PLOS ONE

RESEARCH ARTICLE

Fine-scale genetic structure and rare variant frequencies

Laurence Gagnon[1,2], **Claudia Moreau**[1,2], **Catherine Laprise**[1,2,3], **Simon L. Girard**[1,2,4]*

1 Département des Sciences Fondamentales, Université du Québec à Chicoutimi, Saguenay, Québec, Canada, **2** Centre Intersectoriel en Santé Durable, Université du Québec à Chicoutimi, Saguenay, Québec, Canada, **3** Centre Intégré Universitaire en Santé et Services Sociaux du Saguenay–Lac-Saint-Jean, Saguenay, Québec, Canada, **4** Centre de recherche CERVO, Université Laval, Québec, Québec, Canada

* simon2_girard@uqac.ca

OPEN ACCESS

Citation: Gagnon L, Moreau C, Laprise C, Girard SL (2024) Fine-scale genetic structure and rare variant frequencies. PLoS ONE 19(11): e0313133. https://doi.org/10.1371/journal.pone.0313133

Editor: Nejat Mahdieh, Shaheed Rajaei Hospital: Rajaie Cardiovascular Medical and Research Center, ISLAMIC REPUBLIC OF IRAN

Received: May 21, 2024

Accepted: October 19, 2024

Published: November 5, 2024

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0313133

Data Availability Statement: The Quebec Regional Reference Sample genotypes are freely available from the institutional repository of BALSAC (DOI: https://doi.org/10.5683/SP3/A5O8DT) and unaffected individuals from the Saguenay-Lac-St-

Abstract

In response to the current challenge in genetic studies to make new associations, we advocate for a shift toward leveraging population fine-scale structure. Our exploration brings to light distinct fine-structure within populations having undergone a founder effect such as the Ashkenazi Jews and the population of the Quebec' province. We leverage the fine-scale population structure to explore its impact on the frequency of rare variants. Notably, we observed an 8-fold increase in frequency for a variant associated with the Usher syndrome in one Quebec subpopulation. Our study underscores that smaller cohorts with greater genetic similarity demonstrate an important increase in rare variant frequencies, offering a promising avenue for new genetic variants' discovery.

Introduction

Common variants are the primary source of variation identified by genetic association studies. However, despite numerous association analyses conducted in the last years, a significant proportion of the genetic predisposition for many diseases still remains unknown. To address this issue, it is crucial to study rare variants, which often have more significant phenotypic effects, to better understand the "missing heritability". Nevertheless, associations with rare variants present a reduction in statistical power due to the scarcity of individuals carrying these alleles [1].

Several challenges have been encountered in rare variant association analyses, such as accounting for population structure which is an important cofounding factor [2]. Various methods have been proposed to address these challenges and adjust for different confounding effects [2]. Considering this, it is thus essential to explore new creative approaches to overcome all these challenges and facilitate the identification of rare variants.

Therefore, this study aims to use the fine-scale genetic structure of population cohorts as a tool to identify rare variants. We propose that rather than correcting for all confounders, it would be more powerful to seek for rare variants in cohorts with greater genetic similarity. Indeed, given their unique structure, populations that had undergone a founder effect (PFE) have the potential to more readily reveal new genetic associations of rare variants that could

Jean asthma familial cohort are available at (DOI: https://doi.org/10.5683/SP3/EXHNLH). The Ashkenazi Jews cohort data is available via dbGaP study accession number phs000448.v1.p1, the Himba cohort data is available via dbGaP study accession number phs001995.v1.p1 and the Hutterites cohort data is available via dbGaP study accession number phs001033.v1.p1 (https://www.ncbi.nlm.nih.gov/gap/). The code used for this study can be found in the following GitHub repository: https://github.com/laugag17/world_pop_with_founder_effect.

Funding: This work was supported by funding from the Canada Research Chair in Genetics and Genealogy held by SLG. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

have implications for human health [3–8]. Hence, we will use some of these populations, specifically the Quebec population, Ashkenazi Jews, Himba and Hutterites, to achieve our goals.

Subjects and methods

This study was approved by the University of Quebec in Chicoutimi (UQAC) Ethics Board. All datasets were accessed on November 11, 2022, and the authors do not have access to any information that could identify individual participants. All methods are summarized in Fig 1.

Cohorts

The data consist of four different cohorts of PFE. We gained access to data from Quebec, Ashkenazi Jews, Himba and Hutterites (S1 Table in S1 File). We also used the data from the 1000 Genomes Project phase 3 as reference groups from Africa (Mende (MSL)), Europe (British, Northern and Western Europe (GBR and CEU)) and East Asia (Han Chinese and Japanese (CHB and JPT)) as outbred control populations.

