Michigan Genomics Initiative Data Freeze 4 Local Ancestry Inferences

Brett Vanderwerff[1,2,*], Lars Fritsche[1,2], Anita Pandit[1,2], Snehal Patil[1,2,3], Matthew Zawistowski[1,2], Michael Boehnke[1,2], Xiang Zhou[1,2], and Sebastian Zöllner[1,2,4]

[1] *Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI 48109, USA*
[2] *Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor, MI 48109, USA*
[3] *Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA*
[4] *Department of Psychiatry, University of Michigan School of Medicine, Ann Arbor, MI 48109, USA*
[*] *To whom correspondence regarding data preparation should be addressed:* brettva@umich.edu

November 17, 2021

1 Data Description

Local ancestry inference (LAI) estimates the genetic ancestry of tracts of DNA in target samples of unknown ancestry by comparing genotypes to a set of reference samples where ancestry is known.

We offer LAI for 60,215 MGI participants included in Data Freeze 4, deconvoluting DNA tracts for European, African, East Asian, Central/South Asian, West Asian, Native American, and Oceanian ancestries. We generate these data by comparing genotypes of MGI participants with samples and super-population labels of the Human Genome Diversity Panel (HGDP) [1]. Our summary of majority global ancestry from LAI suggests Data Freeze 4 contains 53,272 European, 3,798 African, 1,313 East Asian, 852 Central/South Asian, 557 West Asian, and 423 Native American participants (Figure 1).

These LAI data are mapped to the coordinates of build 38 and contained across 3 tab delimited text files described in Table 1.

2 Methods

We described the production and quality control of genotype data and ADMIXTURE- and PCA-based genetic ancestry inference for MGI participants previously [2].

File	Contents
MGI_DataFreeze4.LAI.msp.tsv	Most likely reference population of origin per chromosomal region
MGI_DataFreeze4.LAI.fb.tsv	Probability of reference population assignment per chromosomal region
MGI_DataFreeze4.LAI.sites.txt	Genotype sites included in the LAI analysis

Table 1: Description of files containing LAI data for participants included in Data Freeze 4.

