Improved Allele Frequencies in gnomAD through Local Ancestry Inference

Pragati Kore[1*], Michael W. Wilson[2*], Grace Tiao[2], Katherine Chao[2], Philip W. 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[2,3,4,7], Elizabeth G. Atkinson[1,3,4,8,†]

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. Novo Nordisk Foundation Center for Genomic Mechanisms of Disease, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA
8. 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) is a foundational resource for allele frequency data, widely used in genomic research and clinical interpretation. However, traditional estimates rely on individual-level genetic ancestry groupings that may obscure variation in recently admixed populations. To improve resolution, we applied local ancestry inference (LAI) to over 27 million variants in two admixed groups: Admixed American (n = 7,612) and African/African American (n = 20,250), deriving ancestry-specific allele frequencies. We show that 78.5% and 85.1% of variants in these groups, respectively, exhibit at least a twofold difference in ancestry-specific frequencies. Moreover, 81.49% of variants with LAI information would be assigned a higher gnomAD-wide maximum frequency after incorporating LAI, potentially altering clinical interpretations. This LAI-informed release reveals clinically relevant frequency differences that are masked in aggregate estimates and may support reclassifying some variants from Uncertain Significance to Benign or Likely Benign.

INTRODUCTION

The Genome Aggregation Database (gnomAD) is one of the most widely used resources in the genomics community[1], 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[2,3]. 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[3]. 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[4], the genetic ancestry groups used in gnomAD's aggregations are, by nature, discrete. Specifically, these groups are defined by assigning individuals a population label defined by global

genetic ancestry as trained on reference-panel informed principal components[1], which can obscure finer-scale genetic differences observed across individuals within a genetic ancestry group[5–8]. 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 individual-level 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, rather than entire individuals[9]. By applying LAI, we resolve finer-scale allele frequencies, allowing for better-informed clinical interpretations. This added resolution also increases confidence in classifying variants as benign/likely benign when they are common in any ancestry, thereby lowering the overall burden of variants of uncertain significance across all genetic ancestry groups.

We wish to be clear that 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 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[5,6]. Throughout this article, we

93 use the term “Amerindigenous” to refer to local ancestry tracts inferred to originate from
94 Indigenous American populations, based on reference haplotypes harmonized from AMR-
95 designated samples in the Human Genome Diversity Project (HGDP) and 1000 Genomes Project
96 Phase 3 (1kGP3). To distinguish between individual-level genetic ancestry groupings (e.g.,
97 Admixed American) and ancestry-specific segments within individuals, we refer to frequencies
98 derived from painted tracts as LAI-AMR (Amerindigenous), LAI-AFR (African), and LAI-EUR
99 (European), respectively. Here, we enhance the granularity of gnomAD’s allele frequency data in
100 the inferred Admixed American and African/African American genetic ancestry groups, providing
101 a more accurate representation of genetic variation and its implications for human health.
102
103 **RESULTS**
104 **Release of local ancestry-informed frequencies in gnomAD**
105 Here we release local ancestry-informed frequency data for 14,804,206 bi-allelic single nucleotide
106 polymorphisms (SNPs) of at least MAF 0.1% within the Admixed American (n=7,612) and
107 24,204,574 bi-allelic SNPs within the African/African American genetic ancestry groups
108 (n=20,250), both from gnomAD v3.1. Of the total SNPs in each genetic ancestry group, 9.82% of
109 SNPs in the African/African American group and 11.57% of SNPs in the Admixed American
110 group passed QC and were assigned local ancestry calls. This initial release of gnomAD LAI data
111 contains variants’ estimated alternate allele counts, allele numbers, and allele frequencies
112 partitioned by haplotype-level continental ancestries. These new local ancestry-informed
113 frequencies are released in a downloadable VCF format file and can be explored interactively on
114 the gnomAD browser (https://gnomad.broadinstitute.org/). To access the local ancestry data for a
115 specific variant within the browser, select the "Local Ancestry" tab within the genetic ancestry