Genotyping data cleaning and imputation

Each individual dataset underwent cleaning using PLINK software v1.9, ensuring individuals with at least 95% genotypes among all SNPs were retained [9]. At the SNP level, we retained SNPs with at least 95% genotypes among all individuals, located on the autosomes and in Hardy–Weinberg equilibrium $p > 0.001$ (calculated on each cohort).

Subsequently, all datasets were merged (lifting over to hg19 for the Ashkenazi Jews) to retain only common (intersection in all datasets) bi-allelic SNPs. After the merge, individuals with less than 95% genotypes among all SNPs and SNPs with less than 95% genotypes across all individuals were once again filtered out. The final dataset comprises 199,238 SNPs and 4,259 individuals. Related individuals (PLINK pihat > = 0.25) were filtered out to avoid biases in the population structure definition, resulting in a final sample size of 3,683 subjects [10]. This unusually high threshold was applied to retain two populations with high relatedness (first and second degree) (S2 Table in S1 File). The S1 Fig in S1 File demonstrates the low impact of different genetic relatedness thresholds on the pairwise sum of identity-by-descent (IBD) segments length and number. The merged dataset was imputed on TOPMed imputation server, using the reference panel topmed-r2 after lifting over to hg38 [11]. Postimputation quality control filters were applied to remove SNPs with an imputation quality score <0.3 and only biallelic SNPs were kept for further analyses.

PCA, UMAP and clustering

A Principal Component Analysis (PCA) was done on the merged dataset for visualization and on each individual dataset for clustering using SNPs with a minor allele frequency (MAF) of at least 5%, and after pruning (—indep-pairwise 50 5 0.2) to remove SNPs in linkage disequilibrium.

A Uniform Manifold Approximation and Projection (UMAP) method was performed on the PCA data to investigate the population structure. Pre-processing the data with PCA before performing dimensional reduction with UMAP has been shown to be the most effective computational strategy for exploring population structure within large admixed datasets [12, 13]. UMAP was chosen for its ability to emphasize local data structure while preserving the global structure and for its effectiveness in cluster definition [12, 14, 15]. UMAPs were completed on each individual dataset to form clusters, using the first 5, 4, 7, 8 principal components (PCs) for the Ashkenazi Jews, Quebec, Himba and Hutterites, respectively (S2 and S3

