Improved Allele Frequencies in gnomAD through Local Ancestry Inference

Pragati Kore[1*], Michael W. Wilson[2*], Grace Tiao[2], Katherine Chao[2], Phil Darnowsky[2], Nick Watts[2], Jessica Honorato Mauer[1], Samantha M. Baxter[2], Genome Aggregation Database Consortium, Heidi L Rehm[2,5], Mark J. Daly[2,3,4,6], Konrad J. Karczewski[2,3,4,5], Elizabeth Atkinson[1,3,4,7,†]

1. Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, 77030, USA.
2. Program in Medical and Population Genetics, The Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA.
3. Stanley Center for Psychiatric Research, The Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA.
4. Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA.
5. Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA 02114, USA.
6. Institute for Molecular Medicine Finland, Helsinki, Finland.
7. Jan and Dan Duncan Neurological Research Institute, Texas Children's Hospital, Houston, TX, 77030, USA.

*These authors contributed equally to this work.

[†]**Corresponding author contact information**:

Baylor College of Medicine

One Baylor Plaza, Taub Research Building Room 621

Houston, TX, USA 77019

Email: Elizabeth.atkinson@bcm.edu

Keywords: allele frequency, reference panel, admixture, ancestry

Abstract

The Genome Aggregation Database (gnomAD) has been an invaluable resource for the genomics community[1,2], quantifying allele frequencies across multiple genetic ancestry groups. However, traditional gnomAD allele frequencies report aggregate estimates, overlooking fine-scale genetic differences. This can skew allele frequency estimates, particularly in recently admixed populations[3], and potentially lead to clinically relevant misinterpretations in underrepresented groups[4]. To address this, we have conducted local ancestry inference (LAI) in two gnomAD genetic ancestry groups with high degrees of recent admixture – "Admixed American" (n=7,612) and "African/African American" (n=20,250) – to derive more precise genetic ancestry-specific allele frequencies, enhancing the granularity and accuracy of frequency estimates. We demonstrate that LAI enables identification of important disease-risk variants that occur at higher frequencies within specific genetic ancestry components[5]. Additionally, we found that 85.1% of variants in the African/African American and 78.5% in the Admixed American genetic ancestry groups show at least a twofold difference in ancestry-specific allele frequencies and 81.49% of sites analyzed would receive an increased grpmax value. With the release of this new LAI data, we provide a

higher-resolution view of allele frequency across specific ancestry backgrounds, which is obscured when assessing frequency at the aggregate group level, enabling more precise and accurate genomic interpretations.

Introduction

The Genome Aggregation Database (gnomAD) is one of the most widely used resources in the genomics community[2], offering comprehensive data on allele frequencies across diverse genetic ancestry groups. With its vast catalog of genetic variation derived from hundreds of thousands of individuals, gnomAD has become an invaluable tool for researchers and clinicians, aiding in the interpretation of genetic variants and their potential implications for human health[6]. By providing allele frequency data across diverse genetic ancestry groups, gnomAD helps in the interpretation of the impact of rare and common genetic variants, making it a cornerstone of modern genetic research. Allele frequencies are a critical component of genomics research and clinical genetics, serving as a key metric for understanding the distribution of genetic variants within and between genetic ancestry groups[7]. These frequencies are used to assess the pathogenicity of variants, guide the design of genetic data generation platforms, and inform the development of therapeutic strategies. In particular, allele frequencies stratified by genetic ancestry groups are essential for identifying genetic risks that are enriched in particular groups and understanding the genetic architecture of complex traits. As such, gnomAD's genetic ancestry group allele frequencies have been pivotal in advancing our understanding of human genetic variation and its impact on health and disease.

Despite the utility of gnomAD's allele frequencies, there are limitations inherent in the way these frequencies are computed and how genetic ancestry is represented. While genetic ancestry is truly continuous[8], the genetic ancestry groups used in gnomAD's aggregations are by nature discrete. Specifically, these groups are defined by majority continental genetic ancestry as trained on reference-panel informed principal components[2], which can obscure finer-scale genetic

differences observed across individuals within a genetic ancestry group[9–12]. This limitation is particularly important when considering groups that have high amounts of genetic diversity or those with recent admixture events including the gnomAD groups currently identified as 'Admixed Americans' and 'African/African Americans'. Additionally, the use of these discrete groupings does not capture the full spectrum of an individual's genetic background, which can span multiple continental ancestries.

Local ancestry inference (LAI) offers a solution to some of these limitations by deconvolving allele frequencies based on the specific ancestry of chromosomal segments[13]. By applying LAI, we resolve finer-scale allele frequencies, allowing for better-informed clinical interpretations. Here, we enhance the granularity of gnomAD's allele frequency data in the inferred Admixed American and African/African American genetic ancestry groups, providing a more accurate representation of genetic variation and its implications for human health. Our computational pipeline, written in Hail Batch (see *Methods*), is publicly available via GitHub. The allele frequencies and quality metrics can be downloaded or interactively explored on the gnomAD browser (https://gnomad.broadinstitute.org/).

In all analyses, we exclusively examine genetic ancestry. Assignment of an individual's genetic ancestry is not equivalent to and does not negate how an individual self-identifies. To refer to the samples included in these analyses, we adhere to the current nomenclature in gnomAD, namely "Admixed American" and "African/African American". We wish to acknowledge that this terminology is not ideal, however for consistency with broader efforts utilizing these data, and to be clear on the individuals included in these analyses, we retain it here. We look forward to

continuing to further refine gnomAD allele frequency presentations in future releases to better reflect the continuous nature of genetic ancestry[9,10].

