Concordance and dissonance: A genome-wide analysis of self-declared versus inferred ancestry in 10,250 participants from the HostSeq cohort

René L Warren[1,*], Inanc Birol[1,2], CGEn HostSeq Initiative[1,3,4] https://www.cgen.ca/hostseq-contributing-studies-implementation-committee

[1]Canada's Michael Smith Genome Sciences Centre, Vancouver, Canada
[2]Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia, Canada
[3]McGill Genome Centre, McGill University, Montreal, QC, Canada
[4]The Hospital for Sick Children, Toronto, ON, Canada

*****corresponding author

e-mail: rwarren@bcgsc.ca

Abstract

Accurately characterizing human diversity is foundational to equitable genomics. In this study, we present a large-scale analysis comparing self-declared ancestry with genetically inferred ancestry in 10,250 participants from the pan-Canadian HostSeq cohort. Using the ntRoot algorithm on whole genome sequencing data, we inferred both global and local continental-level ancestry and assessed concordance with self-identified sociocultural categories. High agreement was observed among individuals self-identifying as White (concordance rate=98.8%), Black (97.2%), East Asian (96.1%), and South Asian (89.9%), while substantial discordance was found in those self-identifying as Hispanic (concordance rate=74.6%), Middle Eastern / Central Asian (67.9%) or Indigenous (40.7%). We quantified agreement using Cohen's kappa (κ = -0.01 unweighted; 0.35 weighted) and assessed admixture complexity with Shannon entropy, revealing a strong relationship between discordance and ancestry heterogeneity. Principal component analysis further revealed that tightly clustered genetic profiles often corresponded with lower admixture complexity, whereas broader, overlapping distributions were observed in groups with more heterogeneous ancestry and complex sociocultural histories. These findings underscore the complex interplay between sociocultural identity and genomic data, with discordance patterns reflecting the historical and cultural complexity of human populations. By quantifying this relationship systematically with ntRoot, our approach provides a framework for moving rigid categorical labels toward more nuanced genome-derived ancestry characterization that can improve both scientific rigor and representational equity in genomics.

Author Summary:

Study concept: RLW. Software implementation: RLW. Data analysis: RLW.

Manuscript development: RLW. Manuscript editing: RLW, IB. Funding acquisition: IB.

Introduction

Understanding how individuals self-identify in relation to their genetic ancestry is critical for advancing equity in biomedical research and healthcare (1). Self-declared ancestry reflects complex sociocultural, historical, and geopolitical realities, while genetically inferred ancestry provides a population genetics perspective based on single nucleotide variant (SNV) allele frequency distributions. Although both frameworks aim to describe human diversity, they often diverge in practice.

Previous studies have documented varying degrees of concordance between self-reported and inferred ancestry, particularly in admixed populations or those shaped by colonial histories (2,3). Such dissonance has implications for clinical interpretation, genetic risk assessment, and participant trust in research. Moreover, the uncritical use of race and ethnicity in biomedical contexts risks reinforcing structural inequities when not grounded in sociogenomic nuance (4–6).

Ancestry inference methodologies are typically classified as either global or local. Global Ancestry Inference (GAI) estimates genome-wide proportions of ancestry from broad reference groups, providing an overall summary of an individual's genetic background. In contrast, Local Ancestry Inference (LAI) assigns ancestry to specific genomic segments by leveraging the persistence of ancestral haplotype blocks through recombination (7). As a result, LAI captures admixture complexity at finer resolution and enables locus-specific ancestry mapping. Recently, we developed ntRoot, a scalable platform for both global and local ancestry inference using whole genome sequencing (WGS) data (8). Designed for computational efficiency and high-resolution analysis, ntRoot is optimized for application to population-scale datasets. Leveraging Canada's HostSeq initiative, a unique national resource comprising WGS and linked clinical data from over 10,000 individuals affected by COVID-19 (9), we applied ntRoot to infer both global and local ancestry and assess concordance with self-declared identity categories. This study represents one of the largest ancestry comparison efforts in a Canadian cohort, offering new insights into the intersection of genomic variation and sociocultural identity within a multicultural context.