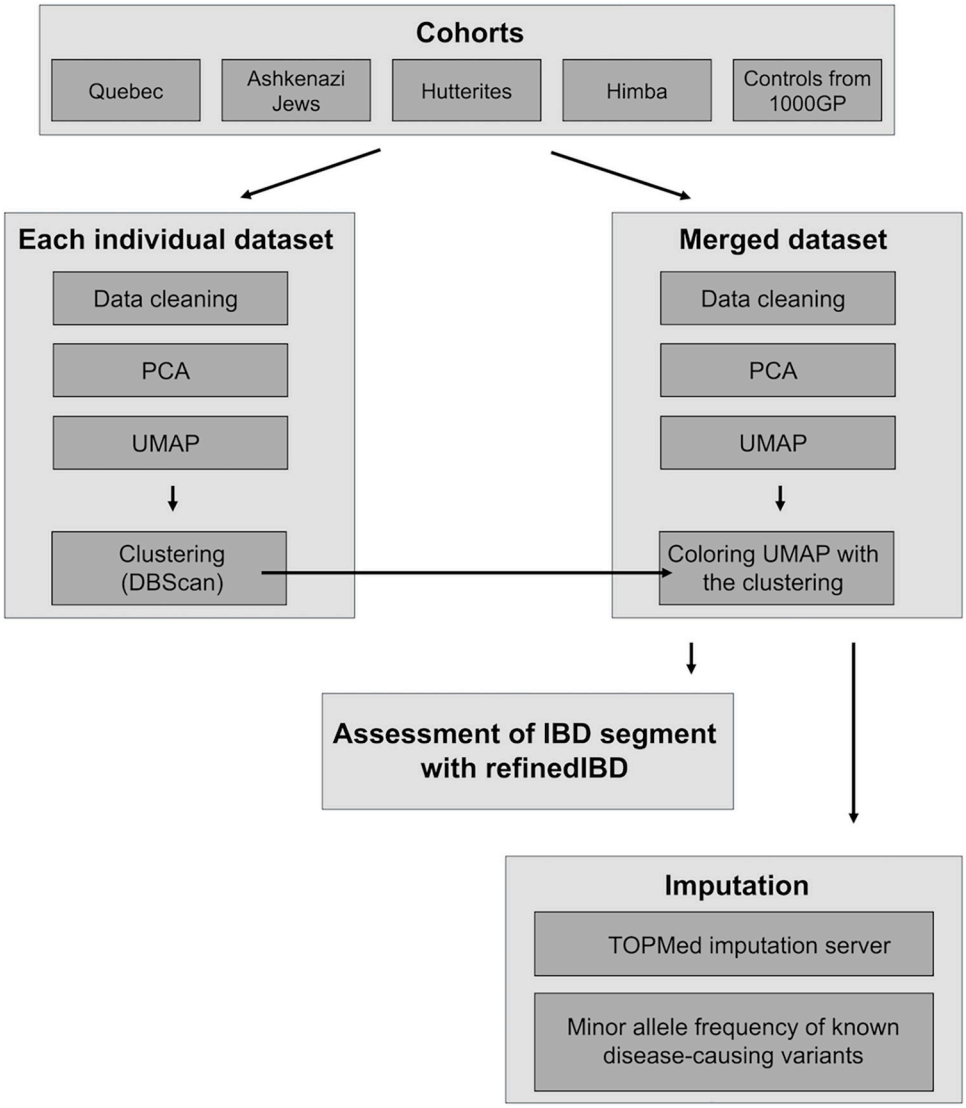

Fig 1. Simplified view of the flow of analysis.

https://doi.org/10.1371/journal.pone.0313133.g001

Figs in S1 File). The number of PCs was determined by the elbow of the scree plot. Another UMAP was done on the first 8 principal components of the PCA of the merged dataset to visualize the population structure (S4 Fig in S1 File). The UMAPs were realized with the R package "umap" v0.9.2.0 [16]. The n neighbors variable was set to the maximal value (number of individuals in the dataset) (S1 Table in S1 File). The min distance value was set to 0.01 for the UMAPs on each individual dataset to better capture the structure and promote clustering; while it was set to 0.9 for the UMAP on the merged dataset to promote dots splitting and ensure good visualization [17].

The UMAPs of the individual dataset were used to form distinct clusters on each population with the DBScan methods. This method was chosen for its ability in density-based clustering, allowing the capture of clusters with various shapes, including non-convex shapes [18]. The "dbscan" R library v1.1–11 was used for clustering with the minPts parameter set to 4, as the data is two-dimensional [19], and a K-distance graph was done on each dataset to select the

epsilon value from the elbow curve [20]. For the Ashkenazi Jews only, the epsilon value was increased to allow for the formation of a larger cluster that can be seen visually (S3A Fig in S1 File) [20]. The Himba and Hutterites only displayed a single cluster each probably, linked to their lifestyle and reflecting their polygyny and endogamy respectively [21, 22], while Ashkenazi Jews and Quebec exhibited 4 and 5 clusters, respectively (S3 Fig in S1 File). These clusters were afterwards used to color the UMAP of the merge dataset.

Statistical analysis

Minor allele frequency (MAF) of known disease-causing variants were computed using PLINK software v1.9 on imputed data. The variants were selected for their previously described association to the specific PFE of Quebec (Saguenay–Lac-Saint-Jean (SLSJ) and the Acadians of Gaspe (S5B Fig in S1 File)) and Ashkenazi Jews [23–28]. The MAF was computed both for entire PFE and for the distinct clusters. The MAF of the entire PFE was calculated based on 1000 permutations of the number of individuals in the clusters exhibiting the highest MAF. A p-value was determined by dividing the number of permutations exceeding the MAF of the corresponding cluster by the total number of permutations (1000). The founder variants (listed in Table 1 and S3 Table in S1 File) were selected because of their well-defined association with specific populations. However, limited literature is available concerning specific variants in the Acadians of Gaspe (Quebec-3) which could explain why we found only one variant of interest in this cluster.

Analysis of IBD segments

The assessment of pairwise IBD segments was performed on the merged dataset of all populations using refinedIBD software v17Jan20 on phased genotypes, which was done using Beagle software version 18May20.d20 [29]. The identified segments were then merged with merge-ibd-segments.17Jan20.102.jar. This software was selected for its robustness and precision in detecting IBD segments [29]. Only segments of 2 cM or more and with a LOD score greater than 3 were retained for further analysis on the level of IBD sharing across the genome.

Results

Fine-scale population structure

We characterized the genetic structure of four PFE alongside three reference groups from the 1000 Genomes Project with the aim of using this structure on variants' frequency analysis. A UMAP analysis was conducted to analyze their genetic structure (Fig 2A). The Himba and Hutterites form one tightly packed cluster each, whereas the Ashkenazi Jews and Quebec

Table 1. Frequency of disease-causing variants known to be associated with a specific population.