Results

Here we release local ancestry-informed frequency data for 14,804,207 bi-allelic single nucleotide polymorphisms (SNPs) within the Admixed American (n=7,612) and 24,204,574 bi-allelic SNPs within the African/African American genetic ancestry groups (n=20,250), both from gnomAD v3.1. This initial release of gnomAD LAI data contains variants' estimated alternate allele counts, allele numbers, and allele frequencies partitioned by haplotype-level continental ancestries. These new local ancestry-informed frequencies are released in a downloadable VCF format file and can be explored interactively on the gnomAD browser (https://gnomad.broadinstitute.org/). To access the local ancestry data for a specific variant within the browser, select the "Local Ancestry" tab within the genetic ancestry group frequency table on the variant's page. If local ancestry data is available for the variant, the ancestral allele count (AC), allele number (AN), and allele frequency (AF) estimates will be displayed **(Supplementary Figure 1**).

Plotting the first two PCs of the full gnomAD collection, we can see the inferred gnomAD Admixed Americans span a large portion of global PC space, with some samples clustering closer to European reference populations, others clustering closer to the African reference samples, and some clustering closer to the East Asian reference individuals, indexing the presence of Amerindigenous ancestry (**Figure 1a**). Resolving global ancestry with ADMIXTURE, we again observe that samples classified as Admixed American by gnomAD exhibit heterogeneous ancestry, with individuals containing differing proportions of two or three-way admixture between

Amerindigenous, African, and European haplotypes. After applying a 5% ancestry inclusion threshold, 5% of individuals derive their genetic ancestry primarily from a single continent (homogeneous AMR), 60% from two continents (AMR/EUR), and 35% from three continents (AMR/EUR/AFR) (**Figure 1b)**. Given this composition, we modeled the Admixed American samples with a three-way reference panel (see *Methods*). The African/African American samples primarily lie on a cline between the EUR and AFR poles in PC space, indicating admixture between largely these two genetic ancestry components. Further exploring African/African American samples' ancestry with ADMIXTURE, we again observed individuals in this group contain primarily African and European components, with 32% homogeneous AFR individuals present and 68% of samples classified as two-way admixed (**Figure 1c**). Therefore, we curated a reference panel comprising two-way continental African and European genetic ancestries for this group.

Local ancestry deconvolution provides additional resolution beyond global inference; we observe that even individuals occupying the same point in PC space have markedly different local ancestry patterns (**Figure 1a)**. By only calculating aggregate rather than ancestry-specific allele frequencies in admixed groups, estimates represent a weighted average and can be distorted from the true ancestry-specific frequencies. This additional resolution can be meaningful for variants of functional importance.

For example, the gnomAD variant 17-7043011-C-T in *SLC16A11*, associated with type 2 diabetes risk, is elevated within the Latino/Admixed American population[14] at a 14% allele frequency compared to the global gnomAD allele frequency of 2% (**Figure 2a**). Breaking down this variant's

frequency into continental allele frequency estimates, we see in fact that the variant occurs at a 45% allele frequency in continental American regions and closer to 0% in the African and European regions. Similarly, the gnomAD variant 22-36265860-A-G in *APOL1*, linked with increased susceptibility to focal segmental glomerulosclerosis (FSGS), HIV-associated nephropathy (HIVAN), and hypertensive end-stage kidney disease (ESKD)[15], shows a 27% allele frequency in African regions of the African American genetic ancestry group[16], compared to a 1% gnomAD-wide ("global") frequency (**Figure 2b**).

LAI is useful in resolving allele frequencies for rare variants as well as common. Variants in the *COL27A1* gene have been shown to cause Steel syndrome and suggest a founder mutation effect in Puerto Rico[17]. Within gnomAD, this causal variant has 0.3% allele frequency within the Admixed American genetic ancestry group while gnomAD's global allele frequency is 0.03%. Separating this variant's gnomAD AMR frequency into continental allele frequency estimates, we see that the variant occurs almost at a 1% allele frequency in Amerindigenous regions and closer to 0% in African and European regions, reinforcing a founder effect hypothesis and again showing the usefulness of local ancestry in large genomic datasets. Another example is the gnomAD variant 7-117606754-G-A in *CFTR*, associated with cystic fibrosis[18]. This variant shows a 0.1% allele frequency in African regions of the African American genetic ancestry group, compared to a 0.007% gnomAD-wide ("global") frequency. These variant examples highlight the importance of more precise allele frequencies in accurately identifying the distribution of genetic risk factors within diverse genetic ancestry groups.

While many variants are similar in their ancestry-specific allele frequency estimates, many are also highly divergent (**Supplementary Figure 2)**. The overall correlation in ancestry-specific allele frequencies between the AFR and EUR tracts in the African/African American group was 0.83. For the Admixed American group, the correlations were 0.80 between AFR and EUR AFs, 0.73 between AFR and AMR AFs, and 0.84 between AMR and EUR AFs. In the African/African American genetic ancestry group, 85.1% of variants with LAI data exhibit at least a twofold difference between ancestry-specific AFs (**Figure 3a**). The majority of these variants (69%) have an African ancestry AF that is at least two times greater than the European ancestry AF, while 16.1% show the opposite pattern. Similarly, in the Admixed American genetic ancestry group, 78.5% of variants demonstrate this level of divergence between ancestry-specific AFs (**Figure 3b**). Among these, 41.6% of variants have an African ancestry AF that is two times greater than those of either other ancestry tracts (European and Amerindigenous), while the remaining variants show a dominant European AF (27.5%) or Amerindigenous AF (9.4%). African ancestry tracts consistently exhibit higher AFs compared to European or Amerindigenous tracts, both when sites that are monomorphic in the other ancestries are considered but also when excluding such loci.

Visualizing the site frequency spectrum **(Figure 4)** highlights the allele frequency distributions across two genetic ancestry groups—African/African American and Admixed American—classified by their ClinVar annotations[19]. The results indicate that most variants across all ClinVar categories, including pathogenic/likely pathogenic, benign/likely benign, uncertain significance, and conflicting classifications of pathogenicity are rare, with AFs concentrated between 0 and 0.1 in all genetic ancestry groups. The broader distribution of benign variants in African/African American groups reinforces the need to include diverse ancestry groups in genomic studies, as it

demonstrates how genetic diversity impacts allele frequency distributions and, by extension, clinical variant interpretation.