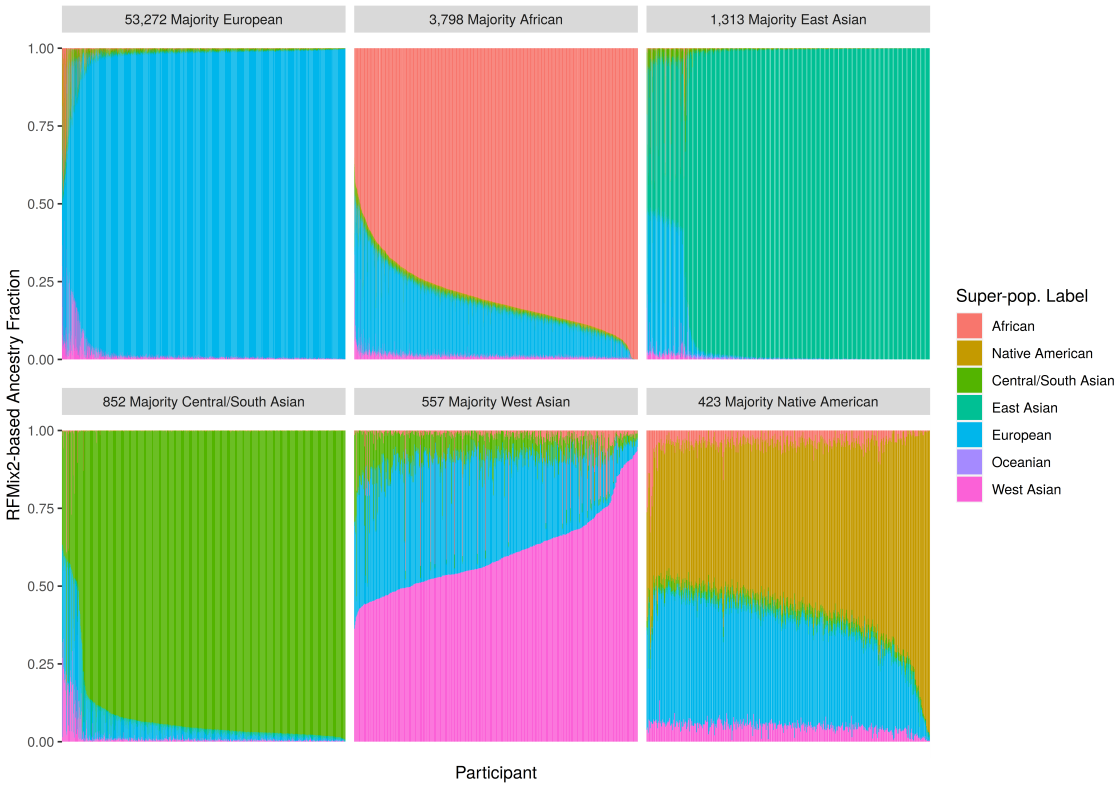

Figure 1: RFMix2-based ancestry of MGI participants. Global ancestry for MGI participants summarized from local ancestry estimated by RFMix2. Each inset is a stacked barplot with a bar for each participant belonging to the respective majority ancestry super-population.

Super-population label	Regional population labels
Africa (100)	Bantu S.E. S.Sotho, Bantu N.E., Bantu S.E. Zulu, Mandenka, Yoruba, San, Bantu S.E. Tswana, Bantu S.W. Herero, Mbuti Pygmies, Bantu S.E. Pedi, Bantu S.W. Ovambo, Biaka Pygmies
Central/South Asia (195)	Balochi, Uygur, Sindhi, Pathan, Burusho, Hazara, Kalash, Makrani, Brahui
East Asia (221)	Lahu, Daur, Japanese, Mongola, She, Tu, Oroqen, Xibo, Naxi, Yakut, Tujia, Cambodians, Miaozu, Hezhen, Han, Yizu, Dai
Europe (153)	Russian, Orcadian, French Basque, French, Tuscan, Adygei, Sardinian, North Italian
Native America (60)	Pima, Maya, Colombians, Surui, Karitiana
Oceania (27)	NAN Melanesian, Papuan
West Asia (158)	Palestinian, Bedouin, Druze, Mozabite

Table 2: HGDP super-population and regional population labels. The regional population labels that comprise each super-population label according to mappings obtained from the Foundation Jean Dausset-CEPH. The number of HGDP reference samples belonging to each super-population label is given in parenthesis.

We prepare a LAI reference panel from a whole genome sequence-based call-set of single nucleotide variants and short insertion-deletions from 914 unrelated HGDP samples that we access from ftp://ngs.sanger.ac.uk/production/hgdp [3]. We filter the HGDP call-set to exclude sites with minor allele count < 2, call-rate $\leq 99\%$, or exact test of Hardy-Weinberg Equilibrium $p \leq 10^{-6}$ before phasing with a call-set of single nucleotide variants and short insertion-deletions from 2,504 1000 Genomes Project samples with Beagle (v5.2) [4, 5].

We assign African, Native American, Central/South Asian, East Asian, European, Oceanian, or West Asian super-population labels to each HGDP sample according to mappings available from the Foundation Jean Dausset-CEPH @ https://cephb.fr/en/hgdp_panel.php. We provide a summary of regional population labels that are grouped to super-population labels by this consolidation in Table 2.

We estimate local ancestry for each MGI participant using RFMix2 (v2.03-r0) [6]. We use a HapMap genetic map accessed from http://bochet.gcc.biostat.washington.edu/beagle/genetic_maps/ and set the -n flag (terminal node size for random forest trees) to 5. We perform a separate RFMix2 run for each autosome.