Population (count)	Associated cluster (count)	Disease	SNP	MAF of the European reference group of 1000GP	Mean MAF of 1000 permutations (P-Value)	MAF of the associated cluster
Ashkenazi Jews (2052)	Ashkenazi Jews-1 (1729)	Familial dysautonomia (ClinVar: 6085)	chr9:108899816: A:G	0.000	0.018 (< 0.001)	0.021
Quebec (941)	Quebec-2 (233)	Spastic ataxia of Charlevoix-Saguenay (ClinVar: 5512)	chr13:23335031: TA:T	0.000	0.008 (<0.001)	0.032
Quebec (941)	Quebec-3 (52)	Usher syndrome type I (ClinVar: 5143)	chr11:17531431: C:T	0.000	0.005 (0.002)	0.039

Only the main population and the associated cluster are shown in the table.

https://doi.org/10.1371/journal.pone.0313133.t001

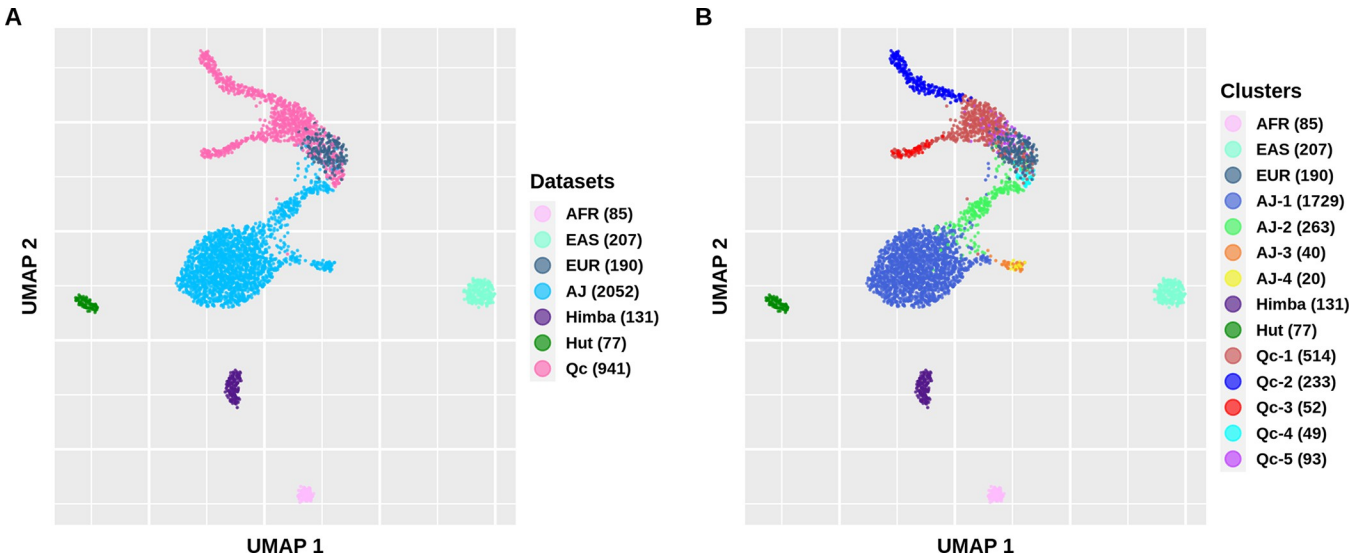

Fig 2. UMAP of the first 8 principal components of the merged dataset. Colored according to the origin of the population (A) and the clusters (on S3 Fig in S1 File) (B). AFR African from the 1000 Genomes Project, EAS East Asian from the 1000 Genomes Project, EUR European from the 1000 Genomes Project, AJ Ashkenazi Jews, Hut Hutterites, Qc Quebec.

https://doi.org/10.1371/journal.pone.0313133.g002

exhibit a more dispersed pattern and are even interconnected. The Ashkenazi Jews and Quebec went through unique histories of migration, isolation and population expansion. The Ashkenazi Jews-1 cluster possibly represents the Ashkenazi Jewish ancestry, i.e. individuals for whom all four grand-parents were of Ashkenazi Jewish ancestry (S5A Fig in S1 File). In comparison, the Ashkenazi Jews-2 cluster appears to represent a more admixed ancestry, due to its connection to the European reference group and Quebec (Fig 2B). This cluster likely contains individuals with only 1 to 3 of their grand-parents with Ashkenazi Jewish ancestry [7]. Moreover, the analysis of the genetic relatedness among and between clusters reveals that the Ashkenazi Jews clusters are distinct and exhibit greater relatedness within than between clusters (S6A Fig in S1 File). As for Quebec, clusters can be associated with specific ethnocultural groups. Specifically, the Quebec-2 and Quebec-3 represent the Saguenay–Lac-St-Jean (SLSJ) and the Acadians of Gaspe, respectively (S5B Fig in S1 File). These two groups are known for having a genetic structure which distinguishes them from the broader Quebec population that can be associated with Quebec-1 cluster [30–32]. The latter may be related to the initial founder effect in Quebec probably reflecting more diversity amongst founders. This is evident in the lower genetic relatedness observed within the Quebec-1 cluster compared to all the other regional clusters (S6B Fig in S1 File). Undoubtedly, populations like the Ashkenazi Jews and Quebec cannot be treated as single entities due to the presence of fine-structure, even if they were initially perceived as "homogenous populations".