We additionally investigated the frequency of ancestry-enriched AFs exceeding those of grpmax (rebranded from popmax)[20], revealing significant insights into local ancestry influences. We found that across sites for both genetic ancestry groups, 22,337,592 sites exhibited a new LAI-based AF that was higher than the existing grpmax, representing 81.49% of the total 27,412,225 sites with available data. Specifically, 74.45% had tract AFs surpassing grpmax for the African/African American group, while the Admixed American group exhibited 64.95% exceeding grpmax. We also compared the tract AFs and grpmax AFs in each genetic ancestry group based on ClinVar classifications (**Supplementary Table 1)**. We found that about 87.69% and 89.32% of ClinVar variants with a higher tract AF than grpmax AF were classified as 'Benign/Likely Benign' in the African/African American and Admixed American genetic ancestry group, respectively. These results highlight the substantial influence of local ancestry-informed AFs on the grpmax comparisons, indicating that ancestry-enriched variants are frequently underestimated when relying solely on grpmax metrics. Using grpmax AF's without consideration of LAI could lead to inaccurate variant classification in clinical studies, particularly when evaluating the pathogenicity of variants in underrepresented populations.

In sum, we anticipate this finer resolution of allele frequencies thanks to local ancestry deconvolution will improve the interpretation of variation by contextualizing variants to their approximate geographical origin and allowing for more accurate allele frequencies to be calculated per genetic ancestry background. This improved resolution can lead to more effective

identification of functionally important risk variants and a more accurate spectrum of allele frequencies. These refined estimates are freely downloadable and integrated into the widely used gnomAD browser for improved interpretation of allele frequencies across continentally defined genetic ancestries.

Discussion

An outstanding question in genomics, with far-reaching clinical and research implications, is the degree of concordance in allele frequencies across different ancestral backgrounds. Allele frequencies vary dramatically across the globe[21,22] due to features such as demography, genetic drift, and natural selection. Deconvolving from which continental ancestry a variant is derived allows us to compute more accurate expected allele frequencies for each continental background, instead of aggregating frequencies across recently admixed individuals regardless of the ancestral origin of the variant's haplotype[23]. The large sample size of gnomAD enables us to examine ancestral allele frequency trends at an unprecedented scale, and with the integration of LAI, at an unprecedented resolution. In some cases, updated allele frequencies that take into account local ancestry differ significantly from their corresponding aggregate allele frequencies, which may impact the clinical interpretation of a variant (**Supplementary Figure 1**).

Our analyses further reveals that a substantial number of variants exhibit highly divergent allele frequencies across ancestry backgrounds: within the African/African American and Admixed American ancestry groups, 85.1% and 78.5% of variants with LAI data, respectively, show at least a twofold difference between ancestry-specific frequencies. Additionally, we found that across both ancestry groups, 81.49% of the 27,412,225 sites analyzed would receive an increased grpmax

value (previously popmax) when using LAI data. When examining these variants in ClinVar, approximately 88% of those with LAI-inferred frequencies higher than grpmax were classified as 'Benign/Likely Benign'. This suggests that many variants, correctly classified as benign using ancestry-informed data, might be underestimated or misclassified when relying solely on global grpmax estimates.

Ancestry-specific allele frequencies can additionally help refine genetic prevalence estimates for recessive diseases. Understanding the estimated proportion of a population that has a causal genotype for a genetic disorder, and whether there is a significantly higher risk in a specific genetic ancestry group or not, is a key question for understanding the global impact of a disease. Local ancestry has the potential to refine these estimates further and provide the community with a more accurate prediction of how many people have a given disease. While a 0.1% frequency cut-off is still more common than most pathogenic disease-causing variants, in the future local ancestry will be able to be applied to more rare variants, which will have an even greater effect on clinical care.

Local ancestry resolved variant allele frequencies help in interpretation of GWAS results[5,9]. Differentiating between aggregate allele frequencies and those enriched in specific genetic ancestry groups can streamline the search for disease-associated variants with elevated importance for particular genetic ancestries[24]. For example, we demonstrate that the gnomAD variant 17-7043011-C-T in *SLC16A11*, associated with type 2 diabetes risk, is observed at a 14% frequency in the Latino/Admixed American population compared to 2% globally. When partitioned by continental ancestry, this variant occurs at a 45% frequency in Amerindigenous regions of admixed genomes, while it is nearly absent in African and European regions. This, and the other vignettes

highlighted, showcases how local ancestry inference resolves frequency discrepancies in both rare and common key variants.

This release marks the first time local ancestry has been systematically incorporated into a globally-representative resource like gnomAD, setting a new standard for genomic inclusion. This release of local ancestry-informed frequency data is currently available for the Admixed American and African/African American genetic ancestry groups. We hope this data can be used to enhance the scientific community's ability to clinically interpret variation and better identify those populations at highest risk for certain diseases.

Main References

1. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature 2020 581:7809* **581**, 434–443 (2020).
2. Chen, S. *et al.* A genomic mutational constraint map using variation in 76,156 human genomes. *Nature 2023 625:7993* **625**, 92–100 (2023).
3. Martin, E. R. *et al.* Properties of Global and Local Ancestry Adjustments in Genetic Association Tests in Admixed Populations. *Genet Epidemiol* **42**, 214 (2017).
4. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics 2019 51:4* **51**, 584–591 (2019).
5. Tan, T. & Atkinson, E. G. Strategies for the Genomic Analysis of Admixed Populations. *Annu Rev Biomed Data Sci* **6**, 105 (2023).
6. Gudmundsson, S. *et al.* Variant interpretation using population databases: Lessons from gnomAD. *Hum Mutat* **43**, 1012 (2022).
7. Lin, M., Park, D. S., Zaitlen, N. A., Henn, B. M. & Gignoux, C. R. Admixed Populations Improve Power for Variant Discovery and Portability in Genome-Wide Association Studies. *Front Genet* **12**, 673167 (2021).
8. Using population descriptors in genetics and genomics research: A new framework for an evolving field. *Using Population Descriptors in Genetics and Genomics Research: A New Framework for an Evolving Field* 1–217 (2023) doi:10.17226/26902.
9. Peterson, R. E. *et al.* Genome-wide Association Studies in Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations. *Cell* **179**, 589–603 (2019).