We estimate the fraction of each MGI participant's genome that originates from each ancestral super-population grouping by summing the number of genetic sites assigned to each super-population by RFMix2 and dividing by 441,410 (2x the number of genetic sites that intersect the HGDP and MGI samples). We define the RFMix2-based majority ancestry of MGI participants as the largest super-population fraction determined by this approach.

We generate a truth set of samples with known local ancestry and phase to estimate LAI accuracy of MGI participants. We simulate 3-way admixture between a founder population of 150 MGI participants inferred European, African, or East Asian by PCA (450 total) using admixture-simulation [7]. We simulate 100 admixed progeny from 8 generations of random-mating between founders. We then infer local ancestry for progeny with RFMix2 using 914 HGDP samples and super-population labels for European, African, East Asian, Central/South Asian, West Asian, Native American, and Oceanian ancestries as reference. We define the LAI concordance rate as the percentage of sites where the inferred ancestry by RFMix2 on simulated data agrees with the truth value out of the total number of sites evaluated. We evaluate LAI concordance rate separately for all sites and just those sites where the truth call or inference is assigned to European, African, or East Asian ancestry.

Majority Ancestry Group	RFMix2 & ADMIXTURE	RFMix2 Only	ADMIXTURE Only
European	52,976	296	78
African	3,760	38	1
East Asian	1,281	32	0
Central/South Asian	849	3	42
West Asian	557	0	223
Native American	412	11	36

Table 3: Comparison of RFMix2- and ADMIXTURE-based majority ancestry labels. The number of MGI participants inferred majority European, African, East Asian, Central/South Asian, West Asian, or Native American by both RFMix2 and ADMIXTURE are given in addition the number of participants inferred to belong to each ancestry group uniquely by either method.

3 Data Quality Evaluation

We compared RFMix2- to PCA-based ancestry inference for 150 MGI participants inferred European, African, or East Asian by PCA (Figure 2). For each MGI participant, the majority ancestry label determined from RFMix2 output was consistent with the PCA-based label. The RFMix2-based ancestry fraction that corresponded to the PCA-based label was above 93% in all expect 6 inferred European ancestry participants where RFMix2 reported increased West Asian ancestry.

We compared RFMix2- to ADMIXTURE-based ancestry for every participant included in Data Freeze 4 (Figure 3). The square of the Pearson correlation coefficient between RFMix2 and ADMIXTURE results was highest for African (.999) and lowest for West Asian (.869) and European (.976) ancestry. We note that in our analysis, ADMIXTURE reports increased levels of West Asian ancestry in MGI participants relative to RFMix2 (data from Oceanian ancestry are not shown; currently Oceanian ancestry is inferred to be present only in trace amounts in the MGI cohort).

In some contrast to our RFMix2 results, majority global ancestry from ADMIXTURE-based analysis suggests Data Freeze 4 contains 53,054 European, 3,761 African, 1,281 East Asian, 891 Central/South Asian, 780 West Asian, and 448 Native American participants (Figure S1). We evaluated the agreement between majority global ancestry labels gotten from RFMix2- and ADMIXTURE-based ancestry inference (Table 3). Disagreement was relatively low for most super-population labels we evaluated with the notable exception that an additional 229 MGI participants were inferred majority West Asian by ADMIXTURE but not RFMix2.

The LAI accuracy as measured by concordance rate between truth and inferred calls at all sites from simulated admixed individuals was 95.57%. We observed a concordance rate of 95.40% for East Asian, 94.07% for African, and 89.33% for European sites.

