spatial population genetics: It's about time. *Annual Review of Ecology, Evolution, and Systematics* **50**, 427–449 (2019).

5. Cavalli-Sforza, L. L. "Genetic Drift" in an Italian population. *Scientific American* **221**, 30–37 (1969).

6. Battey, C. *et al.* Space is the place: Effects of continuous spatial structure on analysis of population genetic data. *Genetics* **215**, 193–214 (2020).

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

8. Ramachandran, S. *et al.* Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15942–7 (2005).

9. Novembre, J. *et al.* Genes mirror geography within Europe. *Nature* **456**, 98–101 (2008).

10. Bycroft, C. *et al.* Patterns of genetic differentiation and the footprints of historical migrations in the Iberian peninsula. *Nature Communications* **10**, 1–14 (2019).

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

12. Rosenberg, N. A. *et al.* Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genetics* **1**, e70 (2005).

13. Bradburd, G. S., Coop, G. M. & Ralph, P. L. Inferring continuous and discrete population genetic structure across space. *Genetics* **210**, 33–52 (2018).

14. Moritz, C. Defining 'evolutionarily significant units' for conservation. *Trends in Ecology & Evolution* **9**, 373–375 (1994).

15. Jay, F. *et al.* Anisotropic isolation by distance: the main orientations of human genetic differentiation. *Molecular Biology and Evolution* **30**, 513–525 (2013).

16. Petkova, D. *et al.* Visualizing spatial population structure with estimated effective migration surfaces. *Nature Genetics* **48**, 94 (2016).

17. McRae, B. H. Isolation by resistance. *Evolution* **60**, 1551–1561 (2006).

18. Charbonneau, H. *et al. 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. Elsey, C. *et al. Eastern Métis: Chronicling and Reclaiming a Denied Past* (Lexington Books, 2021).

21. Moreau, C. *et al.* Native American Admixture in the Quebec Founder Population. *PLoS ONE* **8**, e65507 (2013).

22. Vézina, H. *et al.* An overview of the BALSAC population database. Past developments, current state and future prospects. *Historical Life Course Studies* (2020).

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

24. McInnes, L. & Healy, J. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

25. Diaz-Papkovich, A. *et al.* Umap reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. *PLoS Genetics* **15**, e1008432 (2019).

26. Stanislas-Alfred, L. Bulletin du parler français au Canada (1903).

27. Vézina, H. *et al.* Origines et contributions génétiques des fondatrices et des fondateurs de la population québécoise. *Cahiers Québécois de Démographie* **34**, 235–258 (2005).

28. Kelleher, J. *et al.* Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Computational Biology* **12** (2016).

29. Nelson, D. *et al.* Accounting for long-range correlations in genome-wide simulations of large cohorts. *PLoS Genetics* **16**, e1008619 (2020).

30. Baumdicker, F. *et al.* Efficient ancestry and mutation simulation with msprime 1.0. *Genetics* **220** (2021).

31. Tennessen, J. A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–69 (2012).

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

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

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

35. Rondot, J. Nouvel impact météoritique fossile? La structure semi-circulaire de Charlevoix. *Canadian Journal of Earth Sciences* **5**, 1305–1317 (1968).

36. Moreau, C. *et al.* Deep human genealogies reveal a selective advantage to be on an expanding wave front. *Science* **334**, 1148–1150 (2011).

37. Labuda, D. *et al.* The effective family size of immigrant founders predicts their long-term demographic outcome: From Québec settlers to their 20th-century descendants. *PloS One* **17** (2022).

38. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nature Genetics* **46**, 919–925 (2014).

39. Scerri, E. M. *et al.* Beyond multiregional and simple out-of-Africa models of human evolution. *Nature Ecology & Evolution* **3**, 1370–1372 (2019).

40. Nelson, D. *et al.* Inferring transmission histories of rare alleles in population-scale genealogies. *The American Journal of Human Genetics* **103**, 893–906 (2018).

41. Gagnon, A. & Heyer, E. Intergenerational correlation of effective family size in early Quebec (Canada). *American Journal of Human Biology: The Official Journal of the Human Biology Association* **13**, 645–659 (2001).

42. Gagnon, A. *et al.* Transmission of migration propensity increases genetic divergence between populations. *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists* **129**, 630–636 (2006).

43. Burkett, K. M. *et al.* Correspondence between genomic-and genealogical/coalescent-based inference of homozygosity by descent in large French-Canadian genealogies. *Frontiers in Genetics* **12**, 1–11 (2022).

44. 1000 Genomes Project Consortium and others. A global reference for human genetic variation. *Nature* **526**, 68 (2015).

45. Université du Québec à Chicoutimi. Genealogy of Quebec Population 1621-1993 (2022). URL `https://doi.org/10.5683/SP3/BW7DIG`.

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

47. Statistics Canada. *Table 17-10-0005-01 Population estimates on July 1st, by age and sex* (Government of Canada, 2021).

48. Statistics Canada. *Table 15-10-0003-01 Population by mother tongue and geography, 1951 to 2016* (Government of Canada, 2019).

49. Tang, L. Standardizing population genetics simulations. *Nature Methods* **17**, 876–876 (2020).

50. International HapMap Consortium *et al.* A second generation human haplotype map of over 3.1 million snps. *Nature* **449**, 851 (2007).

51. Ragsdale, A. P. *et al.* Genomic inference using diffusion models and the allele frequency spectrum. *Current Opinion in Genetics & Development* **53**, 140–147 (2018).

52. Abraham, G. *et al.* Flashpca2: principal component analysis of biobank-scale genotype datasets. *Bioinformatics* (2017).

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

54. Menozzi, P. *et al.* Synthetic maps of human gene frequencies in Europeans. *Science* **201**, 786–792 (1978).

55. Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).

56. Petr, M. *et al.* admixr' r package for reproducible analyses using admixtools. *Bioinformatics* **35**, 3194–3195 (2019).

57. Delaneau, O. *et al.* Accurate, scalable and integrative haplotype estimation. *Nature Communications* **10**, 1–10 (2019).

58. Zhou, Y. *et al.* A fast and simple method for detecting identity-by-descent segments in large-scale data. *The American Journal of Human Genetics* **106**, 426–437 (2020).

59. Hudson, R. R. Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology* **23**, 183–201 (1983).

Supporting Materials for

"On the Genes, Genealogies, and Geographies of Quebec"

Luke Anderson-Trocmé, Dominic Nelson, Shadi Zabad, Alex Diaz-Papkovich, Nikolas Baya, Mathilde Touvier, Ben Jeffery, Christian Dina, Hélène Vézina, Jerome Kelleher, Simon Gravel

Contents

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 S17). The regional distribution of the samples included in the study is summarized in Figure 1. The Genizon cohort 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 (`www.cartagene.qc.ca`) 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 (*44*).

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 (*45*).

1.3 Geographical data

Layers of geographical data including rivers, lakes, watersheds, provincial, and federal boundaries were downloaded from the government of Canada geobase `https://open.canada.ca/`.

Polygons from each layer were simplified using ms_simplify function from the rmapshaper R library (*46*) and projected onto the EPSG:4326 coordinate system. 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 2020 Census (*47*) and estimates of the number of individuals who speak French as a primary language from (*48*).

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) and 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`) (*49*).

2.2 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 two

population out-of-Africa model of Tennessen et al. (*31*). The chromosome length and recombination rate for each simulated chromosome was defined by the GRCh37 hapmapII genetic map (*50*). 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 3.62×10^{-8} per basepair per meiosis. This unusually high mutation rate is chosen to match the mutation rate $\mu_{cds} = 2.35 \times 10^{-8}$ used in (*31*) to infer the demographic model, while accounting for the difference in diversity between coding and non-coding sequences. Because the mutation rate varies along the genome, we scaled this mutation rate to match genome-wide expectations using the relative rate of intergenic to coding polymorphism, i.e., $\mu_{int}/\mu_{cds} = 1.53$ as described in (*51*). A summary of model specifications are in table S3.

A GitHub repository with code to run the genome simulation pipeline is available (`https://github.com/LukeAndersonTrocme/genome_simulations`) and extensive documentation of Msprime (`https://tskit.dev/msprime/docs/latest/api.html#msprime.FixedPedigree`).