10. Hero, J. O., Zaslavsky, A. M. & Blendon, R. J. The United States Leads Other Nations In Differences By Income In Perceptions Of Health And Health Care. *Health Aff (Millwood)* **36**, 1032–1040 (2017).

11. Williams, D. R., Priest, N. & Anderson, N. B. Understanding associations among race, socioeconomic status, and health: Patterns and prospects. *Health Psychol* **35**, 407–411 (2016).

12. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics 2019 51:4* **51**, 584–591 (2019).

13. Zhang, J. & Stram, D. O. The Role of Local Ancestry Adjustment in Association Studies Using Admixed Populations. *Genet Epidemiol* **38**, 502 (2014).

14. Williams Amy, A. L. *et al.* Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico. *Nature* **506**, 97 (2014).

15. Kopp, J. B. *et al.* APOL1 Genetic Variants in Focal Segmental Glomerulosclerosis and HIV-Associated Nephropathy. *J Am Soc Nephrol* **22**, 2129 (2011).

16. Genovese, G. *et al.* Association of trypanolytic ApoL1 variants with kidney disease in African Americans. *Science (1979)* **329**, 841–845 (2010).

17. Gonzaga-Jauregui, C. *et al.* Mutations in COL27A1 cause Steel syndrome and suggest a founder mutation effect in the Puerto Rican population. *European Journal of Human Genetics* **23**, 342 (2015).

18. Padoa, C., Goldman, A., Jenkins, T. & Ramsay, M. Cystic fibrosis carrier frequencies in populations of African origin. *J Med Genet* **36**, 41 (1999).

19. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* **42**, (2014).

20. Gudmundsson, S. *et al.* Variant interpretation using population databases: Lessons from gnomAD. *Hum Mutat* **43**, 1012 (2021).

21. Cavalli-Sforza, L. L. The Human Genome Diversity Project: past, present and future. *Nature Reviews Genetics 2005 6:4* **6**, 333–340 (2005).

22. Auton, A. *et al.* A global reference for human genetic variation. *Nature 2015 526:7571* **526**, 68–74 (2015).

23. Marnetto, D. *et al.* Ancestry deconvolution and partial polygenic score can improve susceptibility predictions in recently admixed individuals. *Nature Communications 2020 11:1* **11**, 1–9 (2020).

24. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature 2016 536:7616* **536**, 285–291 (2016).