Methods

Cohort Description

We analyzed data from 10,250 participants enrolled in the HostSeq project, representing a diverse cross-section of the Canadian population. Participants provided self-declared ancestry through structured questionnaires, selecting from categories aligned with national demographic standards.

Genetic Ancestry Inference

We used the ntRoot algorithm (v1.1.5) to perform ancestry prediction from WGS data, categorizing individuals into five continental-level population groups: AFR (African), AMR (Admixed American), EAS (East Asian), EUR (European), and SAS (South Asian). The ntRoot method employs a reference-guided framework optimized for WGS, targeting high-resolution classification. First, each individual read set in CRAM format was converted to fastq using the command: samtools fastq -@ 48 --reference GRCh38_full_analysis_set_plus_decoy_hla.fa -1 SR.1.fastq.gz -2 SR.2.fastq.gz sample.cram. Next, we executed ntRoot (v1.1.5, with parameters -k 55 --draft GRCh38.fa.gz --reads SR -t 48 -Y 0.55 -l 1000GP_integrated_snv_v2a_27022019.GRCh38.phased_gt1.vcf.gz --lai). A total of 10 batched jobs, each containing instructions to process 1,025 samples, ran for a total of 2 months and 3 weeks on server-class systems with Intel(R) Xeon(R) CPU E7-8867 v3 @ 2.50GHz and 2.6TB RAM.

Concordance Rate Estimation

We defined an expected global ancestry for each self-reported category based on sociogenomic assumptions, acknowledging the known admixture and complexity within each category (e.g., "Ashkenazi Jewish" → EUR, "East Asian / Pacific Islander" → EAS, etc.) (Table 1). Concordance was defined as a match between the expected global ancestry and the ancestry inferred from ntRoot, assigned based on the largest predicted LAI fraction. All data handling and analysis were performed using R (v4.3.1) with the dplyr (v1.1.2) package.

Cohen's Kappa Agreement Analysis

To evaluate categorical agreement beyond chance, we used the cohen.kappa() function from the psych package (v2.3.9) on a contingency table comparing expected versus inferred ancestry labels (alpha level = 0.05). We reported both unweighted and weighted kappa scores, the latter accounting for partial similarity across related continental superpopulation groups.

Shannon Entropy (Admixture Complexity)

To assess genomic admixture per individual, we calculated Shannon entropy from continental ancestry fractions (AFR, AMR, EAS, EUR, SAS) using the entropy package (v1.3.1). Entropy scores were then aggregated by self-declared category to identify which groups exhibit more mixed ancestry profiles.

Principal Component Analysis (PCA)

We conducted PCA on the ancestry fractions to visualize similarities and differences in genomic ancestry composition across self-declared identity groups. Principal components were plotted and colored by self-declared category to observe clustering and spread. PCA was computed using the base prcomp() function in R, and visualized using ggplot2 (v3.4.0).

Results

The HostSeq cohort comprises 10,250 individuals recruited across Canada through COVID-19 pandemic-related studies. While the largest proportion of participants (26.8%) self-identified as White (n = 2,743), smaller proportions identified as South Asian (n = 358; 3.5%), East Asian / Pacific Islander (n = 358; 3.5%), Black (n = 213; 2.1%), Hispanic (n = 185; 1.8%), Middle Eastern or Central Asian (n = 162; 1.6%), and Indigenous (First Nations, Métis, Inuit; n = 27; 0.3%). A notable fraction of individuals did not disclose their ancestry (n = 4,723; 46.1%), or selected "Prefer not to answer / do not know" (n = 1,168; 11.4%), "More than one race" (n = 171; 1.7%), or "Other" (n = 78; 0.8%) (Table 1). These uncategorized responses, which together represented the largest fraction (i.e., ~60%) of individuals, mapped across diverse ancestries, suggesting complex identity constructs or reluctance to engage with constrained identity categories, highlighting the limits of categorical identity labels. The aforementioned composition does not reflect the broader demographic diversity of Canada and likely results from limitations in cohort recruitment, persistent barriers to participation among racialized communities, and the constrained nature of categorical identity labels used in biomedical research.