Impact of using population clusters with higher genetic similarity on variants' frequency

To assess the impact of using the fine-scale population structure on variants' frequency, we looked for variants associated to known founder diseases. These variants were selected for their previously reported genetic association among a specific group. Comparing the frequency of these imputed variants within the associated groups (Quebec-2 for SLSJ, Quebec-3 for Acadians of Gaspe, or Ashkenazi Jews-1 for Ashkenazi Jews) (S5 Fig in S1 File), revealed higher allele frequencies in the specific cluster compared with the other clusters or the whole

population (Table 1). They were also nearly absent from the European reference group. Remarkably, within the Quebec populations, the variants associated with spastic ataxia of Charlevoix-Saguenay and Usher syndrome type I exhibit a 4 (P-Value of 1000 permutation <0.001) and 8 (P-Value of 1000 permutation 0.002) fold increase, respectively, when comparing the specific cluster and the whole population allele frequencies. Notably, this was calculated with a much smaller sample size of 4 and 18-fold, respectively. This trend is also observable while investigating the variant of Familial dysautonomia in the Ashkenazi Jews and other diseases associated with a specific population (Table 1 and S3 Table in S1 File).

Discussion

This study demonstrates that leveraging fine-scale population structure to intensify the presence of rare variants inside subpopulation clusters could be a promising avenue towards identifying rare variants of potential clinical interest. This approach might also address some challenges related to rare variants' identification. Firstly, rather than controlling for population structure as a confounder [2], we address the impact of population structure directly by using population subdivisions with more similar genetic background. Secondly, concentrating rare variants in smaller cohorts rises their relative frequency, making them less rare and thereby reducing the necessity for huge cohorts and increasing cost-effectiveness for existing cohorts. Therefore, thinking about incorporating fine-scale population structure while designing rare variants' studies can reduce the need for correction algorithms, thereby addressing some of the challenges associated with rare variant associations. This approach could be useful not only in PFE [33], but also in other isolated or even outbred populations since the presence of clusters in more diverse or admixed populations can have striking effect on variants' frequency [34, 35].

Investigating potentially pathogenic variants that are more frequent within targeted populations has the potential to generate positive impacts on public health at the community level and on the discovery of new genes that could be new therapeutic targets. The present study underscores the importance of focusing on smaller populations' fine-scale genetic structure. Indeed, the Acadians in New Brunswick were recently investigated for the first time regarding the frequency of disease-causing variants and the need to study individuals of Acadian ancestry across Canada Atlantic Provinces [24]. Given their unique population structure [30, 31, 36], it would be valuable to investigate the Acadians of Gaspe for potentially pathogenic alleles that may be more frequent in this population. In fact, despite increasing evidence of a strong and recent founder effect in the Gaspesian population of Quebec, which may have led to notable local changes in allele frequencies, this regional population has been relatively understudied [30]. In contrast, other regional populations of Quebec, such as the SLSJ, and also the Ashkenazi Jews have been the focus of many studies for rare genetic variants that have increased in frequency due to the founder effect [6, 7, 28, 37–39]. Even then, revisiting these populations using fine-structure in population cohorts, might reveal some yet undiscovered rare variants that are more frequent in these specific groups.

As for Hutterites and Himba, they can be studied as a whole as they do not subdivide into clusters. Indeed, Hutterites are known to practice endogamy and live in community and Himba individuals were documented to practice polygyny and live in a pastoralist way [21, 22]. This way of living promotes very close links between individuals and could explain the absence of population subdivision at the level tested. This is also evident in their proportion of pairs sharing an IBD segment across the genome, which reaches the highest level among all PFE (S4 Table in S1 File). However, the Hutterites have been extensively studied for more prevalent diseases within the population [40, 41], whereas the Himba have not, despite their

unique history and the precedence of increased homozygosity [42]. Our findings suggest that the Himba should be a center of interest in rare variant studies. This highlights the issue that genetic studies are predominantly focused on populations of European ancestry [43–45]. As a result, there is a lack of understanding of the genetic diversity present in other populations, which could harbor rare variants of significant importance. By shifting the focus to fine-scale genetic structures—whether analyzing entire populations or their subdivisions—researchers could improve the identification of rare variants by concentrating on individuals with more similar genetic backgrounds. This approach not only enhances our understanding of genetic diversity across different populations but also opens new avenues for discovering rare variants that may contribute to human health and disease.

There are some limitations present in this work. UMAP tends to prioritize local distances, which means that points that are close together in the projection are likely to be similar in the dataset. However, points that are farther apart might not actually be very different [12]. Nevertheless, this allowed us to identify and easily select clusters of very similar individuals. We acknowledge that this method cannot visually represent genetic variation on a continuum, as previously questioned and discussed [46]. Furthermore, the use of imputed data might have led to the loss of some rare variants resulting in an underestimation of the frequency of certain rare variants. Despite this, we were able to demonstrate an increase in frequency through population clusters, although the number of variants tested was likely lower, as evidenced by the detection of only one variant in the Acadians of Gaspe.