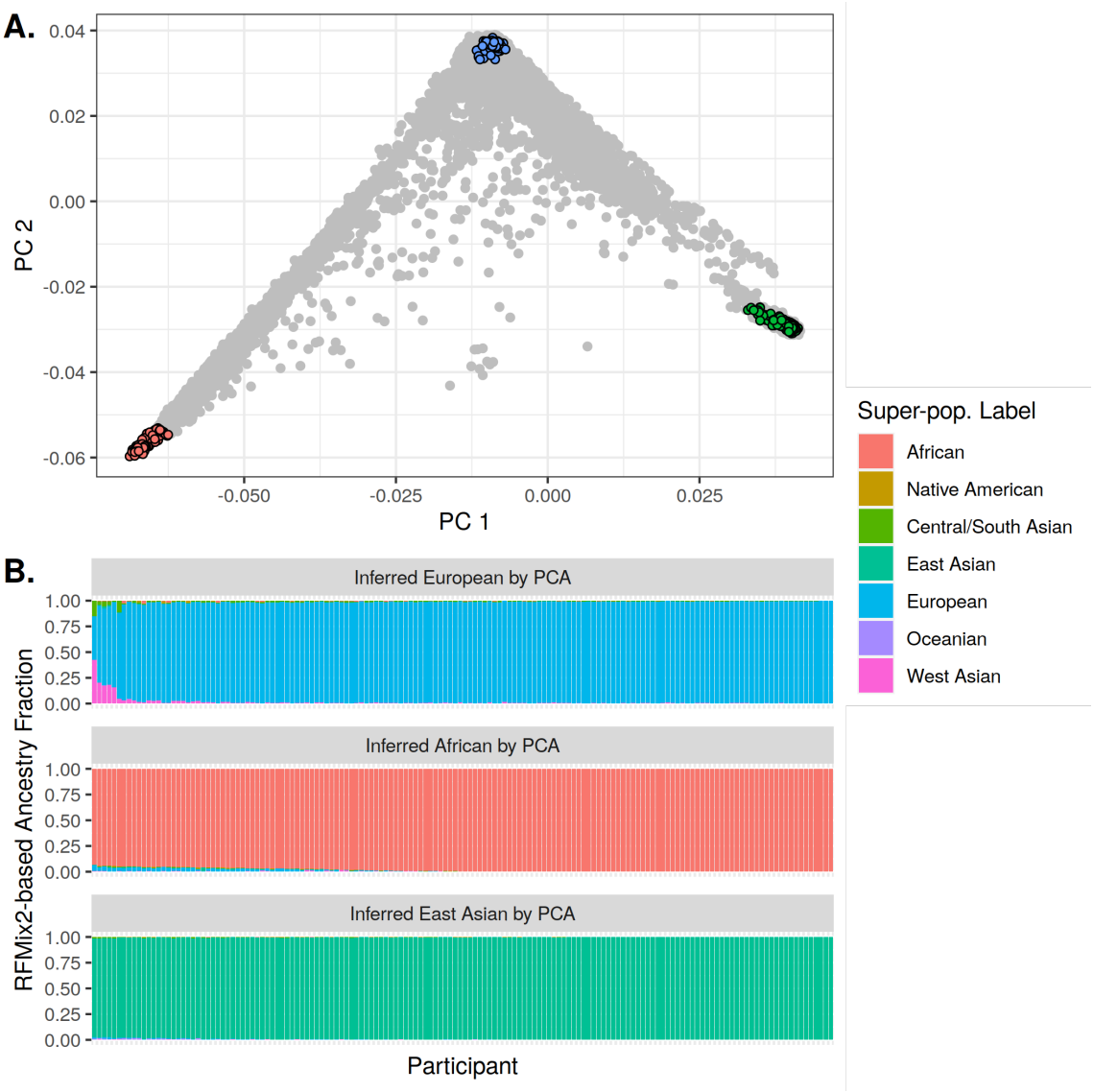

Figure 2: Comparison of RFMix2- and PCA-based ancestry inference. **(A.)** MGI samples projected on the first two principal components of a sample from the Human Genome Diversity Panel. 150 MGI samples inferred African, European, or East Asian (450 total) by PCA are colored with the remaining MGI sample shown in grey. **(B.)** RFMix2-based ancestry for 150 MGI participants inferred European (top panel), African (middle panel), or East Asian (bottom panel) by PCA. Each panel is a stacked barplot with a bar for each participant we compared for RFMix2- and PCA-based ancestry inference results.