Patterns of Discordance

Substantial discordance was observed in certain categories, including individuals of self-declared Hispanic, Indigenous and Middle Eastern / Central Asian backgrounds. Among participants identifying as Hispanic, most were assigned global ancestry labels of AMR (74.6%), EUR (15.1%), or AFR (8.6%) (Fig. 1), and showed appreciable local ancestry contributions from these same groups (Fig. 2). Many individuals identifying as First Nations, Métis, or Inuit clustered within either the AMR (40.7%), or EUR (59.3%) GAI predictions. For Middle Eastern or Central Asian, a similar dispersion across EUR (67.9%), SAS (19.8%), and AFR (12.3%) continental labels suggests limited genomic reference resolution and diverse ancestral backgrounds (Fig. 1A and 1B).

Overall, the average unweighted Cohen's kappa was –0.01, indicating minimal agreement beyond chance across all categories (Table 2, Supplementary Fig. 1). The average weighted kappa was 0.35, reflecting moderate agreement when accounting for partial similarity across ancestry types. Shannon entropy, a measure of admixture complexity where higher values

reflect greater ancestry heterogeneity, revealed that individuals identifying as Sephardic (mean entropy [m.e.]=2.04) and Ashkenazi Jewish (m.e.=1.91), Middle Eastern or Central Asian (m.e.=2.00), Indigenous (m.e.=1.78), White (m.e.=1.72), Hispanic (m.e.=1.71), and "More than one race" (m.e.=1.72) exhibited the highest levels of admixture, and South Asian (m.e.=1.36), Black (m.e.=0.474), and East Asian / Pacific Islander (m.e.=0.471) among the lowest (Table 3, Fig. 2). Principal component analysis reveals distinct clustering among individuals self-identifying as White, East Asian, South Asian, and Black (Supplementary Fig. 2), which is consistent with lower admixture entropy and high concordance with inferred genetic ancestry. In contrast, individuals identifying as Indigenous, Hispanic, or Middle Eastern / Central Asian exhibited broader, overlapping distributions in PCA space, reflecting more complex ancestry profiles and lower overall concordance between self-declared and genome-derived ancestry. These observations are consistent with the concordance table, which shows higher match rates for White (98.8%), Black (97.2%), East Asian / Pacific Islander (96.1%), and South Asian (89.9%) individuals, and lower rates for Hispanic (74.6%), Middle Eastern or Central Asian (67.9%) and Indigenous (40.7%) populations, according to each respective, assigned global ancestry (Table 1).

Proportional Ancestry Patterns

Figure 2 and Supplementary Fig. 3 present the ancestry composition of each self-declared group using stacked ancestry fractions. In Supplementary Figure 3, all individuals are shown in full cohort proportions whereas Fig. 2 normalizes by sample count per group to eliminate size bias. In those figures, the ancestry fractions tallied from LAI for each sample reveal that while categorical self-declared groups including Black, South Asian, and East Asian display tightly clustered and more homogenous genetic profiles, groups such as Hispanic, Indigenous, Ashkenazi and Sephardic Jewish, Middle Eastern and White contain a wider range and appreciable fractions of continental ancestries. This indicates that admixture is intrinsic to these identities, not an artifact of sample distribution.

Discussion

Despite relatively high overall concordance between self-declared identity categories and genome-derived global ancestry predictions, entropy calculations and local ancestry fractions revealed high admixture complexity in several groups. These included individuals identifying as Ashkenazi and Sephardic Jewish, Middle Eastern or Central Asian, Indigenous, White, and Hispanic, as well as those selecting “More than one race,” “Other,” “Prefer not to answer / Do not know,” or who did not self-declare ancestry, mirroring the broad clustering patterns observed in PCA. These results reinforce the association between entropy, discordance, and generally less distinct PCA clustering in groups with more complex admixture histories.