116 group frequency table on the variant's page. If local ancestry data is available for the variant, the
117 ancestral allele count (AC), allele number (AN), and allele frequency (AF) estimates will be
118 displayed (**Supplementary Fig. 1**).
119
120 **Ancestry composition of gnomAD African/African Americans and Admixed Americans**
121 Plotting the first two PCs of the full gnomAD collection, we can see the inferred gnomAD
122 Admixed Americans span a large portion of global PC space, with some samples clustering closer
123 to European reference samples, others clustering closer to the African reference samples, and some
124 clustering closer to the East Asian reference samples, indexing the presence of Amerindigenous
125 ancestry[10] (**Fig. 1a, Supplementary Fig. 2a**). Resolving global ancestry with ADMIXTURE, we
126 again observe that samples classified as Admixed American by gnomAD exhibit heterogeneous
127 genetic ancestry, with individuals containing differing proportions of two or three-way admixture
128 between Amerindigenous (AMR), African (AFR), and European (EUR) haplotypes. After
129 applying a 5% ancestry inclusion threshold, 5% of individuals derive their genetic ancestry
130 primarily from a single continent (homogeneous AMR or EUR or AFR), 60% from two continents
131 (AFR/AMR or AFR/EUR or AMR/EUR), and 35% from three continents (AMR/EUR/AFR) (**Fig.**
132 **1b**). Given this composition, we modeled the Admixed American samples with a three-way
133 reference panel (see *Methods*). The African/African American samples primarily lie on a cline
134 between the EUR and AFR poles in PC space, indicating admixture between largely these two
135 genetic ancestry components. Further exploring African/African American samples' ancestry with
136 ADMIXTURE, we again observed individuals in this group contain primarily AFR and EUR
137 components, with 32% homogeneous AFR individuals present and 68% of samples classified as

138 two-way admixed (**Fig. 1c**). Therefore, we curated a reference panel comprising two-way
139 continental AFR and EUR genetic ancestries for this group.
140
141 Local ancestry deconvolution provides additional resolution beyond global inference; we observe
142 that even individuals occupying the same point in PC space have markedly different local ancestry
143 patterns (**Fig. 1a**). Notably, even across higher principal components (e.g., PC1 to PC12),
144 individuals 'b' and 'c' remain closely clustered, indicating that despite differences in their local
145 ancestry, their overall genetic profiles are still similar when considering finer-scale genetic
146 variation (**Supplementary Fig. 2b-g**). With only genetic ancestry group level rather than ancestry-
147 specific allele frequencies, estimates represent a weighted average of haplotype-derived
148 frequencies and can thus be distorted from the more precise ancestry-specific frequencies. This
149 additional resolution can be meaningful for variants of functional importance.
150
151 **Functional variant case studies**
152 For example, the gnomAD variant 17-7043011-C-T in *SLC16A11*, associated with type 2 diabetes
153 risk, is elevated within the Admixed American genetic ancestry group[11] at a 24% allele frequency
154 compared to the gnomAD-wide gnomAD allele frequency of 2% (**Fig. 2c**). Breaking down this
155 variant's frequency into continental allele frequency estimates, we see in fact that the variant
156 occurs at a 45% allele frequency in LAI-AMR regions and closer to 0% in the LAI-AFR and LAI-
157 EUR regions. Similarly, the gnomAD variant 22-36265860-A-G in *APOL1*, linked with increased
158 susceptibility to focal segmental glomerulosclerosis (FSGS), HIV-associated nephropathy
159 (HIVAN), and hypertensive end-stage kidney disease (ESKD)[12], shows a 27% allele frequency in