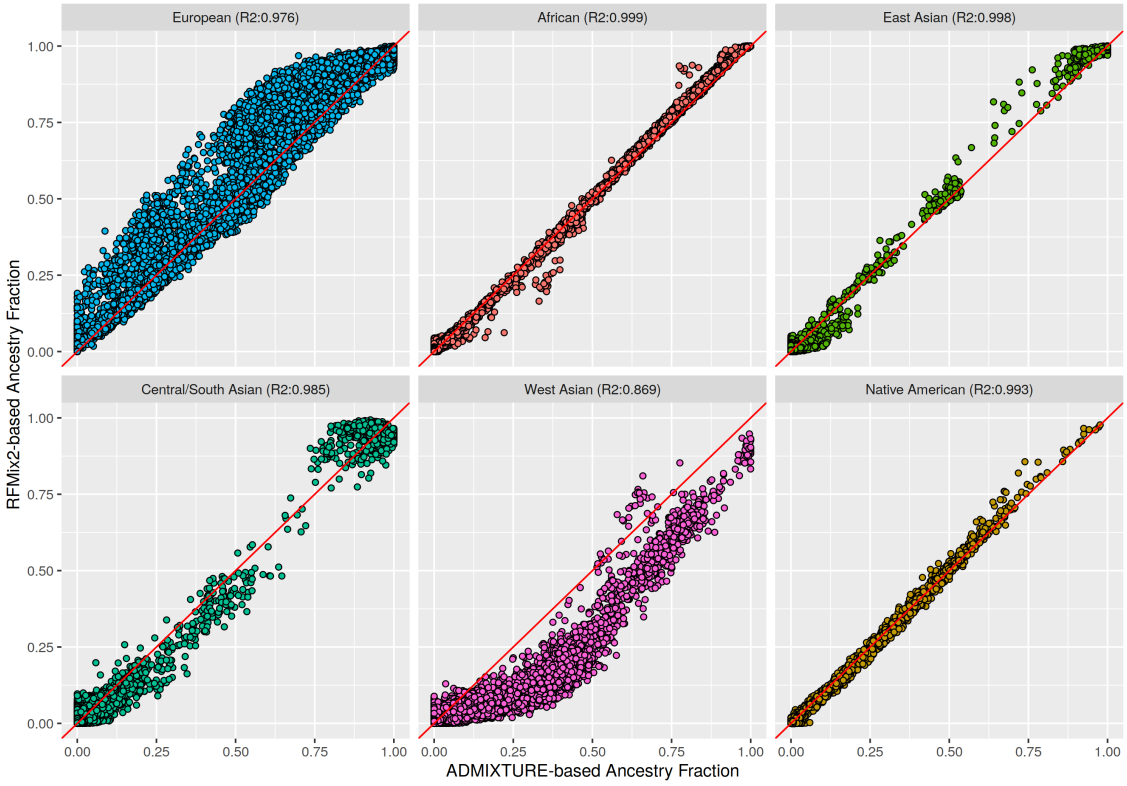

Figure 3: Comparison of RFMix2- and ADMIXTURE-based ancestry inference. Global ancestry summarized from local ancestry estimated by RFMix2 compared to that reported by ADMIXTURE. Each inset contains 60,215 points (one for each MGI participant included in Data Freeze 4). R2, square of the Pearson correlation coefficient of ADMIXTURE and RFMix2 global ancestry. Diagonal red line indicates y=x.

References

[1] Cann, H. M. *et al.* A human genome diversity cell line panel. *Science (New York, N.Y.)* **296**, 261–262 (2002).

[2] Vanderwerff, B. *et al.* Michigan Genomics Initiative Data Freeze 4 Technical Notes. https://precisionhealth.umich.edu/wp-content/uploads/sites/67/2021/08/data_freeze4_tech_notes.pdf.