The analysis presented here affirms the general reliability of self-declared ancestry in broad continental contexts, while also highlighting the need for more nuanced interpretation in admixed or historically complex populations. Our findings are consistent with previous work (2,10) showing that discordance between self-identified and inferred ancestry is most pronounced in groups shaped by colonial histories or recent admixture. Importantly, we emphasize that neither genetic ancestry nor self-identification should be treated as inherently more valid. Each reflects a different lens on identity—biological and sociocultural—and their intersection offers a richer and more complete picture of human diversity.

For genomic research, these findings reinforce the need for flexible, multidimensional approaches to participant categorization, and call for a deliberate commitment to equity-informed study design (4). The inclusion of proportional ancestry analysis underscores that identity categories such as Ashkenazi and Sephardic Jewish, Middle Eastern or Central Asian, Indigenous, White and Hispanic are not genetically homogeneous groups but socially meaningful designations encompassing diverse ancestry profiles. These insights have critical implications for how populations are stratified in biomedical research, particularly in precision medicine frameworks that rely heavily on genetic clustering and inference.

Moreover, the incorporation of local ancestry inference (LAI) using whole genome sequencing (WGS) via ntRoot provides resolution that global ancestry inference (GAI) alone cannot achieve. While GAI summarizes continental-scale proportions, LAI identifies the ancestral origin of specific genomic segments, enabling locus-level insights into admixture and genetic architecture. This is particularly valuable in genome-wide association studies (GWAS), pharmacogenomics, and polygenic risk score development, especially in admixed or

underrepresented populations (11–13). Fine-grained ancestry analysis helps mitigate biases introduced by majority-European training datasets (14) and allows for the identification of ancestry-specific associations with complex traits. As such, LAI represents an essential complement to GAI in efforts to make genomic medicine more equitable and mechanistically precise.

This study has several limitations. First, it relies on categorical mapping assumptions that assign self-declared identities to expected genomic ancestry groups, which may oversimplify complex sociocultural constructs. Second, while ntRoot offers both global and local inference, its resolution is currently limited to continental-scale labels and may obscure the continuous and temporally deep nature of human population structure. As others have noted, these classifications can conflate distinct concepts such as race, ethnicity, geography, and genetic ancestry (15–18), which may limit their interpretability, particularly in diverse or historically admixed populations. Additionally, the self-declared data reflect heterogeneous interpretations of ancestry, race, and ethnicity by participants. Finally, reference panels used for ancestry inference may underrepresent certain populations, particularly Middle Eastern and Indigenous groups, potentially reducing accuracy in these categories (19,20). It is important to note that the Indigenous category in this Canadian cohort may encompass a range of identities, including First Nations, Inuit, and particularly Métis individuals, whose genetic ancestry often reflects historical admixture with European populations, which may partly explain our observations. These findings, therefore, may not generalize to Indigenous populations in other national contexts, and highlight the importance of region-specific considerations, including documented colonial and post-colonial histories, when interpreting genomic diversity within socially defined groups (21–23).

Conclusion

This work represents one of the largest efforts to date to compare self-declared and genome-derived ancestry predictions in a nationally diverse cohort. By combining whole genome-based ancestry inference with participant-reported identity data, we provide a picture of where genomic and sociocultural understandings of ancestry converge and where they diverge. These findings underscore the importance of integrating both biological and self-identified dimensions of

identity in genomic research. Doing so will be critical for developing inclusive, precise, and equitable research in population genomics.

Data and Code Availability

The data used in this study is available for download at CGEn (https://www.cgen.ca/hostseq-application-form) . ntRoot is freely available on GitHub (https://github.com/bcgsc/ntroot).