160 LAI-AFR regions of the African/African American genetic ancestry group[13], compared to a 1%
161 gnomAD-wide frequency (**Fig. 2d**).
162
163 LAI is useful in resolving allele frequencies for rare variants as well as common. Variants in the
164 *COL27A1* gene have been shown to cause Steel syndrome and suggest a founder mutation effect
165 in Puerto Rico[14]. Within gnomAD, the causal variant 9-114195977-G-C[14] has an allele frequency
166 of 0.1% in the Admixed American genetic ancestry group, while the gnomAD-wide allele
167 frequency is 0.006% (**Fig. 2e**). Separating this variant's gnomAD Admixed American genetic
168 ancestry group frequency into continental allele frequency estimates, we see that the variant occurs
169 almost at a 1% allele frequency in LAI-AMR regions and closer to 0% in LAI-AFR and LAI-EUR
170 regions, reinforcing a founder effect hypothesis and again showing the usefulness of local ancestry
171 in large genomic datasets. Another example is the gnomAD variant 7-117606754-G-A in *CFTR*,
172 associated with cystic fibrosis[15]. This variant shows a 0.1% allele frequency in LAI-AFR regions
173 of the African/African American genetic ancestry group, compared to a 0.007% gnomAD-wide
174 frequency (**Fig. 2f**).
175
176 Allele frequency data is used for more than just clinical interpretation and pathogenicity.
177 Population allele frequency (AF) databases have recently been utilized to estimate aggregate
178 carrier frequencies and genetic prevalence, providing insights into the potential number of affected
179 individuals[16]. LAI data enhances these estimations, enabling more precise predictions of
180 populations with potentially higher genetic prevalence. For instance, the pathogenic variant 2-
181 168970131-G-A (p.Arg575Ter) in the *ABCB11* gene, which causes Bile Salt Export Pump (BSEP)
182 deficiency resulting in cholestasis[17], has a frequency of 0.04845% in the Admixed American

183 genetic ancestry group, the highest observed across all gnomAD genetic ancestry groups. The
184 cumulative carrier frequency of all likely pathogenic/pathogenic (LP/P) variants in *ABCB11* across
185 genetic ancestry groups suggests that the 2-168970131-G-A variant leads to a slightly higher
186 carrier frequency in this group when compared to the total estimate[18]. However, analysis of LAI
187 data revealed that this variant was exclusively observed in the LAI-EUR regions of Admixed
188 American samples. Consequently, the elevated carrier frequency, and thus the estimated genetic
189 prevalence, primarily pertains to individuals of higher LAI-EUR ancestry within the broader
190 genetic ancestry inferred label. These variant examples highlight the importance of more precise
191 allele frequencies in accurately identifying the distribution of genetic risk factors within diverse
192 and admixed genetic ancestry groups.
193
194 **Divergence in allele frequencies across ancestry tracts**
195 While many variants are similar in their ancestry-specific allele frequency estimates, many are also
196 highly divergent (**Supplementary Fig. 3**). The overall correlation in ancestry-specific allele
197 frequencies between the LAI-AFR and LAI-EUR tracts in the African/African American group
198 was 0.83. For the Admixed American group, the correlations were 0.80 between LAI-AFR and
199 LAI-EUR AFs, 0.73 between LAI-AFR and LAI-AMR AFs, and 0.84 between LAI-AMR and
200 LAI-EUR AFs. Allele frequency variance was highest within the LAI-AMR tract (0.0447),
201 compared to the LAI-AFR (0.0367) and LAI-EUR tracts (0.0356), suggesting greater
202 heterogeneity among variants assigned to this ancestry. This elevated variance may reflect
203 historical bottlenecks or founder effects that have shaped allele distributions in LAI-AMR
204 haplotypes. Moreover, pairwise Levene's tests for equality of variances between ancestry tracts
205 were all highly significant ($p < 1 \times 10^{-16}$). Strikingly, in the African/African American genetic