2.3 Comparing simulations to ascertained data

To compare the simulated genomes to ascertained genotype data, we downsampled the simulations 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. We restricted our simulations to 4,882 individuals from a total of 5,402 individuals who consented to be linked to the FC pedigree. We excluded 100 individuals who did not have all four grandparents 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 (*52*) 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 FC ancestry (*47, 48*), as such we expect the majority of the participants in our cohorts also derive FC ancestry despite only a fraction of them being linked to the genealogy. For the purposes of visualizing the population structure of FC in Figure 1A and C, we sought to leverage the large number of FC 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 FC ancestry, and excluded 617 individuals from our analyses (see Supplementary Figure S18 and Supplementary Table S1).

3.2 Dimension reduction

Flashpca2 was used to performed our principal component analysis (PCA) (*52*) to generate (Supplementary Figure S1) and the R package uwot (*53*) for our uniform manifold and approximation projection (UMAP) (*24*) 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 (*54*).

3.3 F statistics

F statistics used in Supplementary Figure S2A were computed using the admixture R package (*55, 56*). 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

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 (*57*) and downsampled to a set of common SNPs before computing pairwise IBD using Hap-IBD for all samples (*58*) 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 (i.e., the length of inherited genetic material in centiMorgans) of an individual i to a set P of probands,

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

where $\mathrm{K}^p(i)$ 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

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

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 ancestors 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:

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

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

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:

$$\lambda^{P^A P^B}(i) \approx \frac{1}{4} \sum_{(m,n)} \mathrm{K}^{P^A}(m)\mathrm{K}^{P^B}(n) \tag{S5}$$

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 (*38*):

$$\gamma^{P^A P^B} = \frac{2\Lambda^{P^A P^B}}{\Lambda^{P^A} + \Lambda^{P^B}} \tag{S6}$$

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 desendents 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

$$d(M_{a\to b}) = \sum_{i\in I_{a\to b}} w_i d_i \tag{S7}$$

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 :

$$\delta_{a\to b} = \frac{|M_{a\to b}|}{N_a} \tag{S8}$$

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.

$$\delta_{A\to b} = \sum_{a\in A} \delta_{a\to b}. \tag{S9}$$

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 S16. 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)

$$T' = T \cap S. \tag{S10}$$

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

$$c(b) = \frac{|T'|}{|T|}, \tag{S11}$$

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

$$\omega_m(b, a) \tag{S12}$$

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

$$\Omega_m(b, T') = \sum_{a \in T'} \omega_m(b, a), \tag{S13}$$

and

$$\Omega_m(b, T) = \sum_{a \in T} \omega_m(b, a), \tag{S14}$$

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

$$\eta_m(b) = \frac{\Omega_m(b, T')}{\Omega_m(b, T)}. \tag{S15}$$

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

$$\epsilon_m(b) = \frac{\eta_m(b)}{c(b)}. \tag{S16}$$

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:

$$\omega_{IBD}(b,a) \equiv g(a,b) = \frac{1}{|P^a||P^b|} \sum_{i \in P^a, j \in P^b} IBD(i,j) \tag{S17}$$

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 :

$$\omega_{mig}(b,a) \equiv \delta_{a \to b} = \frac{|M_{a \to b}|}{N_a} \tag{S18}$$

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 :

$$\omega_{coal}(b,a) \equiv \gamma^{ba} = \frac{2\Lambda^{ba}}{\Lambda^b + \Lambda^a} \tag{S19}$$

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.