Funding

This study is supported by the Canadian Institutes of Health Research (CIHR) [PJT-183608]. The research presented here has been conducted using CGEn's HostSeq Databank, funded by the Government of Canada through Genome Canada, under Project ID 'DACO-5'.

References

1. Bonham VL, Green ED, Pérez-Stable EJ. Examining How Race, Ethnicity, and Ancestry Data Are Used in Biomedical Research. JAMA. 2018 Oct 16;320(15):1533.

2. Bryc K, Durand EY, Macpherson JM, Reich D, Mountain JL. The Genetic Ancestry of African Americans, Latinos, and European Americans across the United States. Am J Hum Genet. 2015 Jan;96(1):37–53.

3. Wojcik GL, Graff M, Nishimura KK, Tao R, Haessler J, Gignoux CR, et al. Genetic analyses of diverse populations improves discovery for complex traits. Nature. 2019 Jun;570(7762):514–8.

4. Committee on the Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research, Board on Health Sciences Policy, Committee on Population, Health and Medicine Division, Division of Behavioral and Social Sciences and Education, National Academies of Sciences, Engineering, and Medicine. Using Population Descriptors in Genetics and Genomics Research: A New Framework for an Evolving Field [Internet]. Washington, D.C.: National Academies Press; 2023 [cited 2024 Aug 27]. Available from: https://www.nap.edu/catalog/26902

5. Morning A. The Nature of RaceHow Scientists Think and Teach about Human Difference [Internet]. University of California Press; 2011 [cited 2025 May 21]. Available from: https://academic.oup.com/california-scholarship-online/book/13767

6. Roth WD. Racial Mismatch: The Divergence Between Form and Function in Data for Monitoring Racial Discrimination of Hispanics*. Soc Sci Q. 2010 Dec;91(5):1288–311.

7. Tang H, Peng J, Wang P, Risch NJ. Estimation of individual admixture: analytical and study design considerations. Genet Epidemiol Off Publ Int Genet Epidemiol Soc. 2005;28(4):289–301.

8. Warren RL, Coombe L, Wong J, Kazemi P, Birol I. Human ancestry inference at scale, from genomic data [Internet]. 2024 [cited 2024 Sep 6]. Available from: http://biorxiv.org/lookup/doi/10.1101/2024.03.26.586646

9. Yoo S, Garg E, Elliott LT, Hung RJ, Halevy AR, Brooks JD, et al. HostSeq: a Canadian whole genome sequencing and clinical data resource. BMC Genomic Data. 2023 May 2;24(1):26.

10. Via M, Gignoux CR, Roth LA, Fejerman L, Galanter J, Choudhry S, et al. History Shaped the Geographic Distribution of Genomic Admixture on the Island of Puerto Rico. Relethford J, editor. PLoS ONE. 2011 Jan 31;6(1):e16513.

11. Marnetto D, Pärna K, Läll K, Molinaro L, Montinaro F, Haller T, et al. Ancestry deconvolution and partial polygenic score can improve susceptibility predictions in recently admixed individuals. Nat Commun. 2020 Apr 2;11(1):1628.

12. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet. 2019 Apr;51(4):584–91.

13. Mathieson I, Scally A. What is ancestry? Flint J, editor. PLOS Genet. 2020 Mar 9;16(3):e1008624.

14. Popejoy AB, Fullerton SM. Genomics is failing on diversity. Nature. 2016 Oct 13;538(7624):161–4.

15. Cerdeña JP, Grubbs V, Non AL. Genomic supremacy: the harm of conflating genetic ancestry and race. Hum Genomics. 2022 May 19;16(1):18.

16. Gouveia MH, Meeks KAC, Borda V, Leal TP, Kehdy FSG, Mogire R, et al. Subcontinental genetic variation in the All of Us Research Program: Implications for biomedical research. Am J Hum Genet. 2025 Jun 5;112(6):1286–301.

17. Grundler MC, Terhorst J, Bradburd GS. A geographic history of human genetic ancestry. Science. 2025 Mar 28;387(6741):1391–7.

