nature portfolio

Peer Review File

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

Corresponding Author: Professor Simon Girard

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

A version of this paper was originally rejected for publication by Nature Communications, however that decision was reconsidered after appeal by the authors.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Morin et al, analyze the fine scale structure of the Saguenay–Lac-Saint-Jean (SLSJ) region of Quebec, known for its high frequency of some rare genetic disease. Although the founder events and demographic processes that lead to the increase of frequency of these disorders have been extensively studied, the fine scale genetic structure is less known. This is important if subregion have a different genetic makeup for genetic counseling. Further it is now demonstrated that substructure can lead to biases estimation in Gwas studies.
Their study use the extensive demographic genealogical database Balsac that covers the whole region. More precisely, they study the genealogies of 80 348 probands married between 1921 and 1960. And they use the genetic database the CARTaGENE cohort that includes whole genome genotyping of 29,337 individuals, among them 9405 matched in the Balsac genealogical database.

Their study show a clear genetic substructure. This is clearly important since this population has always been described as quite homogenous.

(Remarks on code availability)
They first demonstrate a high correlation for these 9405 individuals between genetic and genealogical kinship. This result enables them to use the whole demographic database to infer genetic substructure in the region. In order to achieve these results, they develop a new method to compute kinship on their big genealogical dataset.
Their results show a clear fine scale structure.
The paper is well written, the methodology is impressive.

My remarks are minors:
1. some maps would be usefull.
- a map of the East of Quebec province in the main text would be usefull.
- maps that present the substructure differences in term of founder genetic contribution
- map that present the kinship level.
2. It could be useful to address a little bit more the importance of substructure for GWAS studies

Reviewer #2

(Remarks to the Author)
Reviewer comment for
Fine-scale structure of a whole regional population through genetics and genealogies

Summary: The authors investigate the fine-scale population structure of a French Canadian population by computing two kinship metrics, one expected (from a pedigree) and the other realized. They convert the kinship matrix into a distance matrix, then apply non-linear dimension reduction and clustering algorithms to obtain smaller subpopulations from the data.

In particular, they developed a novel kinship coefficient computing algorithm that offers substantial improvement over existing methods. Although I think the novel algorithm is helpful, I'm not convinced by the clustering/dimension reduction results. Below are my comments.

Major comments
1. Can the authors try out PC initialization? Unlike other branches of genomics where non-linear dimension reduction (hereafter, DR) is widely used, literature in applying non-linear DR to DNA sequence data has been relatively sparse. A few examples I can remember are the All-of-Us paper and another Quebec paper that used a dataset that overlaps with the paper currently being reviewed. Therefore, I believe there is little guidance on how to properly apply non-linear DR or if it's even appropriate to apply these types of DRs. At this point, it would be informative to validate the wisdom from other fields in genomics, such as single-cell RNA sequencing. It is more common to apply PCA first and then feed the PC coordinates to non-linear DR as described in this paper (https://www.nature.com/articles/s41587-020-00809-z). Another paper working with DNA sequence data takes similar steps (https://www.cell.com/ajhg/pdfExtended/S0002-9297(22)00112-4), PCA first and then UMAP. It seems that initializing the non-linear DR with PCA makes a substantial difference. I would perhaps do the following two analyses:
a) Initialize with PCA, then feed PCA to the non-linear DR
b) Initialize with PCA, but feed the original distance matrix to non-linear DR. i.e., do the exact same analysis as in the draft but add PC initialization.
For pedigree PC, there are several options including the paper in which the authors cited (https://gsejournal.biomedcentral.com/articles/10.1186/s12711-025-00994-y).

2. How robust is the HDBSCAN/UMAP based classification? Visually, the points are very crowded and putting lines between them feels very artificial. One could try to adjust the parameters to check robustness, which the authors have already done to some extent, but we have a better alternative in this case. Since the migration patterns of Quebecois are already documented to a certain extent, it will be possible to run a population genetic simulation using msprime or SLiM (https://tskit.dev/) (or perhaps another suitable simulation scheme) . Then, authors could apply their pipeline to the simulated data to see if it recovers the demographic pattern codified into the simulation. The resulting tree sequence is amenable for kinship extraction (e.g. https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1011537) which should be amenable to the paper's pipeline.

3. What is the time scaling of the new algorithm? Since it has to materialize an N-by-N kinship matrix, it should be at least O(N^2). I wonder if this is indeed the case.

(Remarks on code availability)
Their software github repo is well-documented. I have not attempted to run the analysis of the paper.

Reviewer #3

(Remarks to the Author)
In this work Morine et al investigate the fine population structure of the Saguenay-Lac-Saint-Jean region of Quebec (Canada) integrating extensive genotype data and genealogical records. To do this they optimize an algorithm to efficiently compute kinship coefficients from genealogies. The study is thorough, methodologically sound, and well written; however, I have some reservations regarding its suitability for publication in Nature Communications. Indeed, while the study provides a detailed and quite convincing analysis of the Saguenay–Lac-Saint-Jean population (see below for specific comments on the presented analyses), the topic may be too specialized for the broad readership of Nature Communications. Moreover, the extent to which the proposed approach can be generalized remains unclear. The authors benefit from exceptionally rich datasets, both in terms of genotype information and genealogical records, which may not be available in other cases. Consequently, the broader applicability of their findings and methods might be limited. It is also worth noting that the authors do not present new data, but rather reanalyze previously generated genotype and genealogical datasets. While the methodological optimization is valuable, the lack of novel data somewhat limits the overall originality and impact of the study.