FIGURES

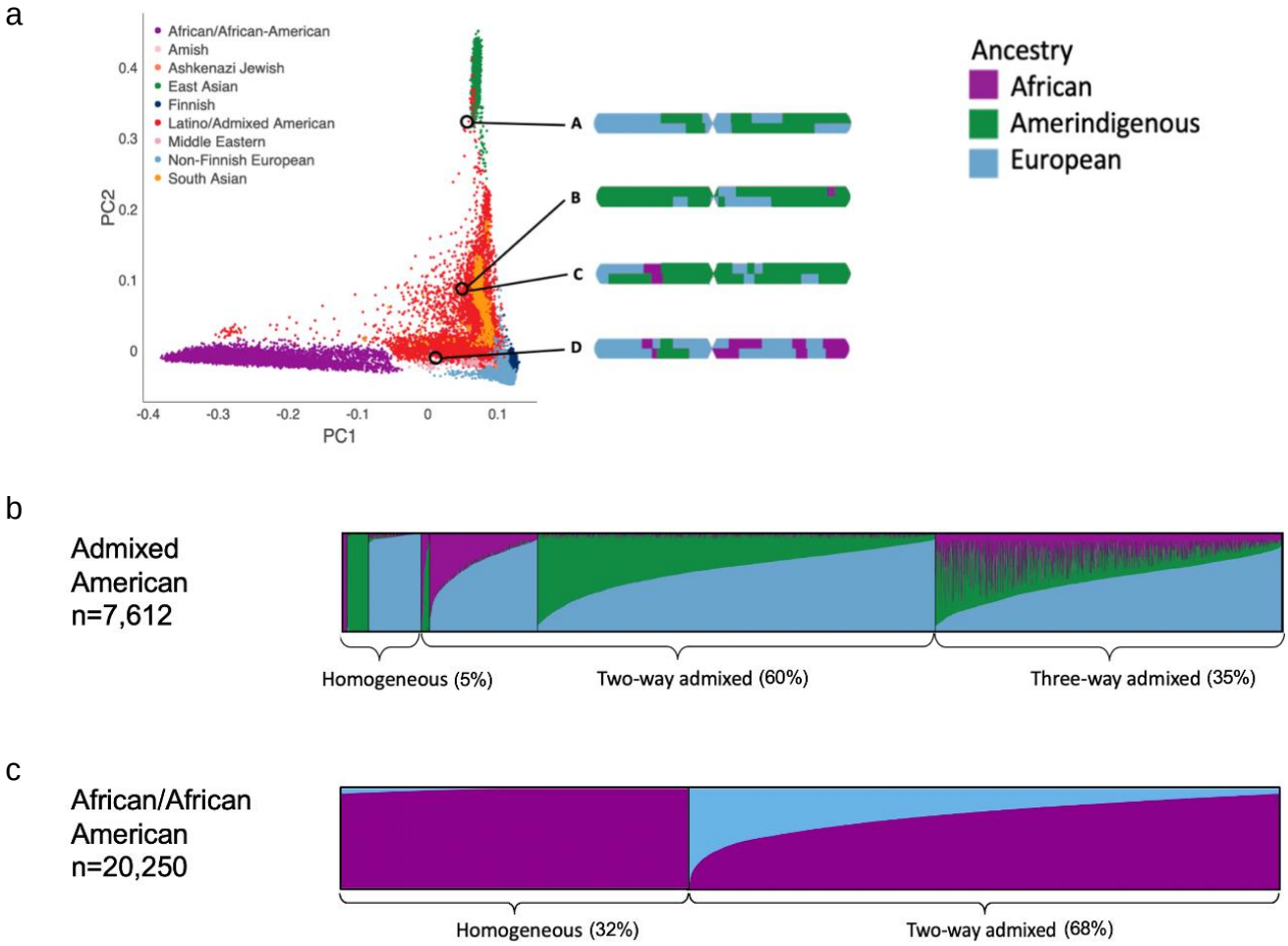


Figure 1. Population structure of the gnomAD Admixed American and African/African American genetic ancestry groups. a. PCA plot for all gnomAD samples with the painted karyograms of a single chromosome from an individual within each circle. Note that individuals B and C have strikingly different local ancestry patterns despite appearing in the same location in PC space. **b.** ADMIXTURE plot for the gnomAD Admixed American genetic ancestry group displaying global ancestry proportions. 5% of individuals derive ancestry from a single continent, 60% from two, and 35% from three. **c.** ADMIXTURE plot for the African/African American genetic ancestry group showed admixture primarily from two continents, with 32% homogeneous AFR individuals.

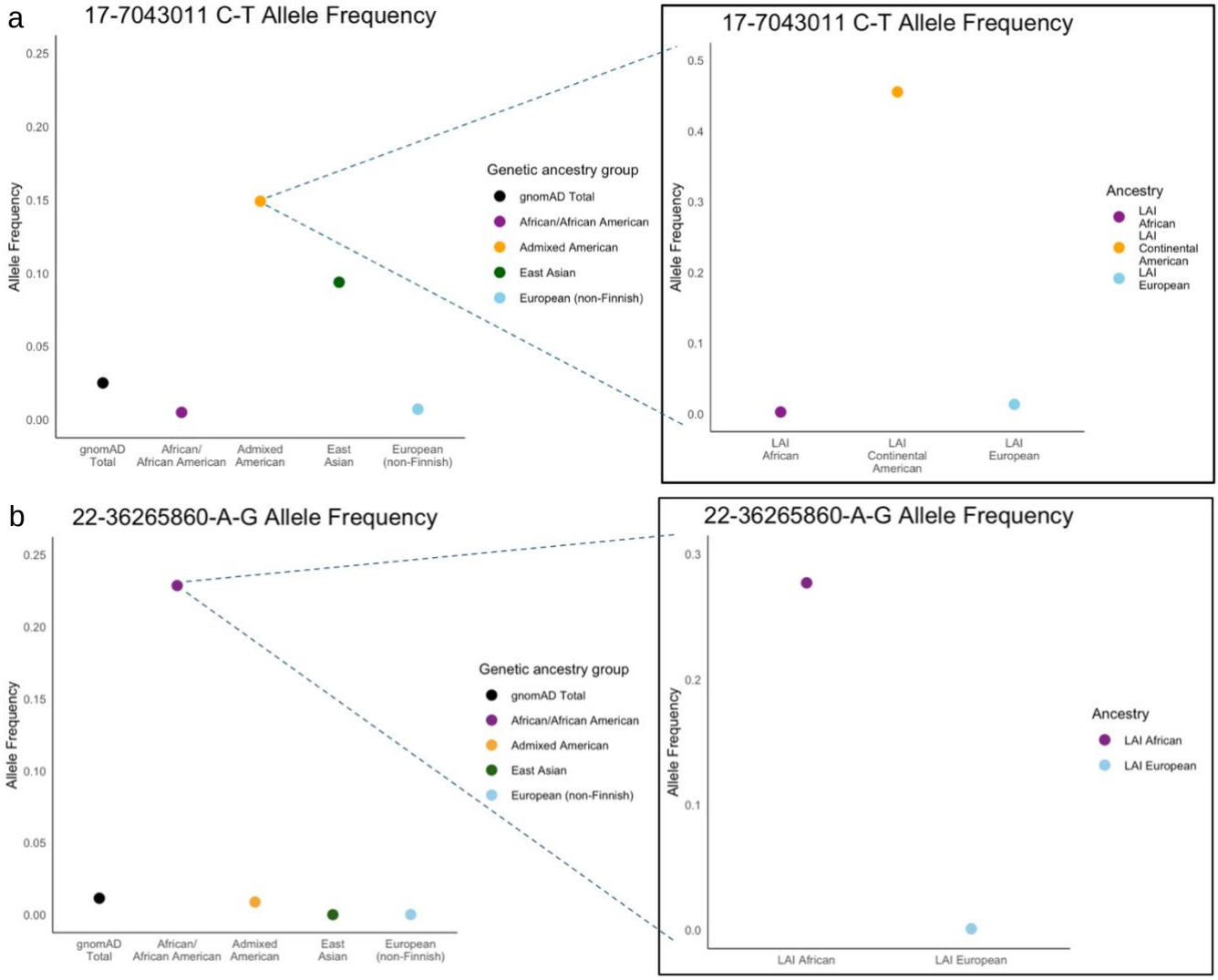


Figure 2: Local ancestry better resolves ancestral allele frequencies. a. A variant within the *SLC16A11* gene (rs117767867), associated with increased risk of diabetes[16], has divergent frequencies estimated in the Admixed American genetic ancestry group based on aggregate versus local ancestry. **b.** Frequency of a variant within the *APOL1* gene (rs73885319), linked with increased susceptibility to FSGS, HIVAN, and ESKD[17], in the African/African American genetic ancestry group based on aggregate and local ancestry. Note the striking heterogeneity in ancestry-specific AF estimates that is not able to be resolved with just the aggregate frequency estimate.