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

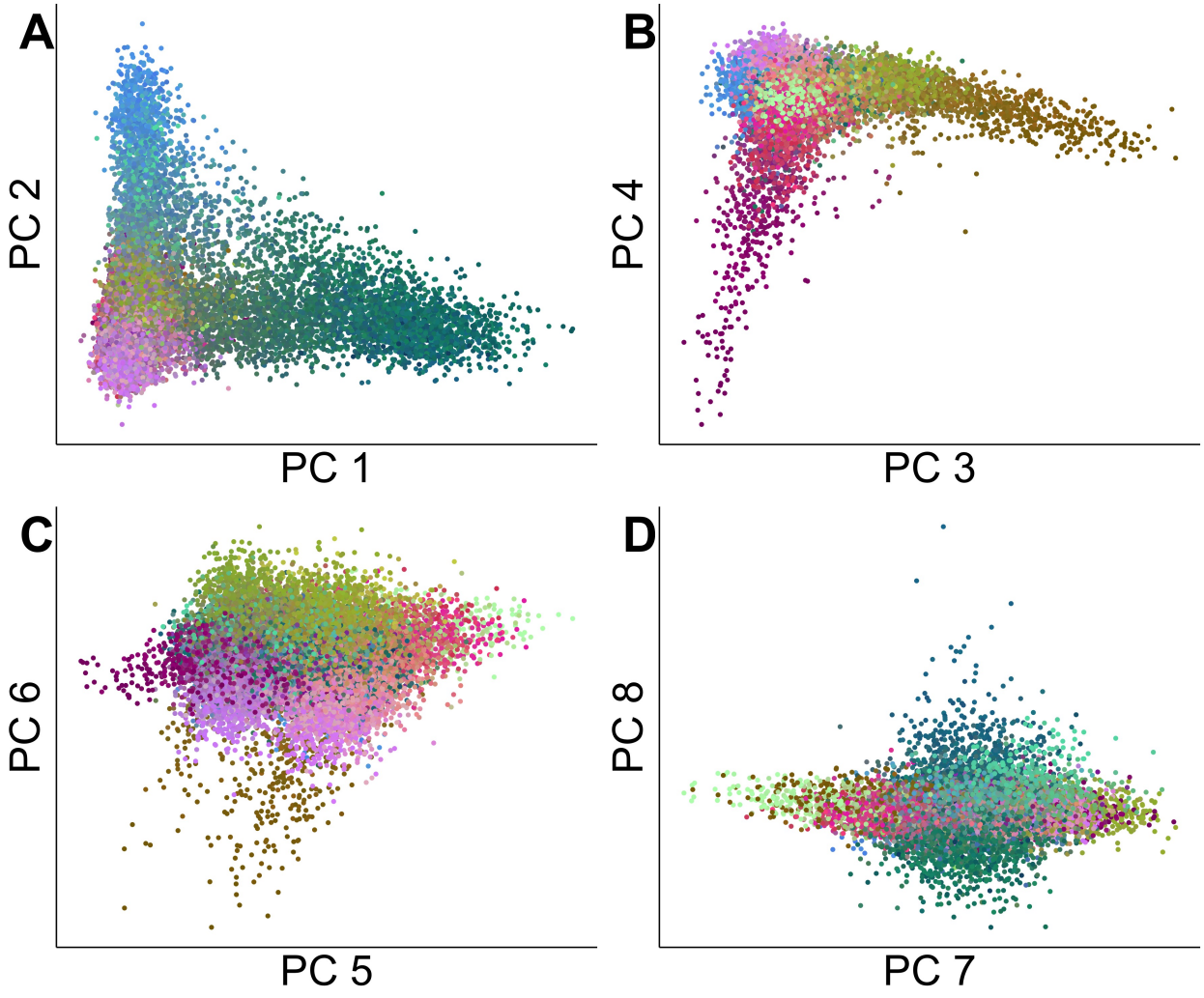
Supporting tables and figures

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

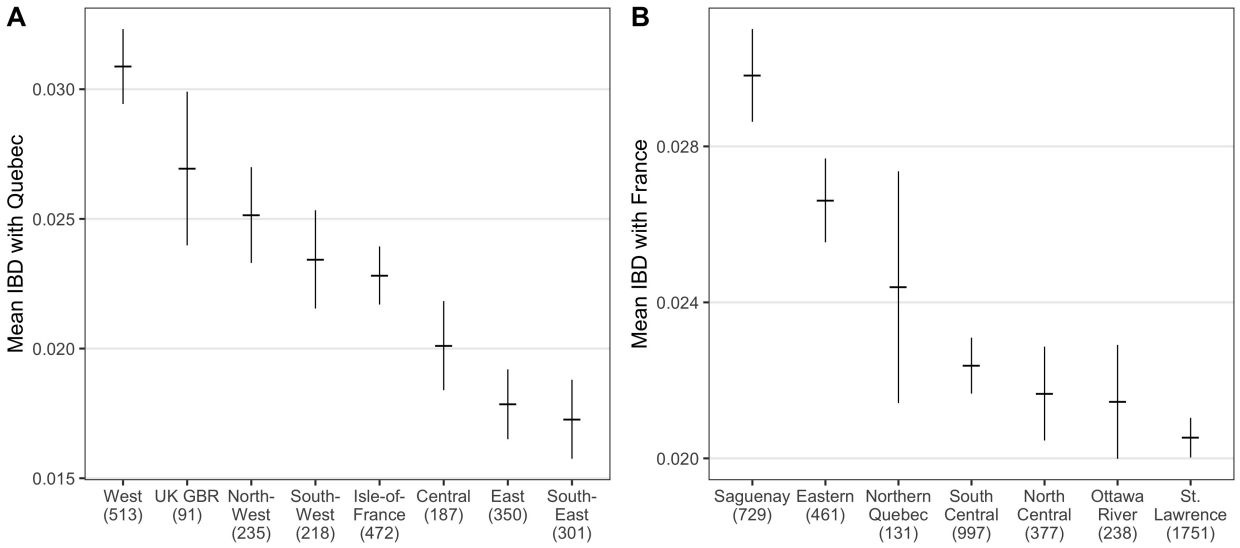
Table S1: Sample sizes of genotyped cohorts and their ancestries. A total of 20,451 individuals genealogically linked or genetically inferred French-Canadian (FC) ancestry are used for visualizing population structure using PCA and UMAP.

South-East
1, 4, 5, 6, 7, 13, 26, 38, 42, 69, 73, 74, 83, 84
South-West
9, 11, 12, 24, 30, 31, 32, 33, 34, 40, 46, 47, 48, 64, 65, 66, 81, 82
West
16, 17, 22, 29, 35, 44, 49, 53, 56, 72, 79, 85, 86
North-West
2, 14, 27, 50, 59, 60, 61, 62, 76, 80
Central
3, 15, 18, 19, 23, 28, 36, 37, 41, 43, 45, 63, 87
Isle-of-France
75, 77, 78, 91, 92, 93, 94, 95
East
8, 10, 21, 25, 39, 51, 52, 54, 55, 57, 58, 67, 68, 70, 71, 88, 89, 90

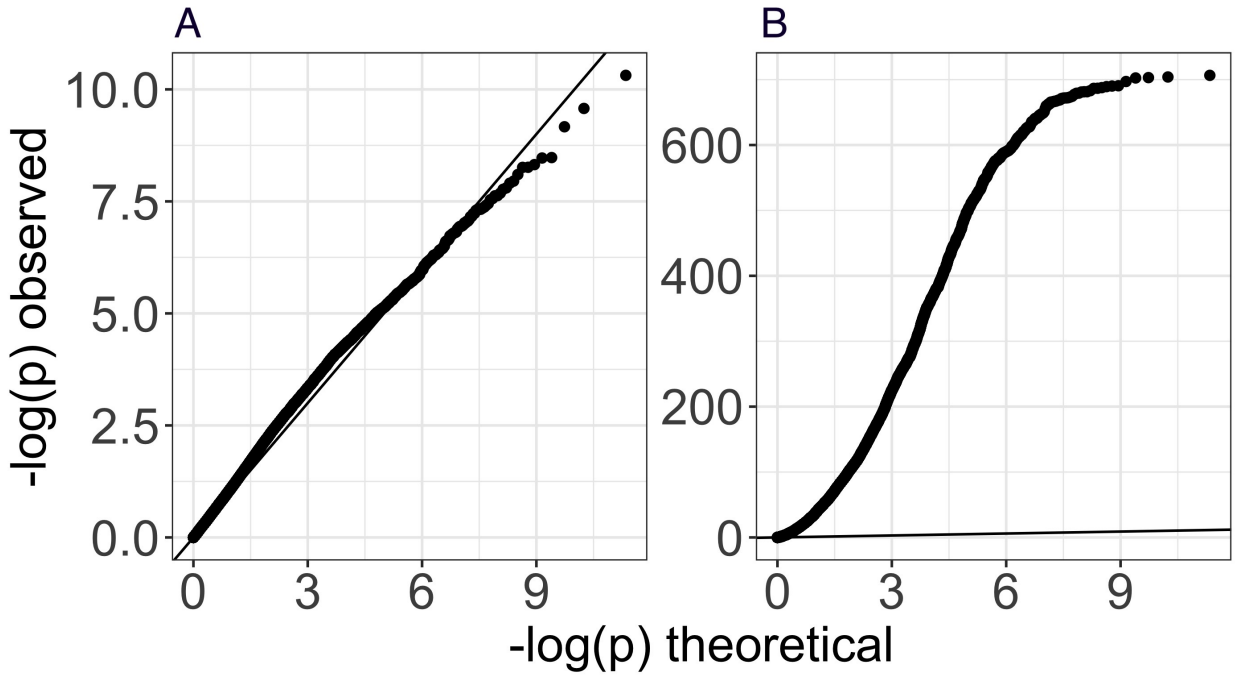
Table S2: French administrative departments within each of the seven regions.



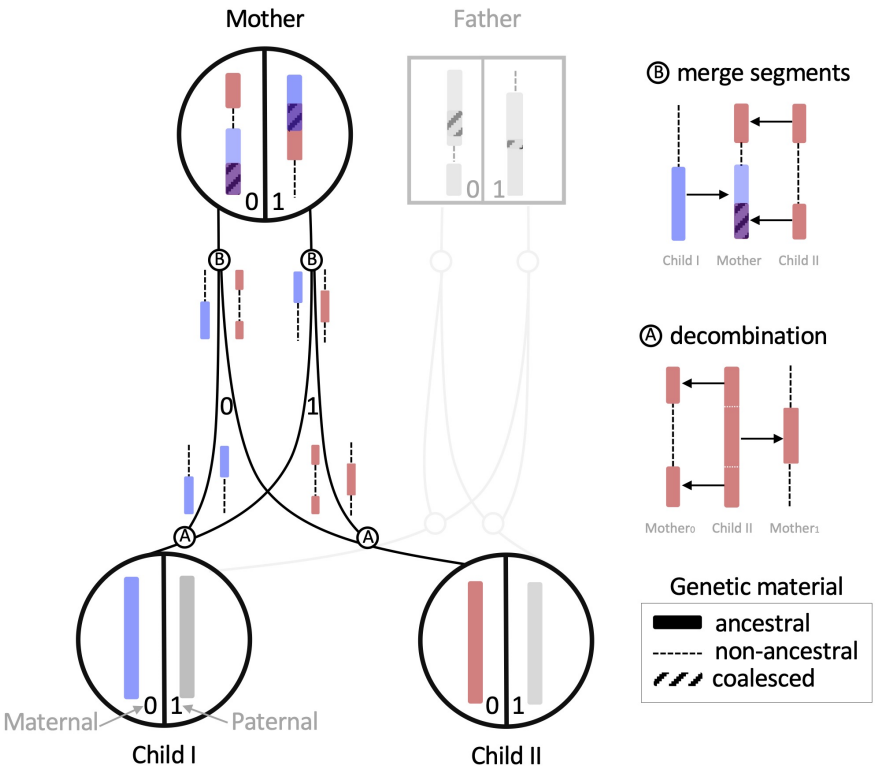
Supplementary Figure S1: **Principal Component Analysis of French Canadians** (**A-D**) The top eight principal components of the genotype data included in the analyses.