206 ancestry group, 85.1% of variants with LAI data exhibit at least a twofold greater difference
207 between ancestry-specific AFs (**Fig. 3a**). The majority of these variants (69%) have an LAI-AFR
208 AF that is at least two times greater than the LAI-EUR AF, while 16.1% show the opposite pattern.
209 Similarly, in the Admixed American genetic ancestry group, 78.5% of variants demonstrate this
210 level of divergence between ancestry-specific AFs (**Fig. 3b**). Among these, 41.6% of variants have
211 an LAI-AFR AF that is two times greater than those of either other ancestry's tracts, while the
212 remaining variants show a dominant LAI-EUR AF (27.5%) or LAI-AMR AF (9.4%). Overall,
213 LAI-AFR tracts consistently exhibit higher AFs compared to LAI-EUR or LAI-AMR tracts, both
214 when sites that are monomorphic in the other ancestries are considered but also when excluding
215 such loci. To complement this analysis, we also evaluated ancestry-specific depletion - variants
216 whose AF in one ancestry tract is less than or equal to half that of the others (**Fig. 3c-d**) providing
217 a more comprehensive genome-wide view of how allele frequencies diverge across ancestries.
218 When restricting to monomorphic sites in the depleted tracts, 62.2% of variants in the
219 African/African American group and 26.2% in the Admixed American group met this depletion
220 criterion.
221
222 **Local ancestry-informed frequencies can better resolve variant clinical relevance**
223 Visualizing the site frequency spectrum (**Fig. 4, Supplementary Fig. 4, Supplementary Table**
224 **1**) highlights the allele frequency distributions across two genetic ancestry groups classified by
225 their ClinVar annotations[19]. Of the LAI variants, only a small fraction have corresponding ClinVar
226 annotations: 131,886 out of 14,804,206 (0.9%) in the Admixed American and 183,164 out of
227 24,204,574 (0.8%) in the African/African American genetic ancestry groups. Because single-
228 submitter and no-criteria entries may be less reliable, we applied an additional QC filter, retaining

229 only annotations with ≥ 2 ClinVar stars, resulting in variant counts of 73,087 and 96,851,
230 respectively. We additionally examined the number of variants with a tract-specific allele
231 frequency at least twice as high as the other ancestral tract frequencies: the African/African
232 American group has 75,756 variants when including all shared alleles (66,100 variants when
233 considering only monomorphic variants), and the Admixed American group has 53,588 variants
234 when including all shared alleles (24,683 variants when considering only monomorphic
235 variants)(**Supplementary Fig. 5**). The results indicate that most variants across all ClinVar
236 categories — including pathogenic/likely pathogenic, benign/likely benign, uncertain significance,
237 and conflicting classifications of pathogenicity — are rare, with AFs concentrated between 0 and
238 0.1 in all genetic ancestry groups, regardless of ClinVar filtering (**Supplementary Fig. 6**). To
239 further investigate if benign classifications may be driven primarily by frequency thresholds, we
240 zoomed-in on benign variants and again on benign high-confidence ClinVar variants (≥2 stars)
241 with allele frequency <0.1 and <0.01 (**Supplementary Fig. 7**). We conducted a two-sample
242 Kolmogorov–Smirnov test on benign/likely benign ClinVar variants in the African/African
243 American group, comparing the SFS between LAI-AFR and LAI-EUR. The test yielded a KS
244 statistic of 0.278 with a p-value near zero (p = 10^-100), indicating a significant difference between
245 the two distributions. The broader distribution of benign variants in African/African American
246 groups reinforces the need to include diverse ancestry groups in genomic studies, as it
247 demonstrates how genetic diversity impacts allele frequency distributions and, by extension,
248 clinical variant interpretation.
249
250 We additionally investigated the frequency of ancestry-enriched AFs exceeding those of grpmax
251 (formerly popmax)[2], which measures the highest allele frequency observed in any gnomAD

genetic ancestry group. As grpmax is often used to inform a variant classification of benign/likely benign, better resolving allele frequencies across ancestral haplotypes with LAI provides a more precise estimate of frequencies that may change clinical interpretation. 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 2)**. 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.