Despite my view that this manuscript may not be appropriate for publication in Nature Communications, I include below some remarks and suggestions concerning the analyses carried out, specifically on the assessment of genetic structure and dimensionality reduction methods.
The authors investigate the genetic structure of the population through IBD sharing (realized kinship), aiming to efficiently compare genetic and genealogical relationships. It would, however, be informative to also explore the genetic structure directly from genotype data (e.g. SNP-based analyses), for instance through principal component analysis (PCA) or estimates of FST between clusters. Such analyses could provide complementary insights into the fine-scale population structure shaped by watercourses and the east-west migration patterns described in the study. In addition, the authors employ UMAP projections to summarize both genetic and genealogical data. While UMAP is a powerful dimensionality reduction method, it can sometimes introduce distortions that exaggerate the apparent separation between clusters. I therefore recommend that the authors also perform a PCA and subsequent clustering on the PCA coordinates (e.g. using k-means clustering) to verify whether the genetic structure observed in the UMAP plots is robust.

Overall, this is a carefully executed study that makes good use of both genetic and genealogical data to characterize the fine-scale population structure of the SLSJ region. The methodological work is solid, and the manuscript is clearly written. However, the study mainly reanalyzes existing datasets and focuses on a highly specific regional case. As a result, the

broader generalizability and potential impact of the findings appear limited. While the methodological contribution and the integration of rich genealogical resources is appreciable, I am not convinced that the current version of the manuscript meets the breadth and impact expected for publication in Nature Communications. I would, however, encourage the authors to consider submission to a more specialized genetics or population genomics journal.

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #2

(Remarks to the Author)
The authors have made substantial efforts to revise their manuscript and I only have one major concern left. It's an overreaching claim to say that any magnitude of population structure can induce spurious associations in association studies. It is in fact the realized allele frequency stratification that creates excess false positives. For instance, Edge et al. is a good reference to this point (https://pubmed.ncbi.nlm.nih.gov/24481204/). Also, see Rosenberg et al. (https://academic.oup.com/genetics/article/173/3/1665/6060953).

If the authors want to keep up with this claim, they should run simulations to confirm their findings: since they have a pedigree, simulate genomes conditioned on the pedigree and run GWAS on them. I'll recommend using msprime (https://tskit.dev/msprime/docs/stable/pedigrees.html), which can easily generate genomes conditioned on the pedigree. Then, using the genotypes from the data, one can further simulate traits and conduct GWAS to check the claim.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
The authors have provided convincing arguments addressing the most delicate point raised in my previous report, namely the specificity of the results and their limited generalizability. Their response is detailed and, this time, persuasive. In light of this, considering the opinions of the other referees and the revisions introduced in the new version of the manuscript, I now believe that the work is appropriate for publication in Nature Communications.

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Version 2:

Reviewer comments:

Reviewer #2

(Remarks to the Author)

The new benchmark looks excellent. I do not have any concerns left. It's a great paper, and best to my knowledge, one of the few paper that clearly demonstrates the significance of regional population structure in human populations.

(Remarks on code availability)

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.
In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.
The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.
To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/

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

We appreciate that the editor sent our work for peer review and are pleased that the reviewers recognised its technical soundness and the value of integrating genealogical and genetic data with an optimised kinship algorithm. We respectfully request an appeal of the decision, as we believe we can address the reviewers' concerns regarding the broader applicability of our approach.

General comments:

In particular, Reviewer 3 emphasised that the method has only been demonstrated in a uniquely rich, region-specific case with exceptional depth of genealogical and genotype data, making it difficult to judge whether the approach generalises beyond Saguenay-Lac-Saint-Jean.

We have substantially revised the manuscript to present the work with broader generalisation and relevance for populations with less comprehensive data resources. We believe that access to such rich data offers a unique opportunity to understand the complexity of fine structure in human populations. We show in our work that we can detect fine structure within what was considered a 'homogeneous' and very recent founder population like SLSJ. It is reasonable to expect that similar recent structure exists in all populations, signals that are often obscured by coarse, ancient structure under the standard stratification correction methods used in most GWAS and polygenic risk score analyses. Relying solely on PCA of common SNPs has been shown to leave residual stratification, which can lead to serious biases and false associations. Our study now emphasises why addressing this issue is critical. In addition to the revised scope of the manuscript asked by reviewer #3, we made the following changes:

- Add maps of eastern Quebec and genetic contribution as requested by reviewer #1.
- Replace UMAPs of expected and realised kinship matrices by UMAPs initialised on MDS as requested by reviewer #2.
- Replace supplementary Fig. 3 to reflect the 10-dimensional UMAP used to identify the SLSJ cluster as raised by reviewer #2.

We have addressed these points and all others in our following detailed responses to the reviewers. We hope that the revised manuscript will now be considered of broad interest to the Nature Communications readership.

Additionally, Reviewers 2 and 3 highlighted the need for more robust validation of the clustering and dimensionality reduction strategy (e.g., PCA initialisation, simulation-based validation, and orthogonal assessments of structure from genotype-derived SNP analyses).