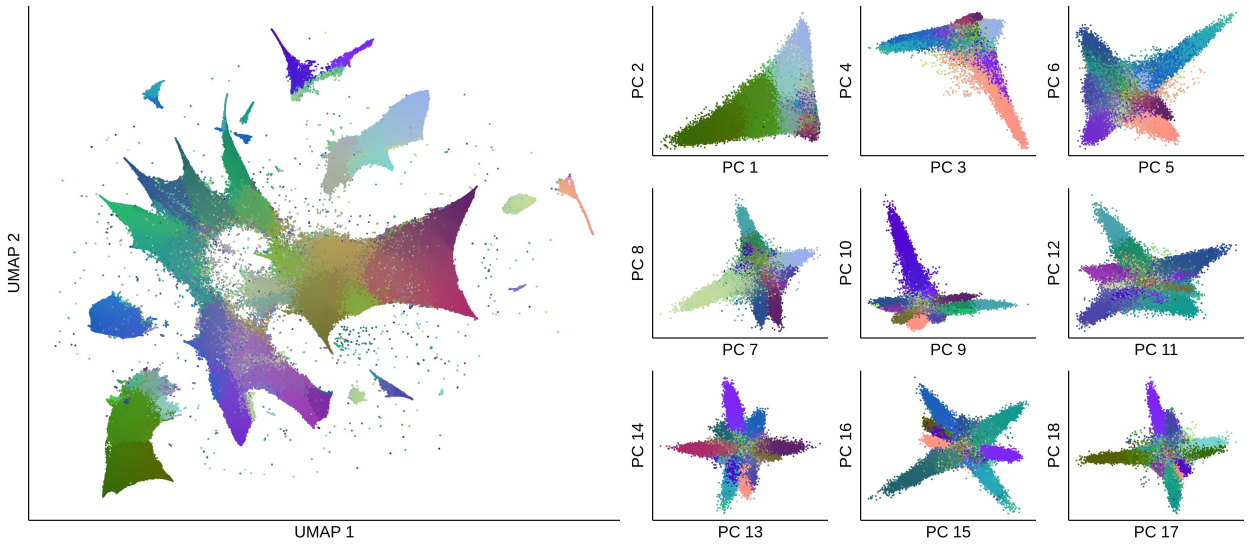
Supplementary Figure S2: **Rates of identity by descent in different regions.** The average IBD sharing between individuals living in Quebec, France, and England. The error bars in these plots indicate the standard error of the mean length (cM) of IBD sharing across individuals from the region indicated 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 Western and Northern France than with individuals from Central and Southern France. We note that in addition to having a small sample size and large error bars, the GBR cohort is less comparable to French cohorts because their genomic data are generated using different platforms. (**B**) Individuals from France have higher rates of IBD with individuals from Saguenay and Eastern Quebec than with individuals from the Ottawa River and Central Quebec.



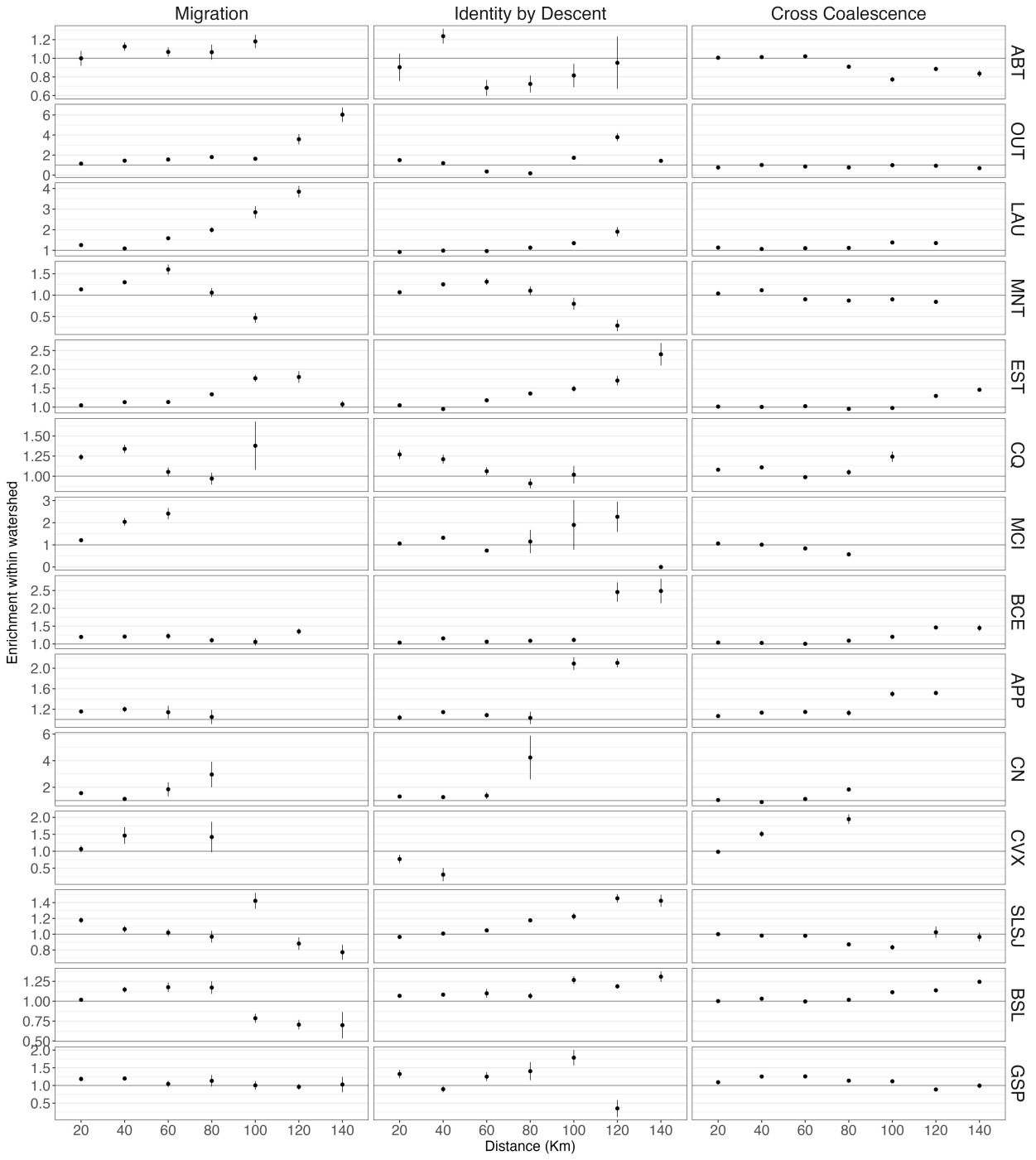
Supplementary Figure S3: **Comparison of Quebec, French and British regions using F_4-statistics (A)** All 441 combinations of regions in Quebec and Europe (France and Britain) were compared using an F_4-statistic: F_4(Qc1, Qc2, Eu1, Eu2). The QQ plot shows no enrichment of significant *p*-values, consistent 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.



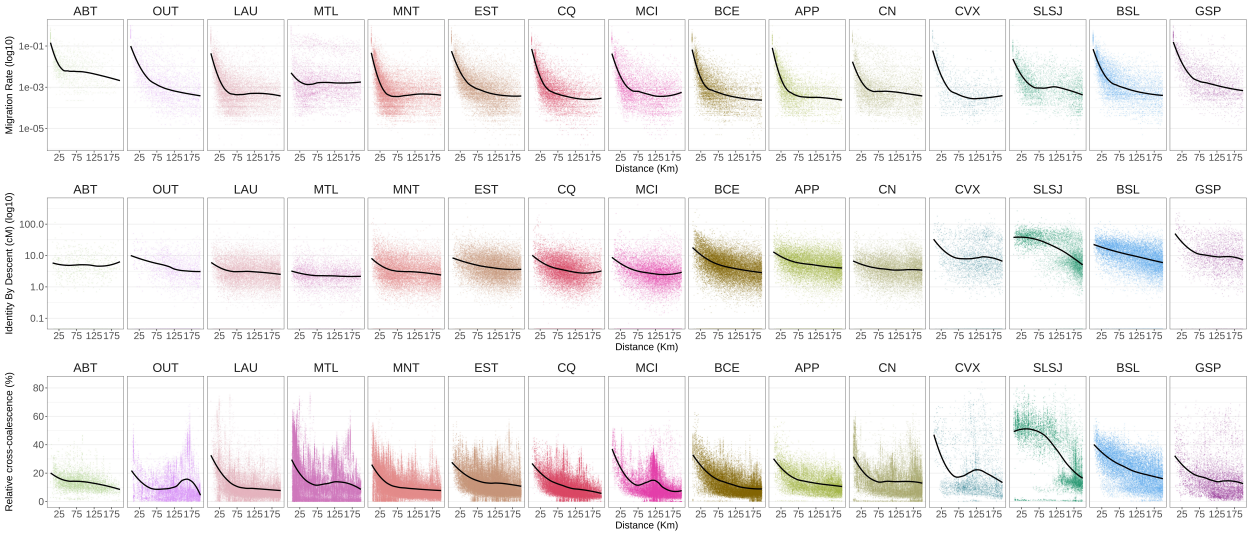
Supplementary Figure S4: **Illustrating decombination** (**A**) Ancestry simulations in backwards time *decombine* genetic material from children based on a user specified recombination map. (**B**) Decombined segments are then merged in a common ancestor. Not all of the genetic material contributed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