To further explore the clinical relevance of LAI-informed allele frequencies, we focused on variants below the commonly used 5% allele frequency threshold, which is treated as stand-alone benign evidence under ACMG's "BA1" guideline[20] (**Supplementary Table 3**). Comparing the former grpmax AF to LAI-informed tract AFs in this critical range revealed that for the Admixed American group, 563,042 of 8,527,238 variants (~6.6%) had a tract AF exceeding 5%. In contrast, 475,698 of 15,775,667 variants (~3.0%) in the African/African American group showed the same pattern. To determine whether these shifts persisted among clinically curated variants, we repeated this analysis using only ClinVar variants with ≥ 2 stars [20]. Among high-confidence variants, 3,164 of 31,485 variants (~10.0%) in the Admixed American group and 2,684 of 42,461 variants (~6.3%) in the African/African American group had tract AFs above 5% that exceeded their grpmax values. These findings highlight that ancestry-specific allele frequency information can better assist in the

identification of likely benign variants. We additionally conducted a similar analysis stratifying by CADD PHRED score thresholds (≥10 and ≥15, corresponding to the top 10 and 5 percent of deleterious variants in the genome)[21], comparing the frequency of variants where LAI-informed AF exceeded 5% and grpmax. For the Admixed American group, 6.1% of variants with CADD ≥10 and 5.7% with CADD ≥15 met this condition. For the African/African American group, corresponding proportions were 2.7% and 2.4%, respectively. These trends reinforce the observation that local ancestry resolution meaningfully alters the interpretation of both potentially benign and functionally consequential variants. 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

298 frequencies vary dramatically across the globe[22,23] due to features such as demography, genetic
299 drift, and natural selection. Deconvolving from which continental ancestry a variant is derived
300 allows us to compute more accurate expected allele frequencies for each continental background,
301 instead of aggregating frequencies across recently admixed individuals regardless of the ancestral
302 origin of the variant's haplotype[24]. The large sample size of gnomAD enables us to examine allele
303 frequency trends on a more global scale, and with the integration of LAI, we achieve a finer
304 resolution of ancestry-specific frequency estimates, improving the utility of the resource. In some
305 cases, updated allele frequencies that take into account local ancestry differ significantly from their
306 corresponding genetic ancestry group-level frequencies, and even update the highest allele
307 frequency observed gnomAD-wide (grpmax), and/or increase frequency estimates past commonly
308 used thresholds for estimating pathogenicity, which may impact the clinical interpretation of a
309 variant (**Supplementary Fig. 1**).
310
311 Our analyses further reveal that a substantial number of variants exhibit highly divergent allele
312 frequencies across ancestry backgrounds: within the African/African American and Admixed
313 American genetic ancestry groups, 85.1% and 78.5% of variants with LAI data, respectively, show
314 at least a twofold difference between ancestry-specific frequencies. Allele frequency differences
315 observed across local ancestry backgrounds may arise from a combination of evolutionary and
316 demographic forces, including genetic drift, founder effects, and natural selection. While we do
317 not attempt to infer the specific historical or selective pressures underlying the observed patterns
318 in this study, the extensive allele frequency divergence we report highlights the importance of
319 modeling ancestry at finer scales. Future work leveraging these local ancestry-informed

frequencies may provide insight into the history of population-enriched variants, particularly as more detailed demographic models and functional data become available.

Additionally, we found that across both genetic ancestry groups, 81.49% of the 27,412,225 sites analyzed would receive an increased grpmax value 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 without this information. This misestimation is especially relevant near clinical decision thresholds. The ACMG guidelines designate a AF ≥5% as strong benign evidence (BA1)[25], yet our analysis shows that many variants below this threshold by grpmax may exceed it in one or more ancestry tracts. Notably, over 10% of high-confidence ClinVar variants in the Admixed American group and 6.3% in the African/African American group cross this critical boundary when LAI is incorporated. These shifts were consistent even when stratifying by predicted functional impact using CADD scores[21]. In both genetic ancestry groups, variants with CADD ≥10 and ≥15 also exceeded the 5% threshold when tract AFs were used in place of grpmax. In terms of direct patient impact, reclassifying these variants as benign spares patients the uncertainty, repeated imaging, and potential invasive interventions they might otherwise undergo. These findings suggest that LAI enhances variant interpretation by identifying variants that, while rare globally, are common in specific ancestries—supporting their benign classification across all populations and refining our understanding of functional variant distribution across populations.