We have addressed these points in detail in our responses to the reviewers and implemented the corresponding changes in the manuscript. Please refer to the specific point-by-point answers for further clarification.

Reviewers' comments:

Reviewer #1 (Remarks to the Author):

Their study show a clear genetic substructure. This is clearly important since this population has always been described as quite homogenous.
Their results show a clear fine scale structure.
The paper is well written, the methodology is impressive.

My remarks are minors:
1. some maps would be usefull.
- a map of the East of Quebec province in the main text would be usefull.

We added this as Fig. 1 in the main text. During this process, we identified that St-Prime had been misclassified and corrected this across all figures and results.

- maps that present the substructure differences in term of founder genetic contribution

We also added this as Fig. 5C in the main text.

- map that present the kinship level.

Because kinship is inherently pairwise, we explored several visualisation approaches to represent it effectively. However, none of the configurations provided a clear or interpretable view without appearing cluttered and confusing. Therefore, we decided not to include this map in the revised manuscript.

2. It could be useful to address a little bit more the importance of substructure for GWAS studies

We have incorporated this into the main text, and the scope of the work is now significantly broader, as also requested by reviewer #3. We sincerely thank the reviewer for this valuable comment, which has greatly improved our manuscript.

Reviewer #2 (Remarks to the Author):

Major comments

1. Can the authors try out PC initialization? Unlike other branches of genomics where non-linear dimension reduction (hereafter, DR) is widely used, literature in applying non-linear DR to DNA sequence data has been relatively sparse. A few examples I can remember are the All-of-Us paper and another Quebec paper that used a dataset that overlaps with the paper currently being reviewed. Therefore, I believe there is little guidance on how to properly apply non-linear DR or if it's even appropriate to apply these types of DRs. At this point, it would be informative to validate the wisdom from other fields in genomics, such as single-cell RNA sequencing. It is more common to apply PCA first and then feed the PC coordinates to non-linear DR as described in this paper (https://www.nature.com/articles/s41587-020-00809-z). Another paper working with DNA sequence data takes similar steps (https://www.cell.com/ajhg/pdfExtended/S0002-9297(22)00112-4), PCA first and then UMAP. It seems that initializing the non-linear DR with PCA makes a substantial difference.

We appreciate the reviewer's suggestion and apologise for the lack of clarity in our methods. The reviewer is correct that initialisation plays an important role in UMAP. While our original analysis used spectral embedding (the default in the umap-learn package) for initialisation, we have now applied MDS initialisation as suggested. Because IBD represents pairwise similarity rather than raw genotype features, MDS provides a more appropriate initialisation than PCA, as it is applied to pairwise relationships directly. This approach is consistent with previous work that applied UMAP to MDS of IBD matrices[1]. UMAP figures were changed accordingly in the main text.

I would perhaps do the following two analyses:
a) Initialize with PCA (MDS), then feed PCA (MDS) to the non-linear DR:

This analysis was performed using varying numbers of MDS components (see Fig. 1 below). To fully capture the population structure represented by the expected kinship matrix, approximately 1,000 MDS components were required. The first components primarily reflect the Charlevoix gradient and the East of Ha! Ha! cluster, which explain most of the variance. Consequently, when too few components are used, fine-scale structure remains masked behind these dominant signals.