18. Lewis ACF, Molina SJ, Appelbaum PS, Dauda B, Di Rienzo A, Fuentes A, et al. Getting genetic ancestry right for science and society. Science. 2022 Apr 15;376(6590):250–2.

19. Greater Middle East Variome Consortium, Scott EM, Halees A, Itan Y, Spencer EG, He Y, et al. Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery. Nat Genet. 2016 Sep;48(9):1071–6.

20. Nievergelt CM, Maihofer AX, Shekhtman T, Libiger O, Wang X, Kidd KK, et al. Inference of human continental origin and admixture proportions using a highly discriminative ancestry informative 41-SNP panel. Investig Genet. 2013;4(1):13.

21. Ellison LK, Ng C, Kewayosh A, Troke N, Elias B, Tjepkema M, et al. Ancestry or identity? The importance of Indigenous engagement in articulating First Nations, Inuit and Métis peoples in the 2006 Canadian Census Health and Environmental Cohort (CanCHEC). Int J Circumpolar Health. 2025 Dec;84(1):2495378.

22. Moreau C, Vézina H, Jomphe M, Lavoie ÈM, Roy-Gagnon MH, Labuda D. When Genetics and Genealogies Tell Different Stories-Maternal Lineages in Gaspesia: Amerindian Admixture in New France. Ann Hum Genet. 2010 Nov;no-no.

23. Verdu P, Pemberton TJ, Laurent R, Kemp BM, Gonzalez-Oliver A, Gorodezky C, et al. Patterns of Admixture and Population Structure in Native Populations of Northwest North America. Long JC, editor. PLoS Genet. 2014 Aug 14;10(8):e1004530.

Tables

Table 1. Concordance rate by self-declared ancestry category. Concordance is defined as the proportion of individuals whose genome-derived global ancestry (GAI, assigned based on their largest local ancestry fraction) matches the expected ancestry for their self-declared identity category. The five continental-level population groups include AFR (African), AMR (Admixed American), EAS (East Asian), EUR (European), and SAS (South Asian).

Self-Declared Category	GAI	n	Concordant (n)	Concordance Rate
Ashkenazi Jewish	EUR	62	62	1.000
Sephardic Jewish	EUR	2	2	1.000
White	EUR	2,743	2,709	0.988
Black	AFR	213	207	0.972
East Asian / Pacific Islander	EAS	358	344	0.961
South Asian	SAS	358	322	0.899
Hispanic	AMR	185	138	0.746
Middle Eastern or Central Asian	EUR	162	110	0.679
Indigenous (First Nations, Métis, Inuit)	AMR	27	11	0.407
Not Disclosed	NA	4,723	NA	NA
Prefer not to answer / Do not know	NA	1,168	NA	NA
More than one race	NA	171	NA	NA
Other	NA	78	NA	NA

Table 2. Cohen's Kappa agreement (Interrater agreement). The five continental-level population groups include AFR (African), AMR (Admixed American), EAS (East Asian), EUR (European), and SAS (South Asian).