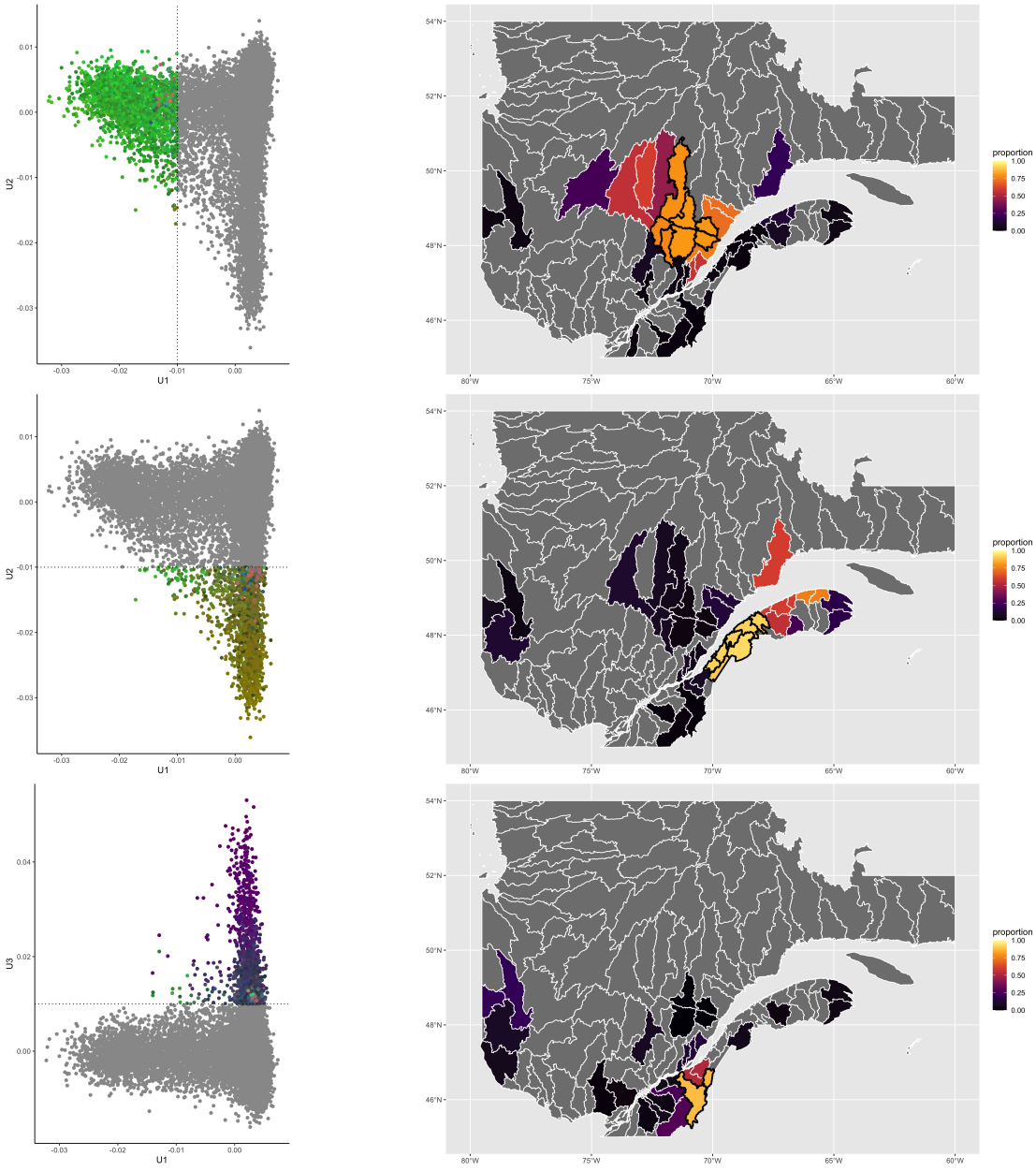
Supplementary Figure S5: **Visualizing 1.4 Million simulated French Canadian genomes** (**left**) A UMAP of the simulated genotype data. (**right**) The top eighteen principal components of the simulated genotype data. Colours were generated from a three dimensional UMAP through converting each x, y, z coordinate into an RGB value unique to each individual (see Supplementary Methods 3.2).



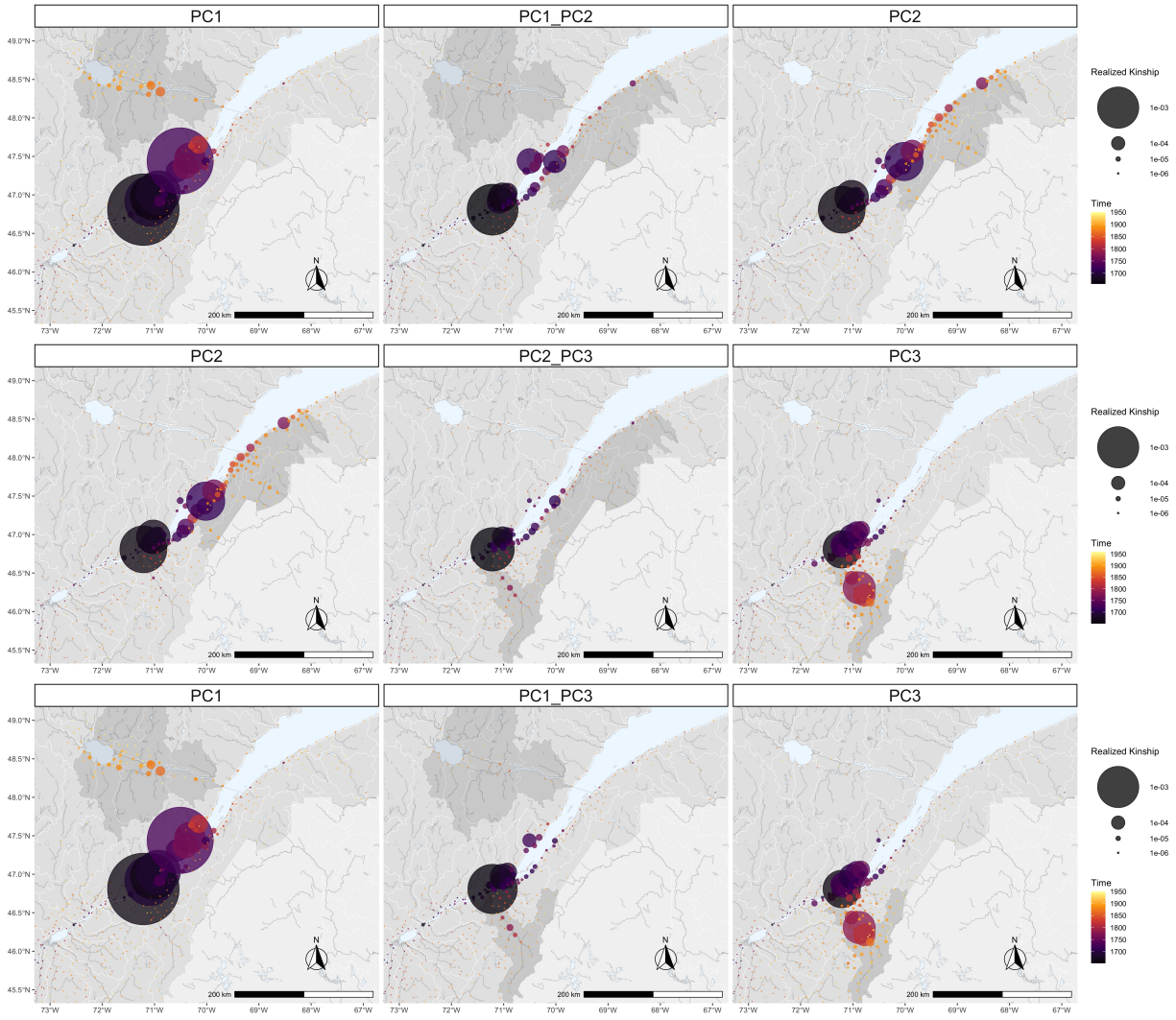
Supplementary Figure S6: **Historical migrations that define regional population structure** Watersheds influence migrations, genetic and genealogical relatedness. Rows of panels indicate a separate region's watershed enrichment comparing the migration rates IBD rates and cross coalescence rates between 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.