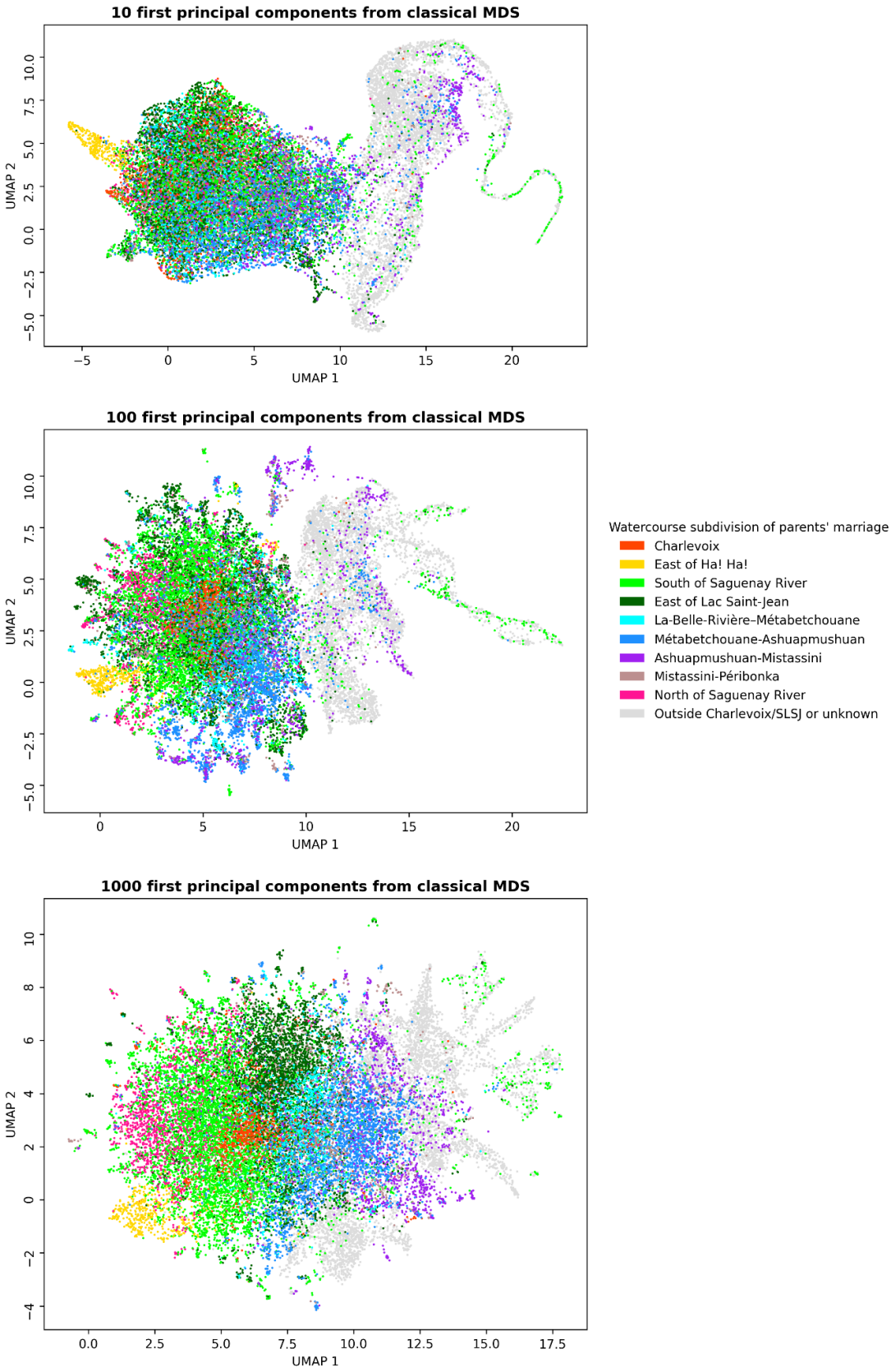


Fig. 1: UMAPs of varying numbers of dimensions of an MDS calculated on genealogical kinship of individuals with parents married in SLSJ in 1931–1960, initialised on MDS 1 and 2.

b) Initialize with PCA (MDS), but feed the original distance matrix to non-linear DR. i.e., do the exact same analysis as in the draft but add PC initialization.

The resulting UMAP is very similar to our previous version initialised with spectral embedding (see Fig. 2 below). Therefore, we included the new figure initialised with MDS in the manuscript, as suggested by Reviewer #2. Considering this and our focus on capturing recent fine-scale structure, which appears easier when using the kinship matrix with UMAP initialised by the first two MDS dimensions, we decided to retain this approach in our manuscript.

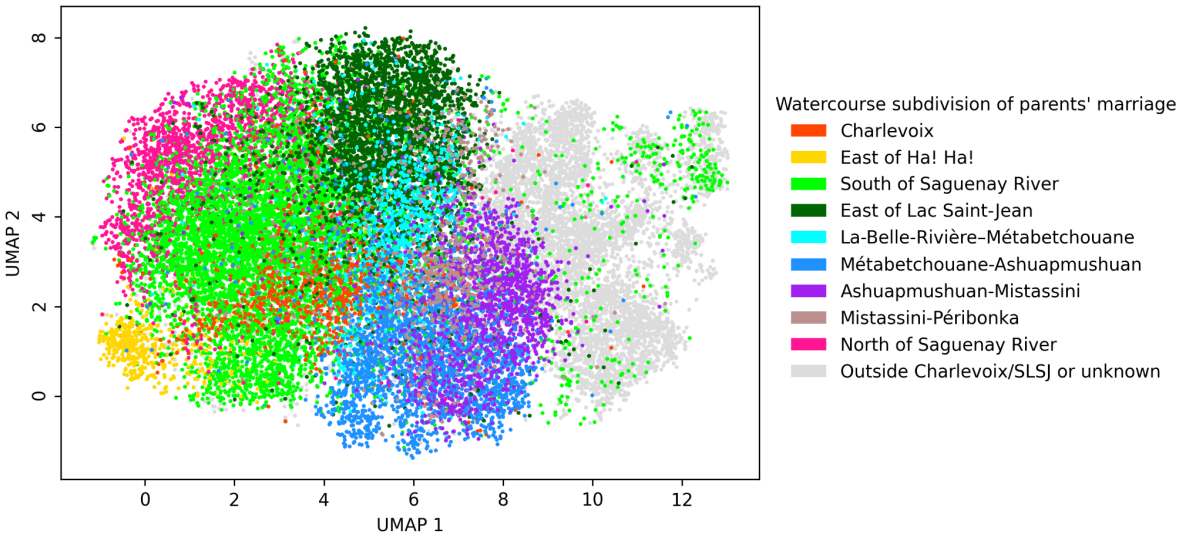


Fig. 2: UMAP of pairwise expected kinship matrix of individuals with parents married in SLSJ in 1931–1960, initialised on MDS 1 and 2.

For pedigree PC, there are several options including the paper in which the authors cited (https://gsejournal.biomedcentral.com/articles/10.1186/s12711-025-00994-y).

We did not use *randPedPCA*, as it appears to incorporate all ancestors in the genealogy and to add a genealogical depth in the response PCA, which was beyond the scope of this study. Our focus was on comparing contemporary probands, i.e. the last generation in the genealogy, based on their genealogical and genotype data.