342 Ancestry-specific allele frequencies can additionally help refine genetic prevalence estimates for
343 recessive diseases. Understanding the estimated proportion of a population that has a causal
344 genotype for a genetic disorder, and whether there is a significantly higher risk in a specific genetic
345 ancestry group or not, is a key question for understanding the global impact of a disease. Local
346 ancestry has the potential to refine these estimates further and provide the community with a more
347 accurate prediction of how many people have a given disease. While a 0.1% frequency cut-off is
348 still more common than many pathogenic disease-causing variants, in the future local ancestry will
349 be able to be applied to more rare variants, which will have an even greater effect on clinical care.
350
351 Local ancestry resolved variant allele frequencies can also help in the interpretation of GWAS
352 results[5,26]. Differentiating between genetic ancestry group-level allele frequencies and those
353 enriched in specific haplotypes can streamline the search for disease-associated variants with
354 elevated importance for particular genetic ancestries[27]. For example, we demonstrate that the
355 gnomAD variant 17-7043011-C-T in *SLC16A11*[11], associated with type 2 diabetes risk, is observed
356 at a 14% frequency in the Admixed American population compared to 2% gnomAD-wide. When
357 partitioned by continental ancestry, this variant occurs at a 45% frequency in LAI-AMR regions
358 of admixed genomes, while it is nearly absent in LAI-AFR and LAI-EUR regions. This, and the
359 other vignettes highlighted, showcases how local ancestry inference better resolves allele
360 frequencies in both rare and common key variants.
361
362 We recognize that genetic ancestry is inherently continuous and not discrete. For the purposes of
363 local ancestry inference, we approximate ancestry as originating from discrete source populations,
364 a necessary simplification for modeling recently admixed populations. Our approach does not

assume that present-day populations are strictly discrete entities but rather leverages the fact that the major sources of genetic variation in recently admixed populations can be well-modeled by a small number of continental reference populations. This assumption holds well for cases of recent admixture, however, as one moves beyond recent admixture events such ancestry becomes increasingly diffuse. We additionally appreciate that continental regions are not monoliths, but contain numerous ancestries therein. Our method is not inherently limited to broad continental ancestries, and in principle, it could be extended to analyze intra-continental variation given appropriate reference data. However, the resolution of such analyses depends on the availability and comprehensiveness of reference panels as well as the degree of genetic differentiation between modeled ancestries[28]. For this reason, we are currently limited in resolving LAI to the continental rather than subcontinental level. We look forward to resolving finer-scale ancestries in the future as reference panels supporting such analyses continue to expand.

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. 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/). We hope this data can be used to enhance the scientific community's ability to clinically interpret variation for those of all genomic backgrounds.