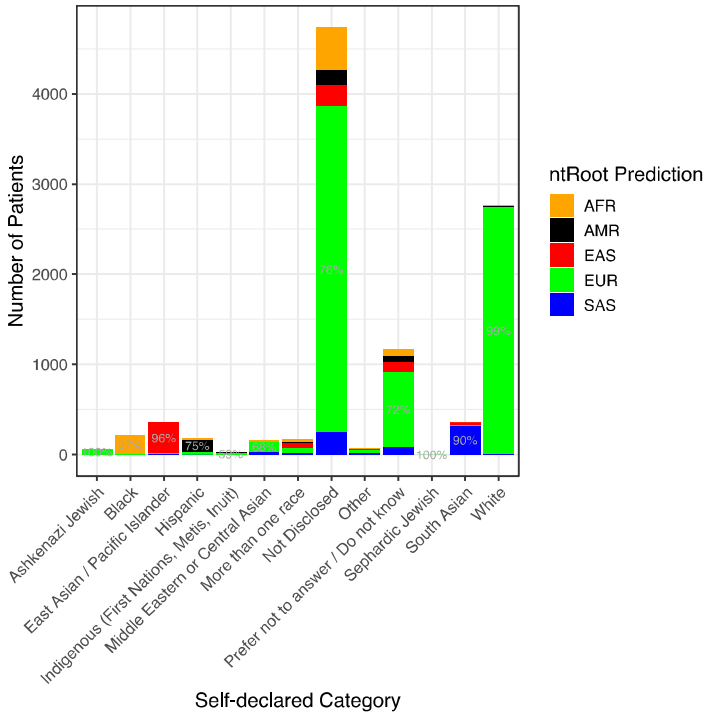
Inferred vs. Expected	**AFR**	**AMR**	**EAS**	**EUR**	**SAS**
AFR	1.000	0.500	0.015	0.400	0.250
AMR	0.000	1.000	0.175	0.688	0.100
EAS	0.000	0.000	1.000	0.312	0.710
EUR	0.000	0.000	0.000	1.000	0.390
SAS	-0.029	0.000	-0.029	-0.059	1.000

Table 3. Shannon entropy and admixture by self-declared ancestry category. Mean entropy represents the average genomic admixture complexity for individuals within each self-declared category, based on the distribution of genome-wide ancestry fractions. Standard deviation (SD) reflects the variability of entropy within each group. Higher entropy indicates greater heterogeneity in continental ancestry proportions, and thus more admixture.

Self-Declared Category	n	Mean Entropy	SD Entropy
Sephardic Jewish	2	2.04	0.045
Middle Eastern or Central Asian	162	2.00	0.096
Ashkenazi Jewish	62	1.91	0.041
Indigenous (First Nations, Métis, Inuit)	27	1.78	0.149
White	2,743	1.72	0.107
More than one race	171	1.72	0.470
Hispanic	185	1.71	0.186
Other	78	1.65	0.399
Not Disclosed	4,723	1.58	0.453
Prefer not to answer / Do not know	1,168	1.53	0.467
South Asian	358	1.36	0.334
Black	213	0.47	0.465
East Asian / Pacific Islander	358	0.47	0.284

Figures

A



B

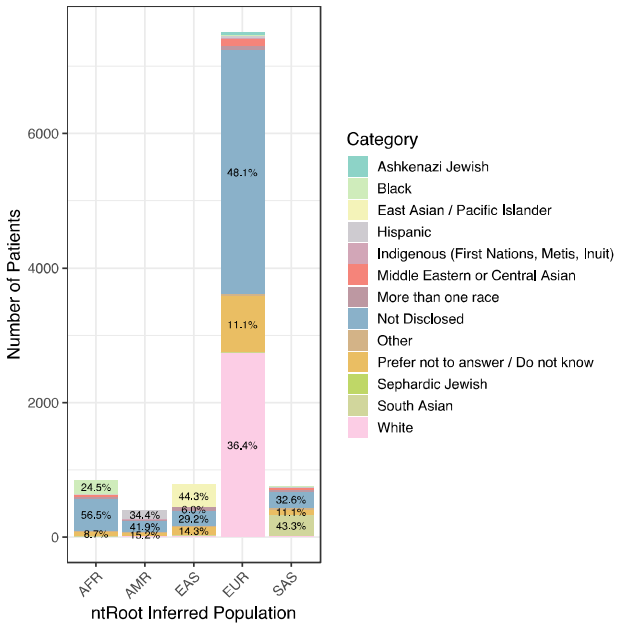


Fig. 1. Inferred versus self-declared ancestry in the HostSeq cohort. **A)** by ntRoot predictions of the main continental ancestry and **B)** by self-declared ancestry categories in the HostSeq study.