2. How robust is the HDBSCAN/UMAP based classification? Visually, the points are very crowded and putting lines between them feels very artificial. One could try to adjust the parameters to check robustness, which the authors have already done to some extent, but we have a better alternative in this case. Since the migration patterns of Quebecois are already documented to a certain extent, it will be possible to run a population genetic simulation using msprime or SLiM (https://tskit.dev/) (or perhaps another suitable simulation scheme) . Then, authors could apply their pipeline to the simulated data to see if it recovers the demographic pattern codified into the simulation. The resulting tree sequence is amenable for kinship extraction (e.g. https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1011537) which should be amenable to the paper's pipeline.

The reviewer is right that the HDBSCAN clustering feels artificial and can't be visually validated on our actual Supplementary Fig. 3. We apologise for the confusion caused by this figure. It gives the impression that the regions are very close to each other, but this is due to the use of two-dimensional UMAPs for visualisation purposes. In reality, the clustering was performed using UMAPs with ten dimensions combined with HDBSCAN, this was stated in the text (line 161). When visualising these ten dimensions, it becomes clear that the SLSJ region is indeed distinct from the

other regions in both our IBD data (realised kinship) and genealogical data (expected kinship). We have replaced Supplementary Fig. 3 with a visualisation showing all ten UMAP dimensions used for the HDBSCAN clustering. As stated in the text, the SLSJ cluster was chosen as the cluster containing the highest number of individuals of known SLSJ origin i.e. with grandparents married in SLSJ.

Please also refer to our response to reviewer #3 for a related question, and see Fig. 3 below, which illustrates the overlap between the HDBSCAN cluster based on the 10-dimensional UMAP and the SLSJ region identified through PCA of common SNPs.

3. What is the time scaling of the new algorithm? Since it has to materialize an N-by-N kinship matrix, it should be at least O(N^2). I wonder if this is indeed the case.

We have added a paragraph under the pseudocode in the supplementary materials to describe the time complexity. It is the same as for Kirkpatrick's recursive-cut exact algorithm (https://doi.org/10.1093/bioinformatics/bty725), i.e. $O(N^2G)$ where N is the number of individuals in the generation and G is the number of generations. Our improvement lies in the parallelisation of this process.

While reassessing the time complexity of our hybrid algorithm, we noticed that by sorting the individuals during the preprocessing, the production of vertex cuts had a time complexity of $O(N \log N)$ instead of $O(N)$ as we expected, where N is the number of individuals in the genealogy. Since the ordering of the individuals does not matter within the intermediate generations, we have removed this sorting from our pseudocode and from the GeneaKit implementation.

Reviewer #3 (Remarks to the Author):

The study is thorough, methodologically sound, and well written; however, I have some reservations regarding its suitability for publication in Nature Communications. Indeed, while the study provides a detailed and quite convincing analysis of the Saguenay–Lac-Saint-Jean population (see below for specific comments on the presented analyses), the topic may be too specialized for the broad readership of Nature Communications. Moreover, the extent to which the proposed approach can be generalized remains unclear. The authors benefit from exceptionally rich datasets, both in terms of genotype information and genealogical records, which may not be available in other cases. Consequently, the broader applicability of their findings and methods might be limited.

It is also worth noting that the authors do not present new data, but rather reanalyze previously generated genotype and genealogical datasets. While the methodological optimization is valuable, the lack of novel data somewhat limits the overall originality and impact of the study.

We respectfully disagree with the reviewer that our work is a reanalysis of existing datasets. To our knowledge, this is the first comprehensive characterization of fine-scale structure across the entire SLSJ population, made possible by our exceptionally rich genealogical dataset. Previous studies were limited to subsets of the region due to the absence of tools capable of handling such large-scale genealogical data. Although we use the same genotype Quebec dataset as the McClelland *et al.* study[2] (recently accepted in Nature Communications), our objectives and contributions are different. The McClelland study highlights the broad diversity of Quebec and the opportunities offered by the biobank for health research, whereas our work uncovers the fine-scale genetic structure of the contemporary SLSJ population, a population shaped by a strong founder effect and long assumed to be homogeneous. Our work is therefore complementary, not redundant. Moreover, we believe that the richness of Quebec's data is not a limitation but a strength: it serves as a model for understanding genetic fine structure linked to recent demographic history, which occurs not only in Quebec, but possibly in many populations worldwide[3,4]. Therefore, the standard stratification correction strategies used in most GWAS and polygenic risk score analyses might not correct for this fine structure which can lead to biases and false associations.

Our goal is not to claim universal applicability of our method, but to demonstrate that fine-scale structure exists even at very small geographic scales and that accounting for it is relevant and necessary for genetic studies. Nevertheless, our approach can be applied to other populations wherever genotype data are available. Since our results are consistent on genealogical kinship and genomic IBD, this allows for the generalisation in any biobank or single population at low costs since only genotyping is required.