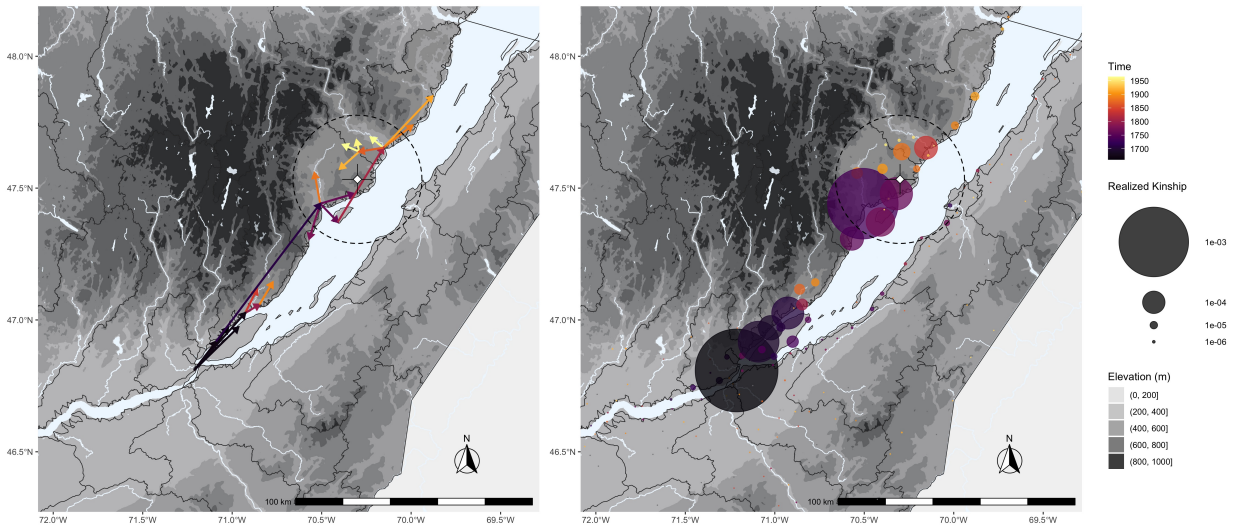
Supplementary Figure S7: **Migration rates, Identity by descent rates, and relative cross coalescence rates decay with distance for each region in Quebec.** The y axis is $\omega_m(b, a)$ for all metrics m for all reference towns b and distal towns a. The x axis is the distance in kilometres between 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 S16. The solid black line indicates a loess fit line using local polynomial regression fitting for each panel separately.



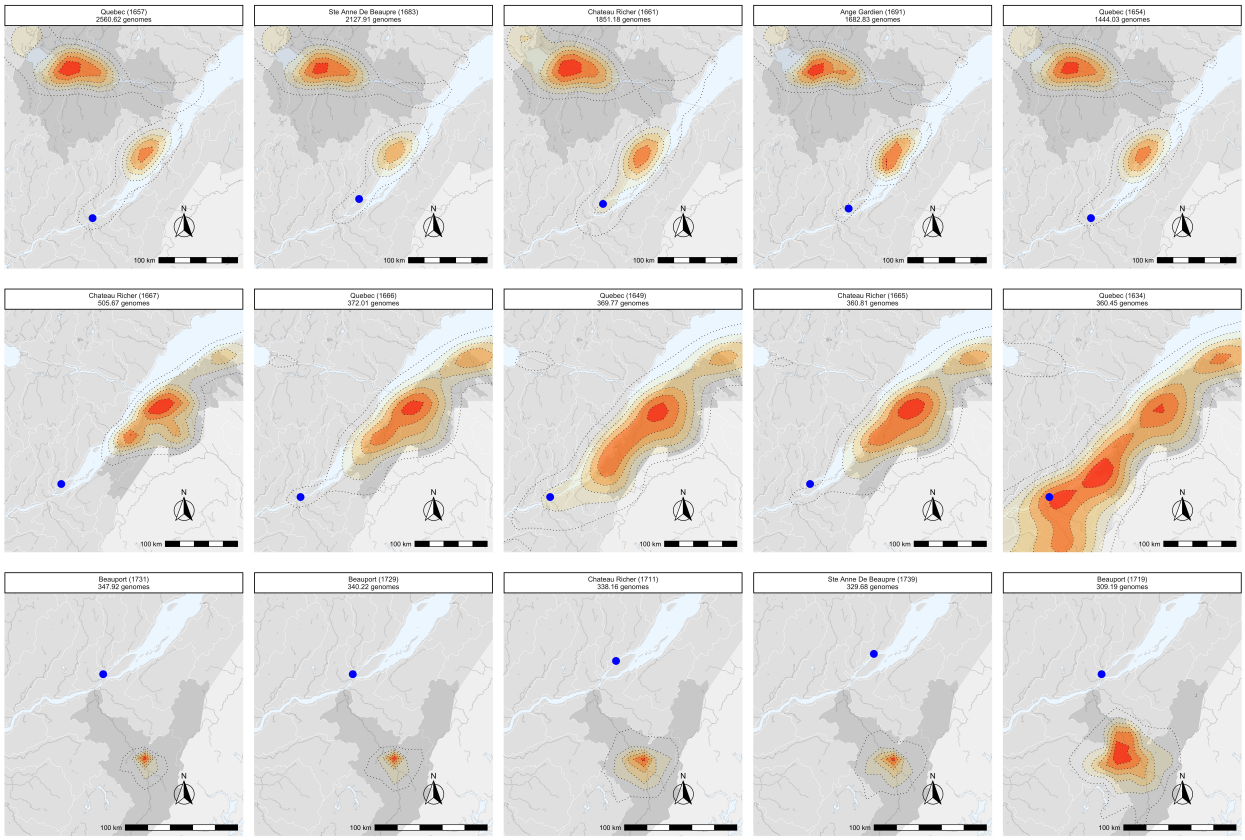
Supplementary Figure S8: **Selecting watersheds with individuals driving principal components** We define a threshold of 0.01 along each of the top three principal components and compute the fraction of individuals in each watershed that are beyond this limit. We choose watersheds with more than 75% of individuals beyond the threshold (black contours).