In conclusion, we suggest a novel approach, parallel to the already existing strategies, that leverages the fine-scale genetic structure of a population cohort to refine analysis models for rare variant studies. This cost-effectiveness strategy would help to enhance the value of existing large cohorts. We believe that cohorts composed of fewer individuals with a common genetic background would help in discovering new rare genetic associations, as they would be easier to find given their increased frequency.

Supporting information

S1 File. Includes S1-S6 Figs and S1-S4 Tables.
(DOCX)

Acknowledgments

This work was made possible by the Digital Research Alliance of Canada which provided access to storage and computing resources. We are extremely grateful to all participants of this research. We would like to thank Hélène Vézina, Damian Labuda and their team for the Quebec Regional Reference Sample cohort constitution. LG received scholarship from the Fonds de recherche du Québec—Santé and the Canadian Institutes of Health Research. CL is the director of the Centre Intersectoriel en Santé Durable (http://www.uqac.ca/santedurable), the chairholder of the Canada Research Chair in the Genomics of asthma and allergic diseases (http://www.chairs.gc.ca) and co-holder of the Chaire en santé durable du Québec (http://www.chairesantedurable.ca).

Author Contributions

Conceptualization: Laurence Gagnon, Claudia Moreau, Simon L. Girard.

Data curation: Laurence Gagnon.

Formal analysis: Laurence Gagnon.

Funding acquisition: Simon L. Girard.

Investigation: Laurence Gagnon.

Methodology: Laurence Gagnon, Claudia Moreau.

Resources: Catherine Laprise.

Supervision: Claudia Moreau, Simon L. Girard.

Writing – original draft: Laurence Gagnon.

Writing – review & editing: Claudia Moreau, Catherine Laprise, Simon L. Girard.

References

1. Goswami C, Chattopadhyay A, Chuang EY. Rare variants: data types and analysis strategies. Ann Transl Med. 2021 Jun; 9(12):961. https://doi.org/10.21037/atm-21-1635 PMID: 34277761
2. Chen W, Coombes BJ, Larson NB. Recent advances and challenges of rare variant association analysis in the biobank sequencing era. Front Genet. 2022 Oct 6; 13. https://doi.org/10.3389/fgene.2022.1014947 PMID: 36276986
3. Casals F, Hodgkinson A, Hussin J, Idaghdour Y, Bruat V, de Maillard T, et al. Whole-exome sequencing reveals a rapid change in the frequency of rare functional variants in a founding population of humans. PLoS Genet. 2013; 9(9):e1003815. https://doi.org/10.1371/journal.pgen.1003815 PMID: 24086152
4. Lencz T, Yu J, Khan RR, Flaherty E, Carmi S, Lam M, et al. Novel ultra-rare exonic variants identified in a founder population implicate cadherins in schizophrenia. Neuron. 2021; 109(9):1465–78. https://doi.org/10.1016/j.neuron.2021.03.004 PMID: 33756103
5. Southam L, Gilly A, Süveges D, Farmaki AE, Schwartzentruber J, Tachmazidou I, et al. Whole genome sequencing and imputation in isolated populations identify genetic associations with medically-relevant complex traits. Nat Commun. 2017 May 26; 8(1):15606. https://doi.org/10.1038/ncomms15606 PMID: 28548082
6. Scriver CR. Human Genetics: Lessons from Quebec Populations. Annual Review of Genomics and Human Genetics. 2001; 2:69–101. https://doi.org/10.1146/annurev.genom.2.1.69 PMID: 11701644
7. Guha S, Rosenfeld JA, Malhotra AK, Lee AT, Gregersen PK, Kane JM, et al. Implications for health and disease in the genetic signature of the Ashkenazi Jewish population. Genome Biology. 2012 Jan 25; 13(1):R2. https://doi.org/10.1186/gb-2012-13-1-r2 PMID: 22277159
8. Carmi S, Hui KY, Kochav E, Liu X, Xue J, Grady F, et al. Sequencing an Ashkenazi reference panel supports population-targeted personal genomics and illuminates Jewish and European origins. Nat Commun. 2014 Sep 9; 5(1):4835. https://doi.org/10.1038/ncomms5835 PMID: 25203624
9. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet. 2007 Sep; 81(3):559–75. https://doi.org/10.1086/519795 PMID: 17701901
10. Wang J. Effects of sampling close relatives on some elementary population genetics analyses. Mol Ecol Resour. 2018 Jan; 18(1):41–54. https://doi.org/10.1111/1755-0998.12708 PMID: 28776944
11. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature. 2021 Feb; 590(7845):290–9. https://doi.org/10.1038/s41586-021-03205-y PMID: 33568819
12. Diaz-Papkovich A, Anderson-Trocmé L, Ben-Eghan C, Gravel S. UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. PLOS Genetics. 2019 Nov 1; 15(11):e1008432. https://doi.org/10.1371/journal.pgen.1008432 PMID: 31675358
13. Diaz-Papkovich A, Zabad S, Ben-Eghan C, Anderson-Trocmé L, Femerling G, Nathan V, et al. Topological stratification of continuous genetic variation in large biobanks. bioRxiv; 2023. p. 2023.07.06.548007.
14. Diaz-Papkovich A, Anderson-Trocmé L, Gravel S. A review of UMAP in population genetics. J Hum Genet. 2021 Jan; 66(1):85–91. https://doi.org/10.1038/s10038-020-00851-4 PMID: 33057159
15. McConville Santos-Rodríguez, Piechocki Craddock. N2D: (Not Too) Deep Clustering via Clustering the Local Manifold of an Autoencoded Embedding. In: 2020 25th International Conference on Pattern Recognition (ICPR). 2021. p. 5145–52.
16. McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:180203426. 2020;