What is new in our study:

- Kinship coefficient calculation across the entire SLSJ population, described at a level of fine-scale structure that surpasses any previous study within Quebec's regional divisions.
- This was recently made possible to calculate pairwise kinship on billions of pairs of individuals with the new tools available, including the one that we present here, GeneaKit.
- Detecting such fine-scale structure within a relatively small and 'homogeneous' population suggests that fine structure exists in all populations. It underscores the need to control for population stratification in all GWAS and polygenic analyses, requiring adjustment beyond the first 20 principal components generally used and the continental population structure[5].

The authors investigate the genetic structure of the population through IBD sharing (realized kinship), aiming to efficiently compare genetic and genealogical relationships. It would, however, be informative to also explore the genetic structure directly from genotype data (e.g. SNP-based analyses), for instance through principal component analysis (PCA) or estimates of FST between clusters. Such analyses could provide complementary insights into the fine-scale population structure shaped by watercourses and the east-west migration patterns described in the study.

We do agree with the reviewer #3 that SNP-based analyses such as PCA and F_{ST} have been widely used to infer population structure. However, these methods excel at capturing ancient divergence and broad continental patterns, but often lack the resolution needed to detect subtle, recent demographic events such as fine-scale structure. This limitation arises because differences in SNP allele frequency accumulate slowly over time, making them less sensitive to recent demographic events[6]. In contrast, IBD has been shown to directly reflect recent shared ancestry, with segment length correlating to the number of generations since a common ancestor. This property enables IBD-based analyses to uncover finer geographic and genealogical structure within the last few hundred years, as demonstrated in other large-scale studies on the UK Biobank[4] and on New York City residents from the All of Us cohort[3]. Indeed, it was shown that when population structure is recent, it cannot be corrected using principal components of common variants because they are uninformative about recent history[5]. This can result in biases in GWAS, and even stronger biases in polygenic risks assessments.

Considering all these factors, we decided to continue using IBD so that we could apply the same methodology consistently across both genealogical and genotype data types.

In addition, the authors employ UMAP projections to summarize both genetic and genealogical data. While UMAP is a powerful dimensionality reduction method, it can sometimes introduce distortions that exaggerate the apparent separation between clusters. I therefore recommend that the authors also perform a PCA and subsequent clustering on the PCA coordinates (e.g. using k-means clustering) to verify whether the genetic structure observed in the UMAP plots is robust.

The question on the reliability of UMAPs has been addressed above, please refer to reviewer #2 answers to questions 1 and 2 for details.

Regarding clustering: We apologise for the confusion caused by our original Supplementary Fig. 3, which did not accurately represent the analysis. The figure suggested that regions were very close, but this was an artifact of using two-dimensional UMAP for visualisation. In reality, clustering was performed with UMAP using 10 dimensions combined with HDBSCAN. We now provide an updated Supplementary Fig. 3 showing all 10 UMAP dimensions.

To illustrate the differences between HDBSCAN clustering on UMAP and the proposed k-means clustering on PCA of SNPs, Fig. 3 below compares: PCA of common SNPs colored by parents' marriage region (top), k-means clustering on this PCA (k = 3 since this PCA clearly distinguishes a northeastern cluster (SLSJ) from an eastern cluster (Bas-St-Laurent and Côte-du-Sud)) (middle), and HDBSCAN clustering on the 10-dimensional UMAP derived from realised and expected kinship matrices (bottom).The SLSJ cluster (shown in red) is clearly visible along PC1, consistent with Quebec's previously described genetic structure based on common SNPs and genealogical data[7–9]. The agreement between methods is strong: 75% of the 938 individuals identified by HDBSCAN were also captured by k-means clustering on the PCA of common SNPs. This PCA in Fig. 3 below likely does not provide an optimal shape for k-means, as it fails to capture the full continuum of SLSJ ancestry. In contrast, the HDBSCAN clustering consists almost entirely of

individuals located along PC1 in the region with the highest concentration of parental marriages within the SLSJ.