388 **MAIN REFERENCES**

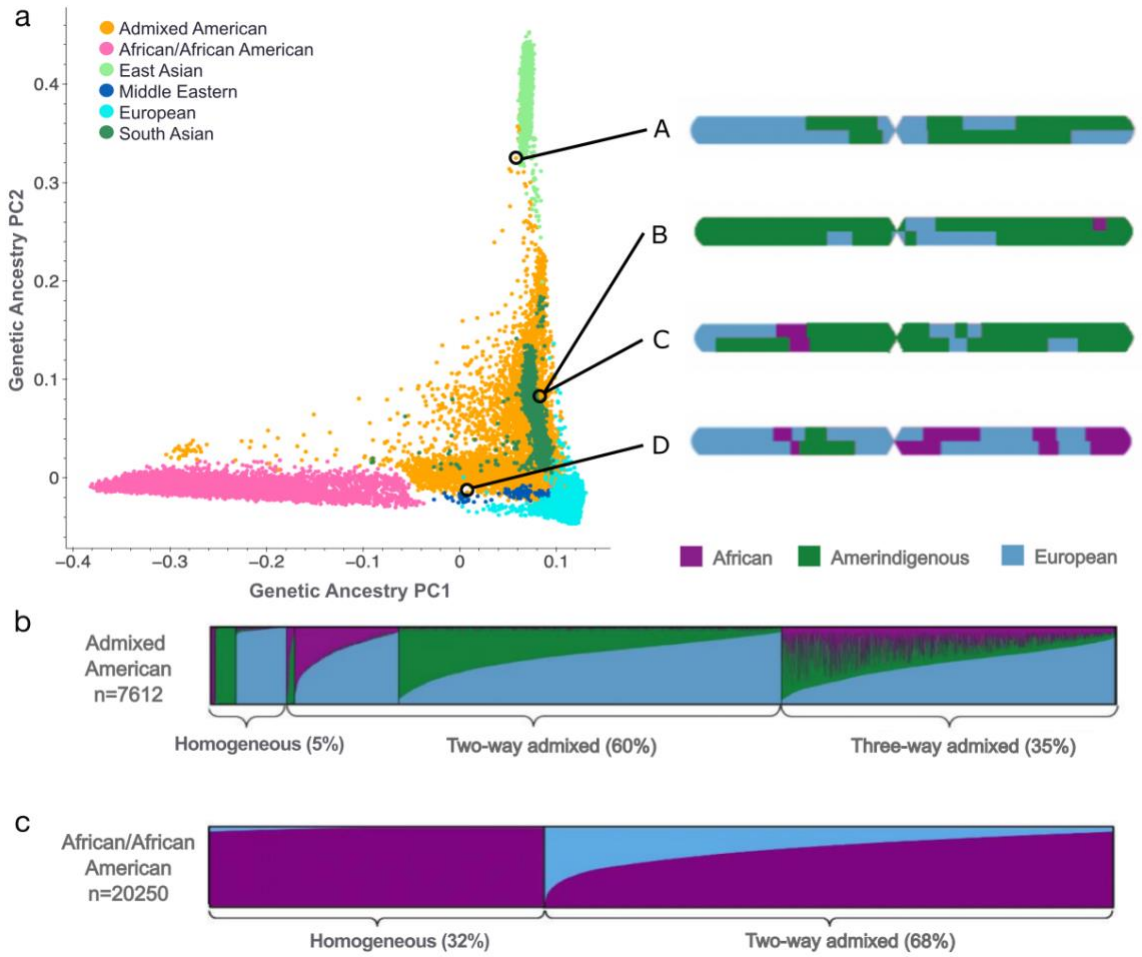
389 1. Chen, S. *et al.* A genomic mutational constraint map using variation in 76,156 human
390 genomes. *Nature 2023 625:7993* **625**, 92–100 (2023).
391 2. Gudmundsson, S. *et al.* Variant interpretation using population databases: Lessons from
392 gnomAD. *Hum Mutat* **43**, 1012 (2022).
393 3. Lin, M., Park, D. S., Zaitlen, N. A., Henn, B. M. & Gignoux, C. R. Admixed Populations
394 Improve Power for Variant Discovery and Portability in Genome-Wide Association
395 Studies. *Front Genet* **12**, 673167 (2021).
396 4. National Academies of Sciences, Engineering, and Medicine. 2023. *Using Population*
397 *Descriptors in Genetics and Genomics Research: A New Framework for an Evolving*
398 *Field*. Washington, DC: The National Academies Press. https://doi.org/10.17226/26902.
399 5. Peterson, R. E. *et al.* Genome-wide Association Studies in Ancestrally Diverse
400 Populations: Opportunities, Methods, Pitfalls, and Recommendations. *Cell* **179**, 589–603
401 (2019).
402 6. Hero, J. O., Zaslavsky, A. M. & Blendon, R. J. The United States Leads Other Nations In
403 Differences By Income In Perceptions Of Health And Health Care. *Health Aff (Millwood)*
404 **36**, 1032–1040 (2017).
405 7. Williams, D. R., Priest, N. & Anderson, N. B. Understanding associations among race,
406 socioeconomic status, and health: Patterns and prospects. *Health Psychol* **35**, 407–411
407 (2016).
408 8. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health
409 disparities. *Nature Genetics 2019 51:4* **51**, 584–591 (2019).