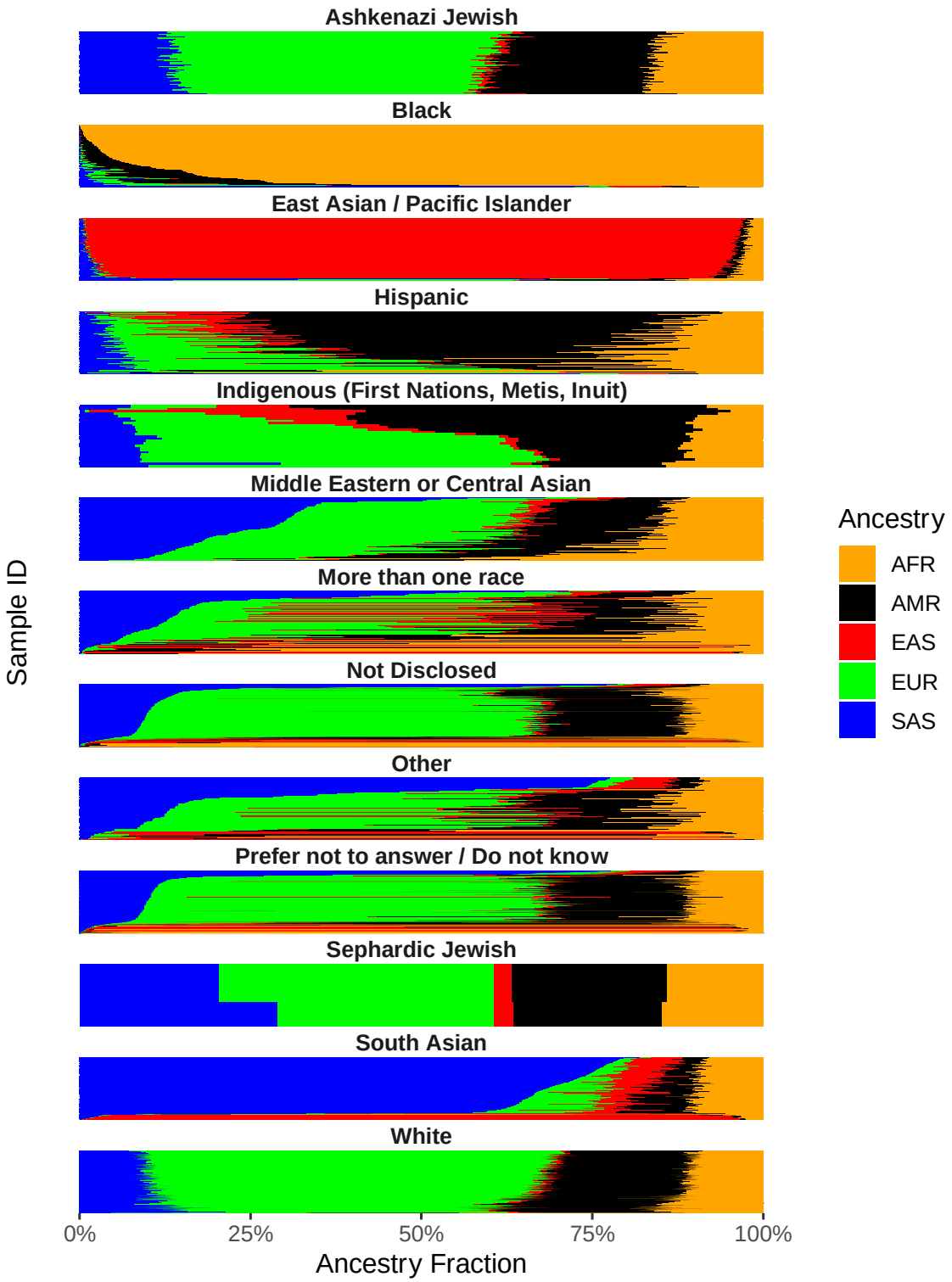


Fig. 2. Predicted ancestry fractions based on LAI (x-axis), for each HostSeq individual sample (y-axis), organized by self-declared identity category.