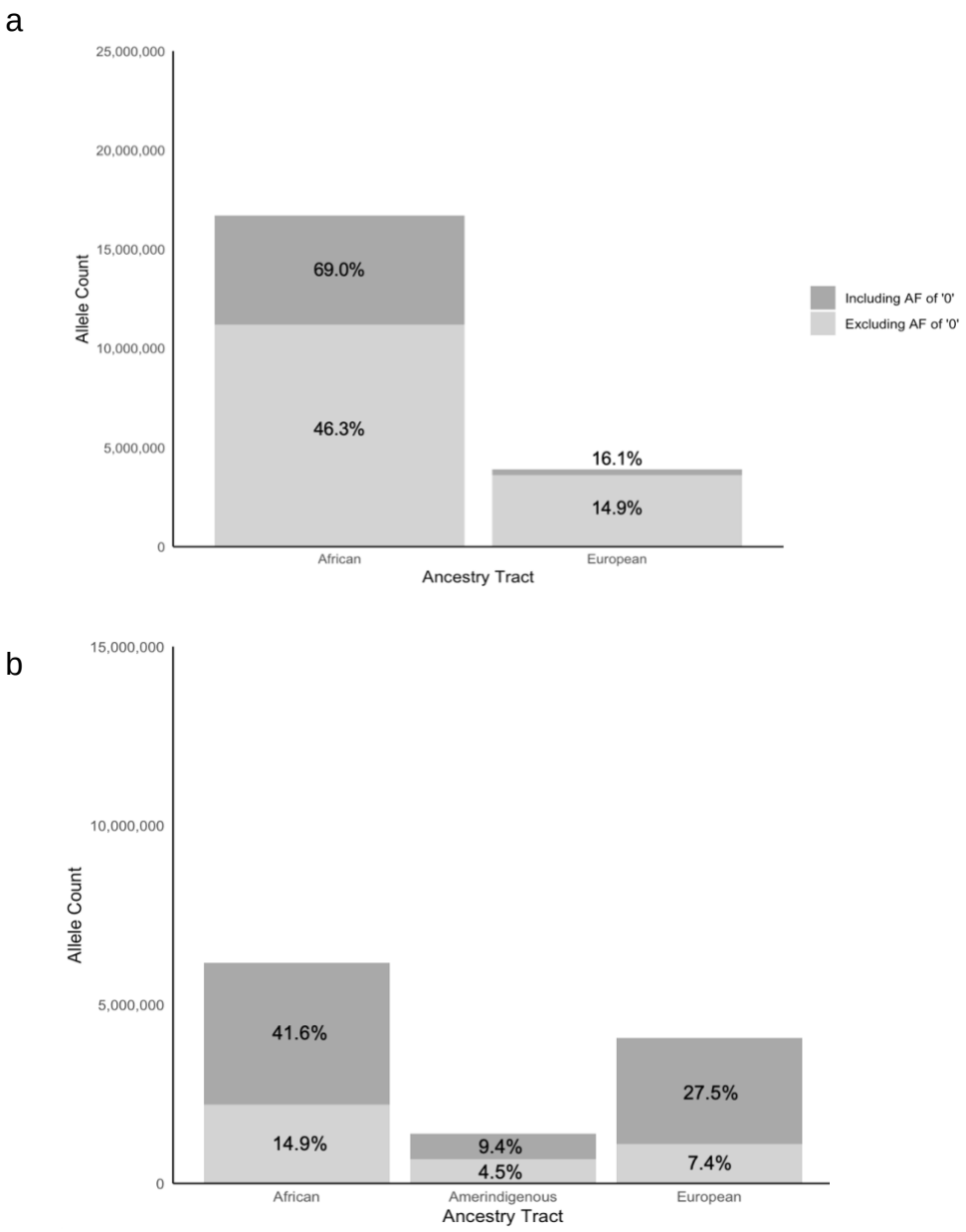


Figure 3: Distribution of variants with specific ancestry tract allele frequencies ≥ 2x other ancestral tracts across genetic ancestry groups. The total percentage for each local-ancestry informed tract is indicated on the bars. **a**. 85.1% of variants with LAI data are at least 2x different between ancestry-specific AFs in the African/African American genetic ancestry group. **b.** 78.5% of variants with LAI data are at least 2x different between ancestry-specific AFs in the Admixed American genetic ancestry group.