17. Allaoui M, Kherfi ML, Cheriet A. Considerably Improving Clustering Algorithms Using UMAP Dimensionality Reduction Technique: A Comparative Study. In: El Moataz A, Mammass D, Mansouri A, Nouboud F, editors. Image and Signal Processing. 2020. p. 317–25.

18. Hahsler M, Piekenbrock M, Doran D. dbscan: Fast Density-Based Clustering with R. Journal of Statistical Software. 2019 Oct 31; 91:1–30.

19. Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In 1996. p. 226–31.

20. Sander J, Ester M, Kriegel HP, Xu X. Density-based clustering in spatial databases: The algorithm gdbscan and its applications. Data mining and knowledge discovery. 1998; 2:169–94.

21. Nimgaonkar VL, Fujiwara TM, Dutta M, Wood J, Gentry K, Maendel S, et al. Low prevalence of psychoses among the Hutterites, an isolated religious community. Am J Psychiatry. 2000 Jul; 157(7):1065–70. https://doi.org/10.1176/appi.ajp.157.7.1065 PMID: 10873912

22. Scelza BA. Female mobility and postmarital kin access in a patrilocal society. Hum Nat. 2011 Dec; 22 (4):377–93. https://doi.org/10.1007/s12110-011-9125-5 PMID: 22388944

23. Ebermann I, Lopez I, Bitner-Glindzicz M, Brown C, Karel Koenekoop R, Jörn Bolz H. Deafblindness in French Canadians from Quebec: a predominant founder mutation in the USH1C gene provides the first genetic link with the Acadian population. Genome Biol. 2007; 8(4):R47. https://doi.org/10.1186/gb-2007-8-4-r47 PMID: 17407589

24. Robichaud PP, Allain EP, Belbraouet S, Bhérer C, Mamelona J, Harquail J, et al. Pathogenic variants carrier screening in New Brunswick: Acadians reveal high carrier frequency for multiple genetic disorders. BMC Medical Genomics. 2022 Apr 29; 15(1):98. https://doi.org/10.1186/s12920-022-01249-1 PMID: 35488281

25. Wallace SE, Mirzaa GM. Resources for Genetics Professionals—Genetic Disorders Associated with Founder Variants Common in the Acadian Population. In: GeneReviews. University of Washington, Seattle; 2022.

26. Consult ARUP. Ashkenazi Jewish Genetic Diseases Panel. 2023. Ashkenazi Jewish Genetic Diseases Panel. Available from: https://arupconsult.com/ati/ashkenazi-jewish-genetic-diseases-panel#clinical-sensitivity

27. Charrow J. Ashkenazi Jewish genetic disorders. Fam Cancer. 2004; 3(3–4):201–6. https://doi.org/10.1007/s10689-004-9545-z PMID: 15516842

28. Cruz Marino T, Leblanc J, Pratte A, Tardif J, Thomas MJ, Fortin CA, et al. Portrait of autosomal recessive diseases in the French-Canadian founder population of Saguenay-Lac-Saint-Jean. Am J Med Genet A. 2023 May; 191(5):1145–63. https://doi.org/10.1002/ajmg.a.63147 PMID: 36786328

29. Browning BL, Browning SR. Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics. 2013 Jun; 194(2):459–71. https://doi.org/10.1534/genetics.113.150029 PMID: 23535385

30. Gagnon L, Moreau C, Laprise C, Vézina H, Girard SL. Deciphering the genetic structure of the Quebec founder population using genealogies. European Journal of Human Genetics. 2024 Jan 1; 32(1):91–7. https://doi.org/10.1038/s41431-023-01356-2 PMID: 37016017