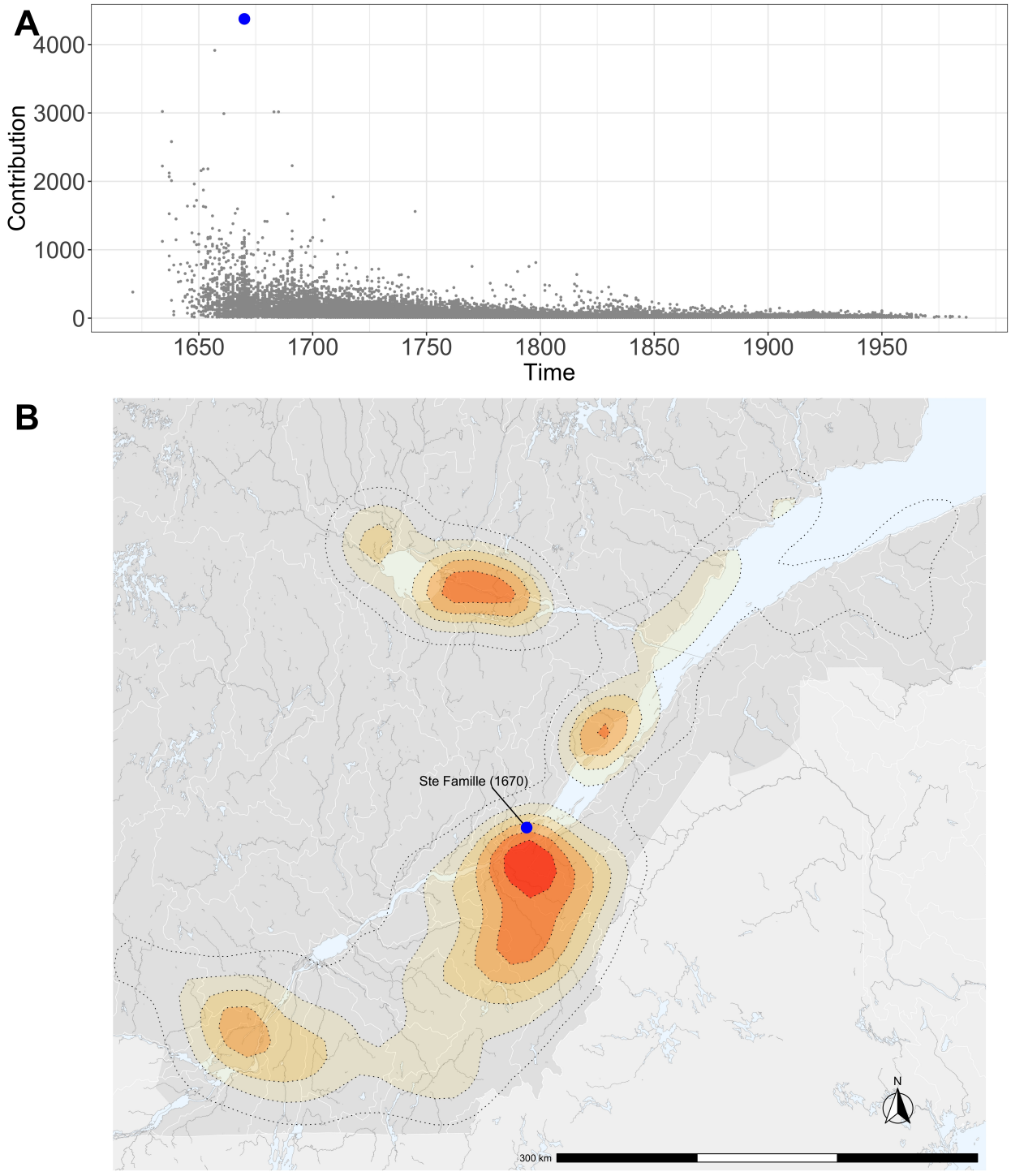
Supplementary Figure S9: **Cross coalescence rates** For each pair of regions defined in Figure S8, we compute the cross coalescence rates by ascending the genealogies of individuals in both regions. The central panels represent the between region cross coalescence rate, whereas the first and last column 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.



Supplementary Figure S10: **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.



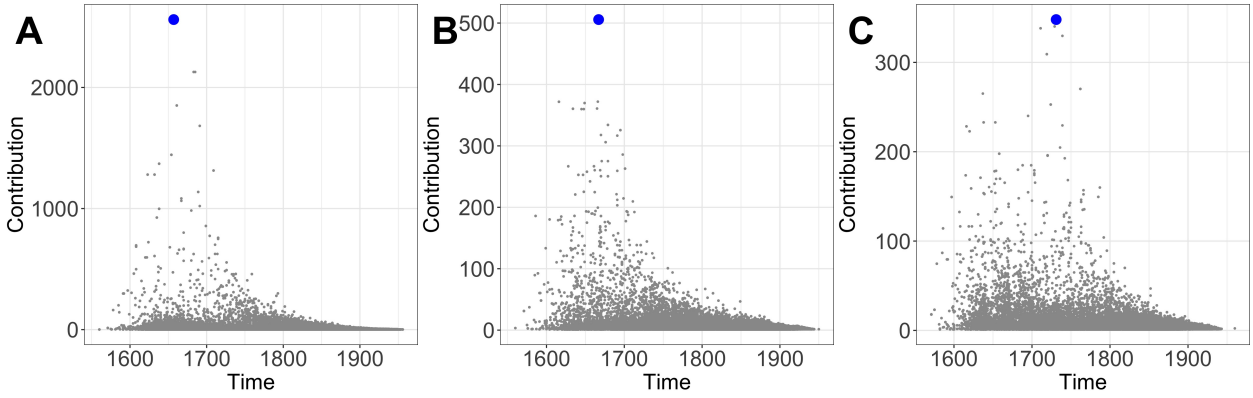
Supplementary Figure S11: **Dispersal range of the top five contributors to each region** For each region **(rows)** enriched in individuals driving each of the top three principal components, we show the dispersal range of the top five super-founders **(columns)**. Even though there is some variation between the dispersal ranges of the super-founders of a region, they broadly cover the same geographic regions. The scaling of the heat maps are not comparable between plots as they were computed separately.



Supplementary Figure S12: **Range dispersal for the top contributing ancestor A** The total genetic contribution of all ancestors in the pedigree. 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.

Parameter	Specification	Citation
demographic model	European ancestry in two population out-of-Africa model	(*31*)
coalescent model	Hudson	(*59*)
mutation rate beyond known pedigree	3.6210^{-8}	(*51*)
recombination map	GRCh37 hapmapII genetic map	(*50*)

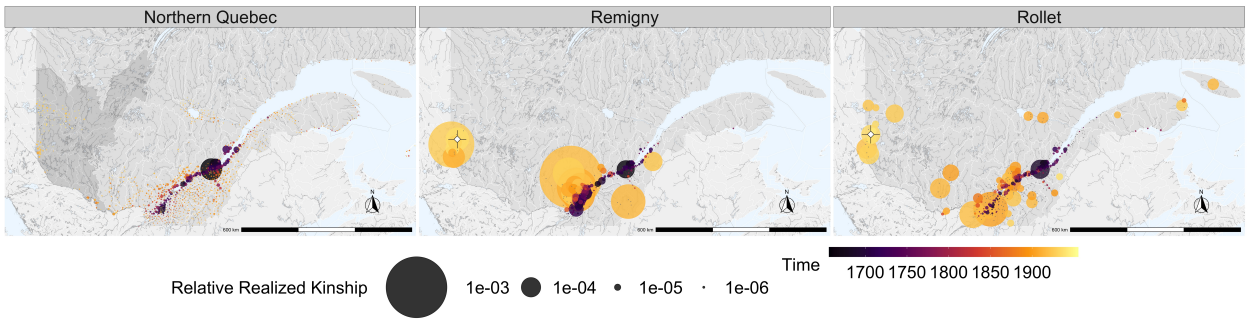
Table S3: The model parameters used to simulate the ancestry of individuals in the French-Canadian pedigree.



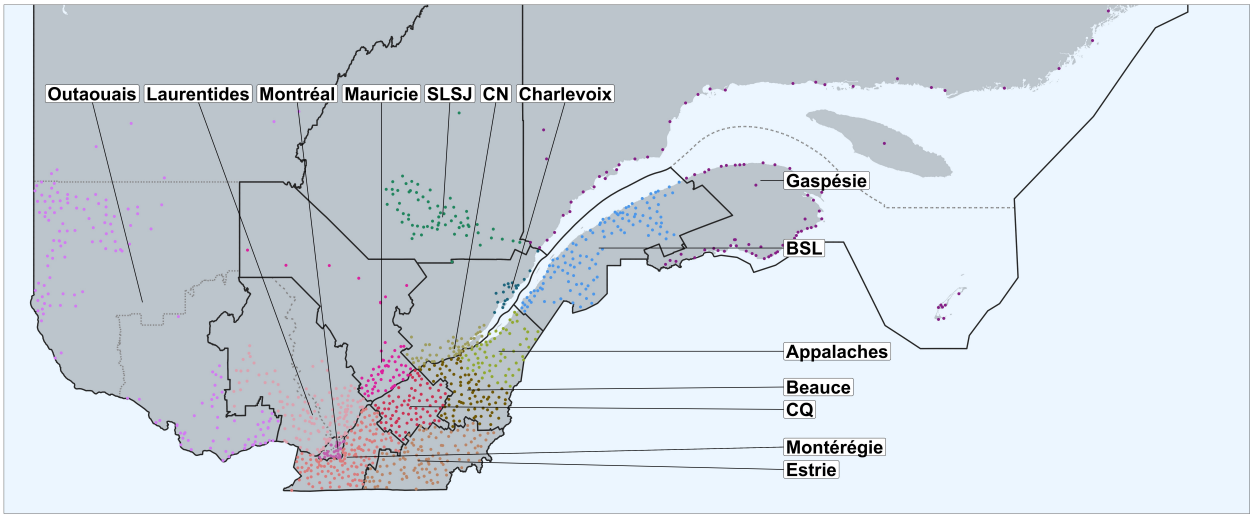
Supplementary Figure S13: **Genetic contribution of ancestors** The genetic contribution of each ancestor to probands living in each of the three regions enriched in the top three principal components. The x axis places each ancestor in time based on their marriage date. The ancestor contributing the most genetic material to each region is indicated by a blue dot.