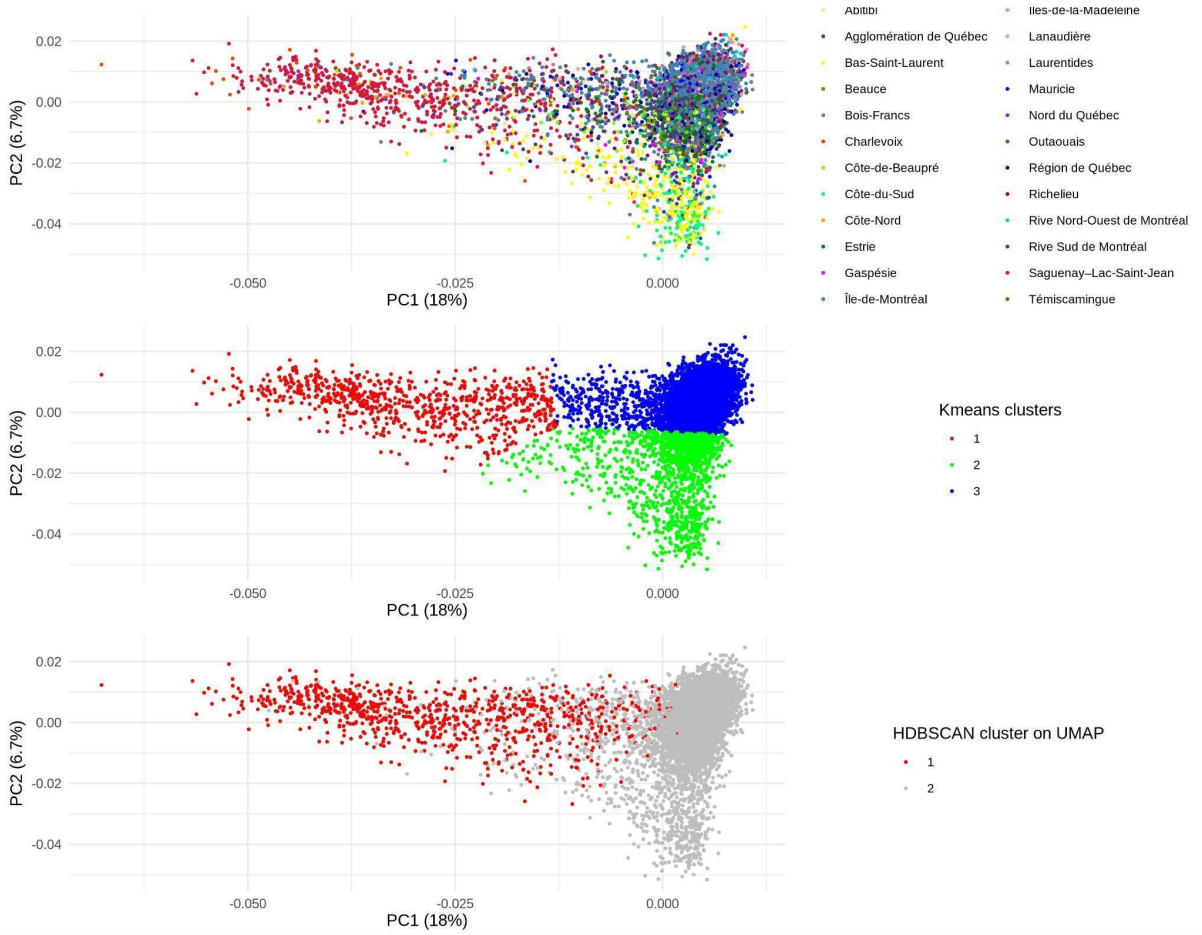


Fig. 3: PCA calculated on common SNPs of more than 7,000 Quebec individuals with genotype and genealogical data colored according to parents' region of marriage (top); K-means clustering (middle); and HDBSCAN clustering on a 10-dimensional UMAP (bottom).

Overall, this is a carefully executed study that makes good use of both genetic and genealogical data to characterize the fine-scale population structure of the SLSJ region. The methodological work is solid, and the manuscript is clearly written. However, the study mainly reanalyzes existing datasets and focuses on a highly specific regional case. As a result, the broader generalizability and potential impact of the findings appear limited. While the methodological contribution and the integration of rich genealogical resources is appreciable, I am not convinced that the current version of the manuscript meets the breadth and impact expected for publication in Nature Communications. I would, however, encourage the authors to consider submission to a more specialized genetics or population genomics journal.

We thank the reviewer for this valuable comment. We have substantially revised the manuscript to present the work with broader generalisation and relevance for populations with less comprehensive data resources. We believe that access to such rich data offers a unique opportunity to understand the complexity of fine-scale structure in human populations globally. Indeed, if we can detect fine-scale structure within what was considered a 'homogeneous' founder population like SLSJ, it is reasonable to expect that similar recent signals exist in all populations, signals that are often obscured by coarse, ancient structure under the standard stratification correction methods used in most GWAS and polygenic risk score analyses[5]. Relying solely on PCA

of common SNPs has been shown to leave residual stratification, which can lead to serious biases and false associations. Our study emphasises why addressing this issue is critical.

Reviewer #4 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

We thank the reviewer #4 for their valuable comments.

References

1. Margaryan, A. *et al.* Population genomics of the Viking world. *Nature* **585**, 390–396 (2020).
2. McClelland, P. *et al.* A multi-ancestry genetic reference for the Quebec population. 2025.05.14.25327536 Preprint at https://doi.org/10.1101/2025.05.14.25327536 (2025).
3. Isshiki, M. *et al.* Genetic disease risks of under-represented founder populations in New York City. *PLOS Genetics* **21**, e1011755 (2025).
4. Nait Saada, J. *et al.* Identity-by-descent detection across 487,409 British samples reveals fine scale population structure and ultra-rare variant associations. *Nat Commun* **11**, 6130 (2020).
5. Zaidi, A. A. & Mathieson, I. Demographic history mediates the effect of stratification on polygenic scores. *eLife* **9**, e61548 (2020).
6. Sellinger, T., Johannes, F. & Tellier, A. Improved inference of population histories by integrating genomic and epigenomic data. *eLife* **12**, RP89470 (2024).
7. Gagnon, L., Moreau, C., Laprise, C., Vézina, H. & Girard, S. L. Deciphering the genetic structure of the Quebec founder population using genealogies. *European Journal of Human Genetics* **32**, 91–97 (2024).
8. Bherer, C. *et al.* Admixed ancestry and stratification of Quebec regional populations. *American Journal of Physical Anthropology* https://doi.org/10.1002/ajpa.21424 (2011) doi:10.1002/ajpa.21424.
9. Roy-Gagnon, M. H. *et al.* Genomic and genealogical investigation of the French Canadian founder population structure. *Human Genetics* **129**, 521–531 (2011).