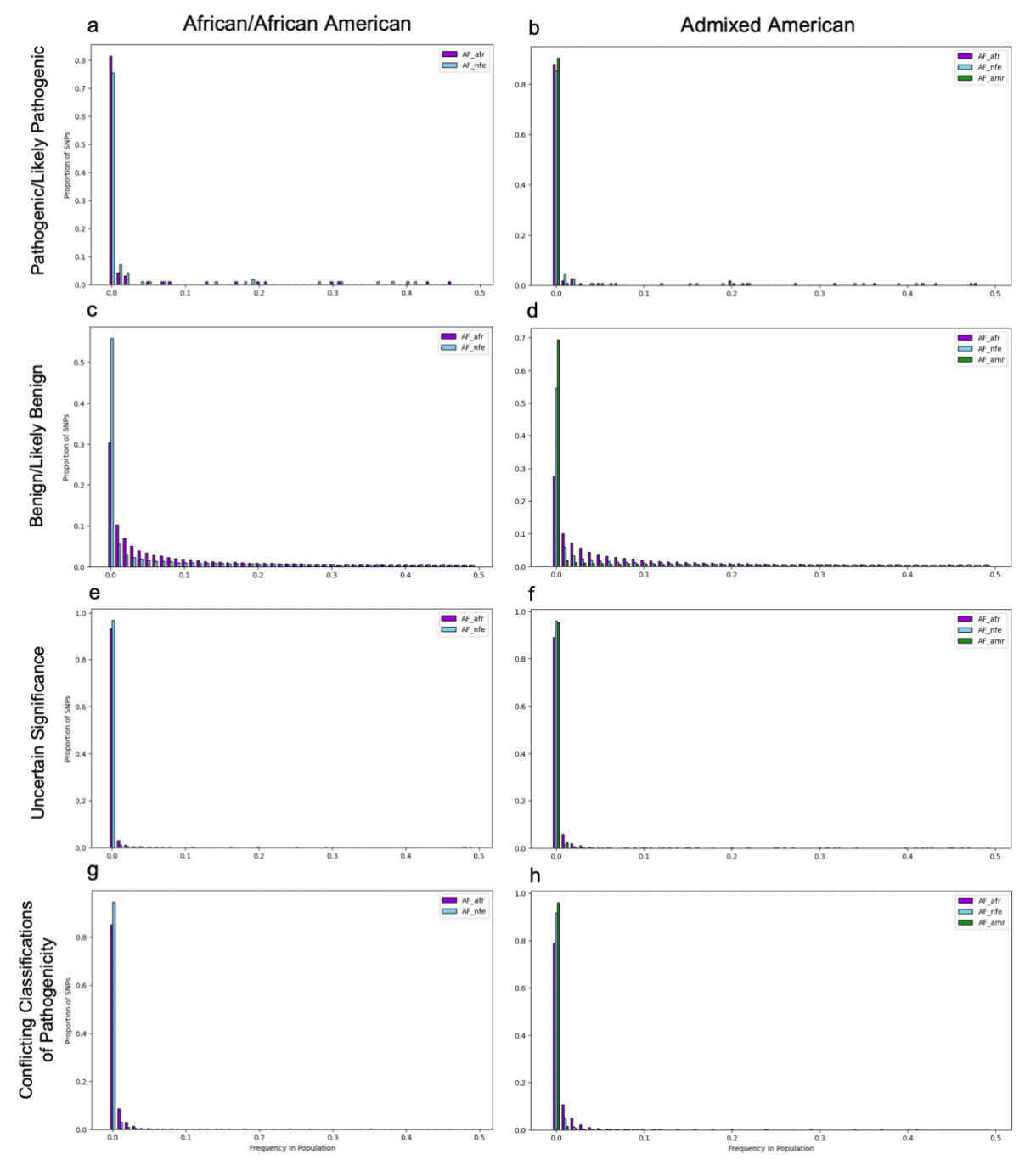


Figure 4: Site Frequency Spectrum for the inferred African/African American and Admixed American genetic ancestry groups based on ClinVar classifications. LAI-derived AFs for component ancestries are indicated with colors.

Methods

Genetic ancestry group labels were assigned to samples as previously described[1], using principal components (PCs) computed from hail's hwe_normalized_pca[2] as features when training a random forest classifier using reference samples with a known continental origin (**Figure 1a**). Specifically, a continental genetic ancestry group was assigned to all samples for which the probability of that continental genetic ancestry group was > 90% according to the random forest model. Samples were subset based on inferred genetic ancestry groups, limiting to bi-allelic variants with low genotype missingness (i.e., call rate > 0.9) and an allele frequency greater than 0.1% within the Admixed American and African/African American gnomAD genetic ancestry groups. The 7,612 Admixed American samples and 20,250 African/African American samples from gnomAD v3.1 used here include all individuals in the database except those from the 1000 Genomes Project (1kGP) and Human Genome Diversity Project (HGDP) joint callset[3], which were used to curate reference panels for LAI in each genetic ancestry group.

To explore the overall population composition of the gnomAD African/African American and Admixed American samples, we conducted principal components analysis (PCA) and unsupervised ADMIXTURE[4] to infer global genetic ancestry proportions. Informed by these global ancestry estimates, we set k=3 for the Admixed American genetic ancestry group (including all reference samples from 1kG/HGDP labeled as belonging to the AFR, EUR, and AMR superpopulations). For the African/African American genetic ancestry group, we set k=2 (using just continental reference samples from the AFR and EUR superpopulations).

To infer local ancestry data in the gnomAD Admixed American and African/African American group, we built an open-source LAI pipeline. **Extended Data Figure 1** shows a schematic of the software pipeline, which includes four primary steps: phasing, LAI, ancestry extraction/dosage calculation, and LAI-informed allele frequency generation. We used Eagle[5] for separately phasing the Admixed American reference panel and for joint phasing of the African/African American group, as it is computationally tractable at scale[5]. We used RFmixV2[6] for LAI, which requires a reference panel representative of the continental ancestries present within the samples. Our reference panel was composed of the harmonized HGDP and 1kGP[3] "AMR" samples to capture the Amerindigenous genetic components in the Admixed American group, "EUR" samples for the European components, and "AFR" samples to capture continental African genetic ancestries. RFmixv2[6] was run with one EM iteration for the Admixed American samples to account for the high amount of admixture present in the three-way reference. For the curated reference for the African/African American group, we excluded the Americans of African Ancestry in Southwest USA (ASW) and African Caribbeans in Barbados (ACB) due to their higher level of admixture. Given the substantially larger sample size of the African/African American genetic ancestry group in gnomAD, reducing admixture within the reference panel and therefore removing the need for an EM iteration dramatically reduces the required computing time.