[3] Bergström, A. *et al.* Insights into human genetic variation and population history from 929 diverse genomes. *Science* (2020).

[4] Byrska-Bishop, M. *et al.* High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios (2021).

[5] Browning, S. R. & Browning, B. L. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. *The American Journal of Human Genetics* **81**, 1084–1097 (2007).

[6] Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. *American Journal of Human Genetics* **93**, 278–288 (2013).

[7] Wright, M. K. Admixture simulation tool. https://github.com/slowkoni/admixture-simulation (2021).

4 Supplementary Figures

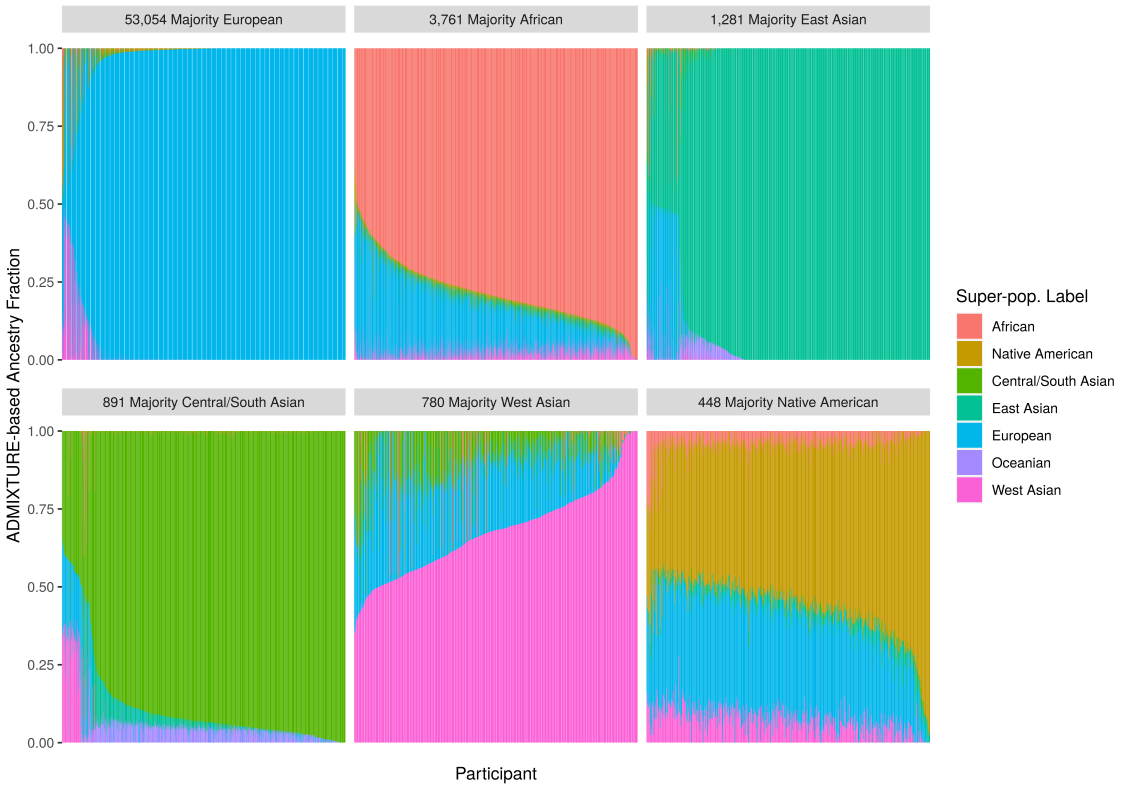

Figure S1: ADMIXTURE-based ancestry of MGI participants. Global ancestry estimated by ADMIXTURE for MGI participants. Global ancestry is defined as the super-population label with the largest reported Q value from ADMIXTURE. Each inset is a stacked barplot with a bar for each participant belonging to the respective majority ancestry population.