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

We appreciate that the editor resubmitted our work for peer review and are pleased that the reviewers recognised the broader applicability of our approach through the updated manuscript.

Reviewers' comments:

Reviewer #2 (Remarks to the Author):

The authors have made substantial efforts to revise their manuscript and I only have one major concern left. It's an overreaching claim to say that any magnitude of population structure can induce spurious associations in association studies. It is in fact the realized allele frequency stratification that creates excess false positives. For instance, Edge et al. is a good reference to this point (https://pubmed.ncbi.nlm.nih.gov/24481204/). Also, see Rosenberg et al. (https://academic.oup.com/genetics/article/173/3/1665/6060953).

We thank the reviewer for their comment. As rightly pointed out, we have adjusted the manuscript to clarify that population structure may lead to spurious associations only through correlation between allele frequencies and phenotypic variation.

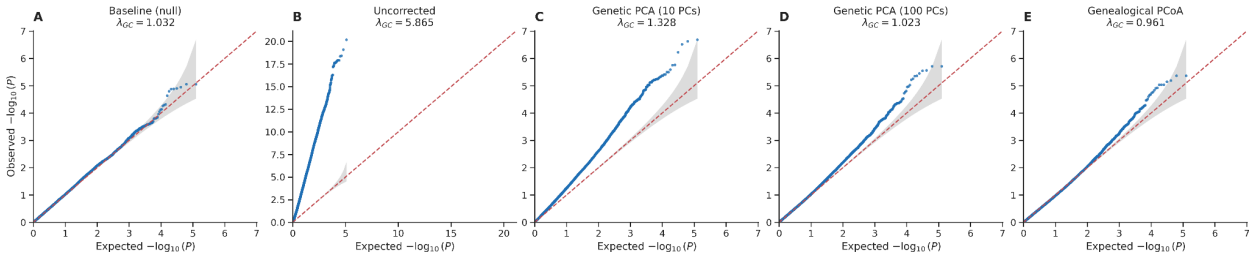
If the authors want to keep up with this claim, they should run simulations to confirm their findings: since they have a pedigree, simulate genomes conditioned on the pedigree and run GWAS on them. I'll recommend using msprime (https://tskit.dev/msprime/docs/stable/pedigrees.html), which can easily generate genomes conditioned on the pedigree. Then, using the genotypes from the data, one can further simulate traits and conduct GWAS to check the claim.

We recognise the value of performing simulations to confirm the presence of population stratification when running GWAS on the SLSJ data. We have added in the manuscript Supplementary Fig. 7 (copied below) and its corresponding Supplementary Methods and References. We first simulated genomes using msprime conditioned on the genealogy for all unrelated SLSJ individuals (above first cousins). Knowing that the correlation between our PCoA 1 of expected kinship coefficients and expected genetic contributions (GC) from Charlevoix founders is high (Pearson's $r = 0.97$, $p < 0.001$), we simulated a continuous trait based on the Charlevoix GC using this formula:

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

where y is the continuous trait, GC is the normalized Charlevoix GC, h^2 is the simulated heritability (0.1), and $N(0, 1)$ is a random noise of normal distribution.
We then ran a GWAS and observed inflation due to stratification on the QQ plot when no correction was applied. These results demonstrate that standard PCA correction may be insufficient for GWAS in populations with fine-scale structure like SLSJ, if a trait happens to be correlated with this fine structure.

Moreover, since SLSJ's fine-scale structure is highly predicted by its Charlevoix ancestry, and this Charlevoix ancestry has been shown to highly correlate with the presence of rare diseases in the region, we think it is reasonable to conclude that some phenotypic variations may correlate with SLSJ's population structure and cause spurious associations if left uncorrected, as shown in our simulation.



Supplementary Fig. 7: PCA does not adequately control for inflation of GWAS on a simulated phenotype correlated with SLSJ fine-scale structure. QQ plots for simulated data (N = 10,000) across 100 Mb of neutral genomic sequence using `msprime` conditioned on the genealogy. The phenotype was simulated to be partially correlated with the expected genetic contribution (GC, computed on the genealogy) from Charlevoix founders (h^2 = 0.1) since this GC was shown to be a major factor influencing the SLSJ fine structure. Panels display association results for: (A) a baseline null phenotype (pure noise); (B) the correlated phenotype with no correction; (C–D) the correlated phenotype corrected using the top 10 and 100 genetic principal components (PCA) of common SNPs (MAF < 0.05); and (E) the correlated phenotype corrected using the top 2 coordinates from a PCoA derived from the genealogical kinship matrix. The genomic inflation factor (λ_{GC}) is reported for each model to quantify test statistic inflation. The red dashed line represents the null expectation ($y = x$), and the gray shaded region indicates the 95% confidence interval.

Reviewer #3 (Remarks to the Author):

The authors have provided convincing arguments addressing the most delicate point raised in my previous report, namely the specificity of the results and their limited generalizability. Their response is detailed and, this time, persuasive. In light of this, considering the opinions of the other referees and the revisions introduced in the new version of the manuscript, I now believe that the work is appropriate for publication in Nature Communications.

We thank the reviewer #3 for their valuable contribution that improved our manuscript.

Reviewer #4 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

We thank the reviewer #4 for their valuable contribution.