31. Roy-Gagnon MH, Moreau C, Bherer C, St-Onge P, Sinnett D, Laprise C, et al. Genomic and genealogical investigation of the French Canadian founder population structure. Human Genetics. 2011 May; 129 (5):521–31. https://doi.org/10.1007/s00439-010-0945-x PMID: 21234765

32. Bherer C, Labuda D, Roy-Gagnon MH, Houde L, Tremblay M, Vézina H. Admixed ancestry and stratification of Quebec regional populations. American Journal of Physical Anthropology. 2011; 144(3):432–41. https://doi.org/10.1002/ajpa.21424 PMID: 21302269

33. Xue Y, Mezzavilla M, Haber M, McCarthy S, Chen Y, Narasimhan V, et al. Enrichment of low-frequency functional variants revealed by whole-genome sequencing of multiple isolated European populations. Nat Commun. 2017 Jun 23; 8(1):15927. https://doi.org/10.1038/ncomms15927 PMID: 28643794

34. Gouveia MH, Bentley AR, Leal TP, Tarazona-Santos E, Bustamante CD, Adeyemo AA, et al. Unappreciated subcontinental admixture in Europeans and European Americans and implications for genetic epidemiology studies. Nat Commun. 2023 Nov 7; 14(1):6802. https://doi.org/10.1038/s41467-023-42491-0 PMID: 37935687

35. Koyama S, Wang Y, Paruchuri K, Uddin MM, Cho SMJ, Urbut SM, et al. Decoding Genetics, Ancestry, and Geospatial Context for Precision Health. medRxiv. 2023 Oct; 2023.10.24.23297096. https://doi.org/10.1101/2023.10.24.23297096 PMID: 37961173

36. Gauvin H, Moreau C, Lefebvre JF, Laprise C, Vézina H, Labuda D, et al. Genome-wide patterns of identity-by-descent sharing in the French Canadian founder population. Eur J Hum Genet. 2014 Jun; 22 (6):814–21. https://doi.org/10.1038/ejhg.2013.227 PMID: 24129432

37. Bchetnia M, Bouchard L, Mathieu J, Campeau PM, Morin C, Brisson D, et al. Genetic burden linked to founder effects in Saguenay-Lac-Saint-Jean illustrates the importance of genetic screening test availability. J Med Genet. 2021 Oct; 58(10):653–65. https://doi.org/10.1136/jmedgenet-2021-107809 PMID: 33910931

38. Laberge AM, Michaud J, Richter A, Lemyre E, Lambert M, Brais B, et al. Population history and its impact on medical genetics in Quebec. Clin Genet. 2005 Oct; 68(4):287–301. https://doi.org/10.1111/j.1399-0004.2005.00497.x PMID: 16143014

39. Risch N, Tang H, Katzenstein H, Ekstein J. Geographic Distribution of Disease Mutations in the Ashkenazi Jewish Population Supports Genetic Drift over Selection. Am J Hum Genet. 2003 Apr; 72(4):812–22. https://doi.org/10.1086/373882 PMID: 12612865

40. Boycott KM, Parboosingh JS, Chodirker BN, Lowry RB, McLeod DR, Morris J, et al. Clinical genetics and the Hutterite population: a review of Mendelian disorders. Am J Med Genet A. 2008 Apr 15; 146A (8):1088–98. https://doi.org/10.1002/ajmg.a.32245 PMID: 18348266

41. Chong JX, Ouwenga R, Anderson RL, Waggoner DJ, Ober C. A Population-Based Study of Autosomal-Recessive Disease-Causing Mutations in a Founder Population. Am J Hum Genet. 2012 Oct 5; 91 (4):608–20. https://doi.org/10.1016/j.ajhg.2012.08.007 PMID: 22981120

42. Swinford NA, Prall SP, Williams CM, Sheehama J, Scelza BA, Henn BM. Increased homozygosity due to endogamy results in fitness consequences in a human population. bioRxiv; 2022. p. 2022.07.25.501261.

43. Fatumo S, Chikowore T, Choudhury A, Ayub M, Martin AR, Kuchenbaecker K. A roadmap to increase diversity in genomic studies. Nat Med. 2022 Feb; 28(2):243–50. https://doi.org/10.1038/s41591-021-01672-4 PMID: 35145307

44. Bustamante CD, Burchard EG, De La Vega FM. Genomics for the world. Nature. 2011 Jul 13; 475 (7355):163–5. https://doi.org/10.1038/475163a PMID: 21753830

45. Popejoy AB, Fullerton SM. Genomics is failing on diversity. Nature. 2016 Oct; 538(7624):161–4. https://doi.org/10.1038/538161a PMID: 27734877

46. Kozlov M. 'All of Us' genetics chart stirs unease over controversial depiction of race. Nature [Internet]. 2024 Feb 23; Available from: https://www.nature.com/articles/d41586-024-00568-w https://doi.org/10.1038/d41586-024-00568-w PMID: 38396099