To extract ancestral segments and calculate ancestry-specific minor allele dosages, we used the tool Tractor[7]. Due to computational constraints, samples from the inferred African American genetic ancestry group were run in batches of 6,750 samples for chromosomes 1 and 3-12, and batches of 3,375 samples for chromosome 2 (**Supplementary Table 2**) for LAI. Since the same reference panel was used across batches, this is not expected to impact results. Finally, we

tabulated local ancestry-informed allele frequencies using a custom Python script that utilizes the Hail library[2]. To evaluate the pipeline's sensitivity, we applied it to a simulated admixed American truth dataset and computed the true positive rates of LAI[8] (**Supplementary Figure 3**). We also validated our pipeline by comparing global ancestry proportions estimated by ADMIXTURE to those from our LAI pipeline results (**Supplementary Figure 4**). The entire pipeline is unified using the Hail Batch Python module for scalability and is available at [github.com/broadinstitute/gnomad_local_ancestry].

Data availability

Data can be downloaded from or browsed interactively at [gnomad.broadinstitute.org].

Code availability

Code is freely available at [github.com/broadinstitute/gnomad_local_ancestry]

Methods References

1. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature 2020 581:7809* **581**, 434–443 (2020).
2. Hail | Genetics. https://hail.is/docs/0.2/methods/genetics.html.
3. Koenig, Z. *et al.* A harmonized public resource of deeply sequenced diverse human genomes. *bioRxiv* (2024) doi:10.1101/2023.01.23.525248.
4. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**, 1655 (2009).
5. Loh, P. R., Palamara, P. F. & Price, A. L. Fast and accurate long-range phasing in a UK Biobank cohort. *Nature Genetics 2016 48:7* **48**, 811–816 (2016).

6. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. *Am J Hum Genet* **93**, 278 (2013).

7. Atkinson, E. G. *et al.* Tractor uses local ancestry to enable the inclusion of admixed individuals in GWAS and to boost power. *Nat Genet* **53**, 195–204 (2021).

8. Honorato-Mauer, J. *et al.* Characterizing features affecting local ancestry inference performance in admixed populations. *bioRxiv* 2024.08.26.609770 (2024) doi:10.1101/2024.08.26.609770.

Acknowledgments

We wish to thank Alicia Byrne for literature review; Jackie Goldstein, Dan King, and Ben Weisburd for pipeline scalability assistance; Marcos Santoro for providing LAI reference panel guidance; Julia Goodrich for pipeline code review. This work was funded by the National Institutes of Health (K01 MH121659 and R01 HG012869 to E.G.A.; P.K. was supported by T32 GM139534-03 and T32 GM136554-05).

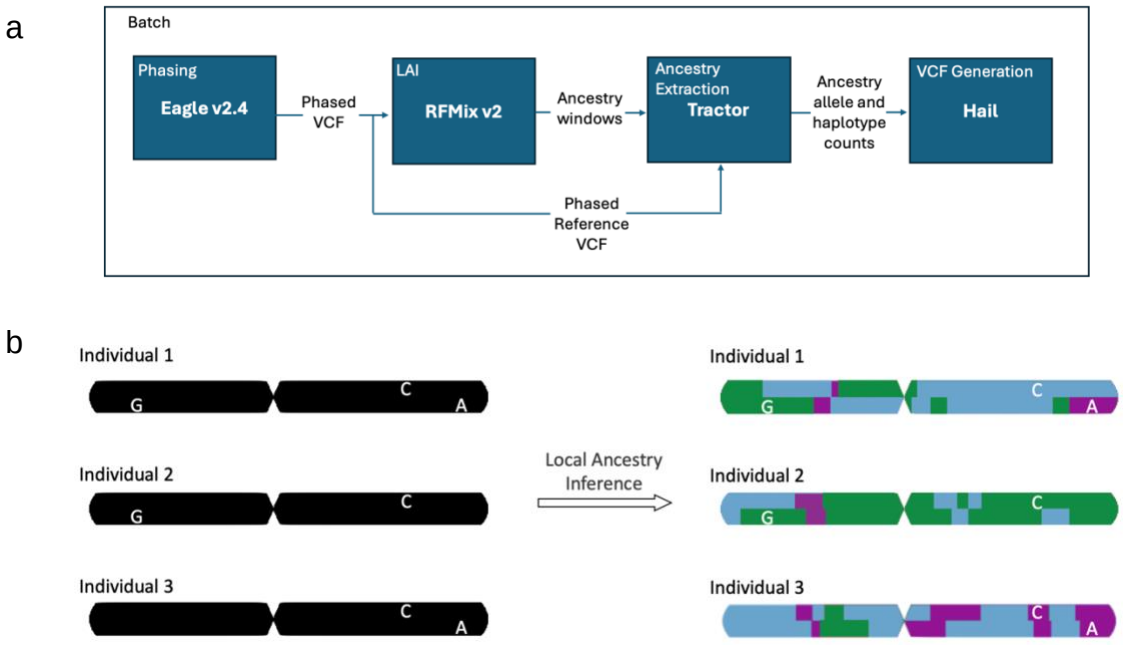
Author Contributions

P.K. and M.W. led data analysis with support from G.T., K.C., S.M.B., J.H.M., H.L.R., M.J.D. and K.J.K. advised on analyses. P.D. and N.W incorporated LAI data in the browser. E.G.A. designed the study and supervised analysis. All authors have read and approve this manuscript.

Competing Interest Declaration

M.J.D is a founder of Maze Therapeutics. K.J.K. is a consultant for Tome Biosciences, AlloDx, and Vor Biosciences, and a member of the scientific advisory board of Nurture Genomics. Other authors declare no competing interests.

EXTENDED DATA FIGURES



Extended Data Figure 1: **Schematics illustrating derivation of local-ancestry informed allele frequencies. a**. LAI pipeline and data generation, **b.** A single chromosome from three individuals, with non-reference alleles at 3 sites of variation (indicated by G, C, and A). After performing LAI, we observe that Allele G is only found on the green ancestry, Allele A only on the purple ancestry, and Allele C can be found across all three.