PC1	PC2	PC3
2561	122	87
2128	93	60
1851	83	64
1683	91	15
1444	74	16
1314	79	14
1137	51	48
984	50	21
124	506	12
134	372	16
146	370	37
119	361	15
201	360	171
119	360	19
118	334	23
66	326	6
94	317	21
8	4	348
8	4	340
9	4	338
9	4	330
9	4	309
6	3	270
520	87	265
7	3	253

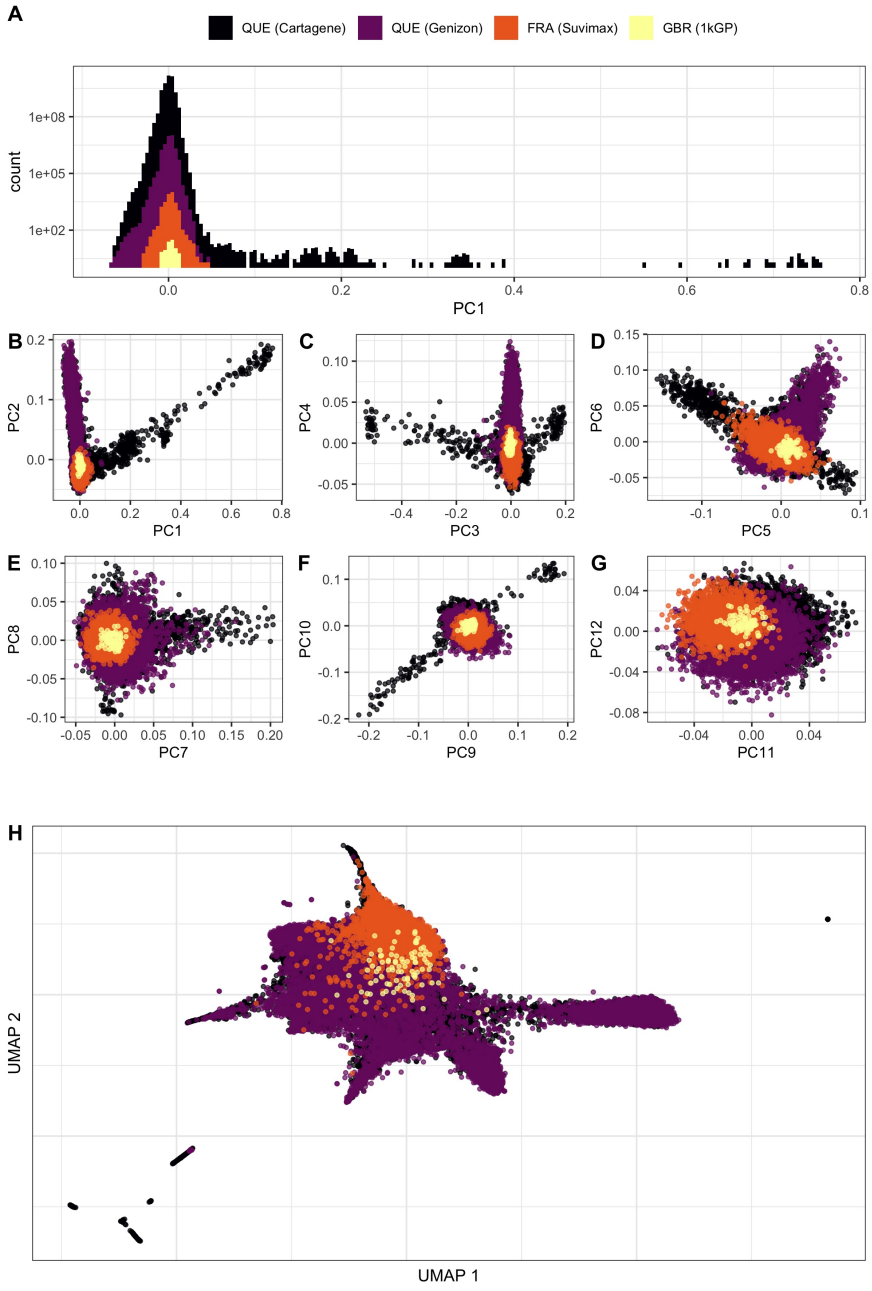
Supplementary Figure S14: **Contributions of regional super-founders** For each region enriched in individuals driving each of the top three principal components, we identify the top ten 'super-founders' based on their realized kinship to each region. We also compute the contributions of these ancestors 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.



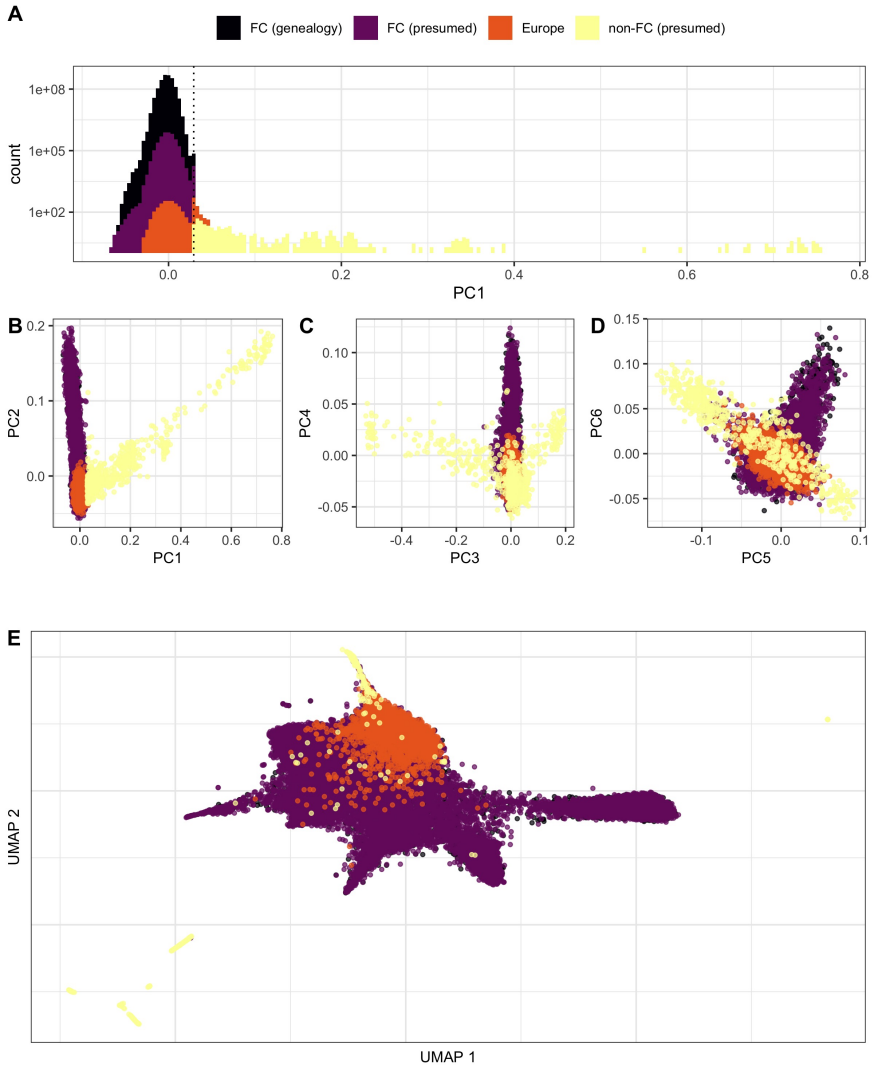
Supplementary Figure S15: **Realized kinship for Remigny and Rollet** Northern Quebec (left panel) shows no major regional founding events other than in Quebec City in the 16th century. When we consider each town separately, bottlenecks become more apparent. The towns of Remigny (centre panel) and Rollet (right panel) have recent founding events from different regions in Quebec despite being twenty kilometres away from each other.



Supplementary Figure S16: **Administrative boundaries used to generate genealogy flow plot** The genealogy flow plot in Figure 1 was generated based on Quebec administrative boundaries. 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.



Supplementary Figure S17: **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.



Supplementary Figure S18: **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 FC ancestry, presumed FC ancestry using a threshold based on the maximum value along the first principal component of genealogically linked individuals, and non-FC 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.