410 9. Zhang, J. & Stram, D. O. The Role of Local Ancestry Adjustment in Association Studies
411 Using Admixed Populations. *Genet Epidemiol* **38**, 502 (2014).
412 10. Sohail, M. *et al.* Mexican Biobank advances population and medical genomics of diverse
413 ancestries. *Nature 2023 622:7984* **622**, 775–783 (2023).
414 11. Williams Amy, A. L. *et al.* Sequence variants in SLC16A11 are a common risk factor for
415 type 2 diabetes in Mexico. *Nature* **506**, 97 (2014).
416 12. Kopp, J. B. *et al.* APOL1 Genetic Variants in Focal Segmental Glomerulosclerosis and
417 HIV-Associated Nephropathy. *J Am Soc Nephrol* **22**, 2129 (2011).
418 13. Genovese, G. *et al.* Association of trypanolytic ApoL1 variants with kidney disease in
419 African Americans. *Science (1979)* **329**, 841–845 (2010).
420 14. Gonzaga-Jauregui, C. *et al.* Mutations in COL27A1 cause Steel syndrome and suggest a
421 founder mutation effect in the Puerto Rican population. *European Journal of Human*
422 *Genetics* **23**, 342 (2015).
423 15. Padoa, C., Goldman, A., Jenkins, T. & Ramsay, M. Cystic fibrosis carrier frequencies in
424 populations of African origin. *J Med Genet* **36**, 41 (1999).
425 16. Hannah, W. B. *et al.* Using genomic databases to determine the frequency and population-
426 based heterogeneity of autosomal recessive conditions. *Genetics in Medicine Open* **2**,
427 101881 (2024).
428 17. Noe, J. *et al.* Impaired expression and function of the bile salt export pump due to three
429 novel ABCB11 mutations in intrahepatic cholestasis. *J Hepatol* **43**, 536–543 (2005).
430 18. ABCB11 - Dashboard | Genetic Prevalence Estimator.
431 https://genie.broadinstitute.org/dashboard/ENSG00000073734/.

432 19. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation
433 and human phenotype. *Nucleic Acids Res* **42**, (2014).
434 20. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting
435 evidence. *Nucleic Acids Res* **46**, D1062 (2017).
436 21. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: Predicting
437 the deleteriousness of variants throughout the human genome. *Nucleic Acids Res* **47**,
438 D886–D894 (2019).
439 22. Cavalli-Sforza, L. L. The Human Genome Diversity Project: past, present and future.
440 *Nature Reviews Genetics 2005 6:4* **6**, 333–340 (2005).
441 23. Auton, A. *et al.* A global reference for human genetic variation. *Nature 2015 526:7571*
442 **526**, 68–74 (2015).
443 24. Marnetto, D. *et al.* Ancestry deconvolution and partial polygenic score can improve
444 susceptibility predictions in recently admixed individuals. *Nature Communications 2020*
445 *11:1* **11**, 1–9 (2020).
446 25. Ghosh, R., Harrison, S. M., Rehm, H. L., Plon, S. E. & Biesecker, L. G. Updated
447 recommendation for the benign stand-alone ACMG/AMP criterion. *Hum Mutat* **39**, 1525–
448 1530 (2018).
449 26. Tan, T. & Atkinson, E. G. Strategies for the Genomic Analysis of Admixed Populations.
450 *Annu Rev Biomed Data Sci* **6**, 105 (2023).
451 27. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature 2016*
452 *536:7616* **536**, 285–291 (2016).

28. Honorato-Mauer, J. *et al.* Characterizing features affecting local ancestry inference performance in admixed populations. *The American Journal of Human Genetics* **112**, 224–234 (2025).

477 **FIGURES**



478

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

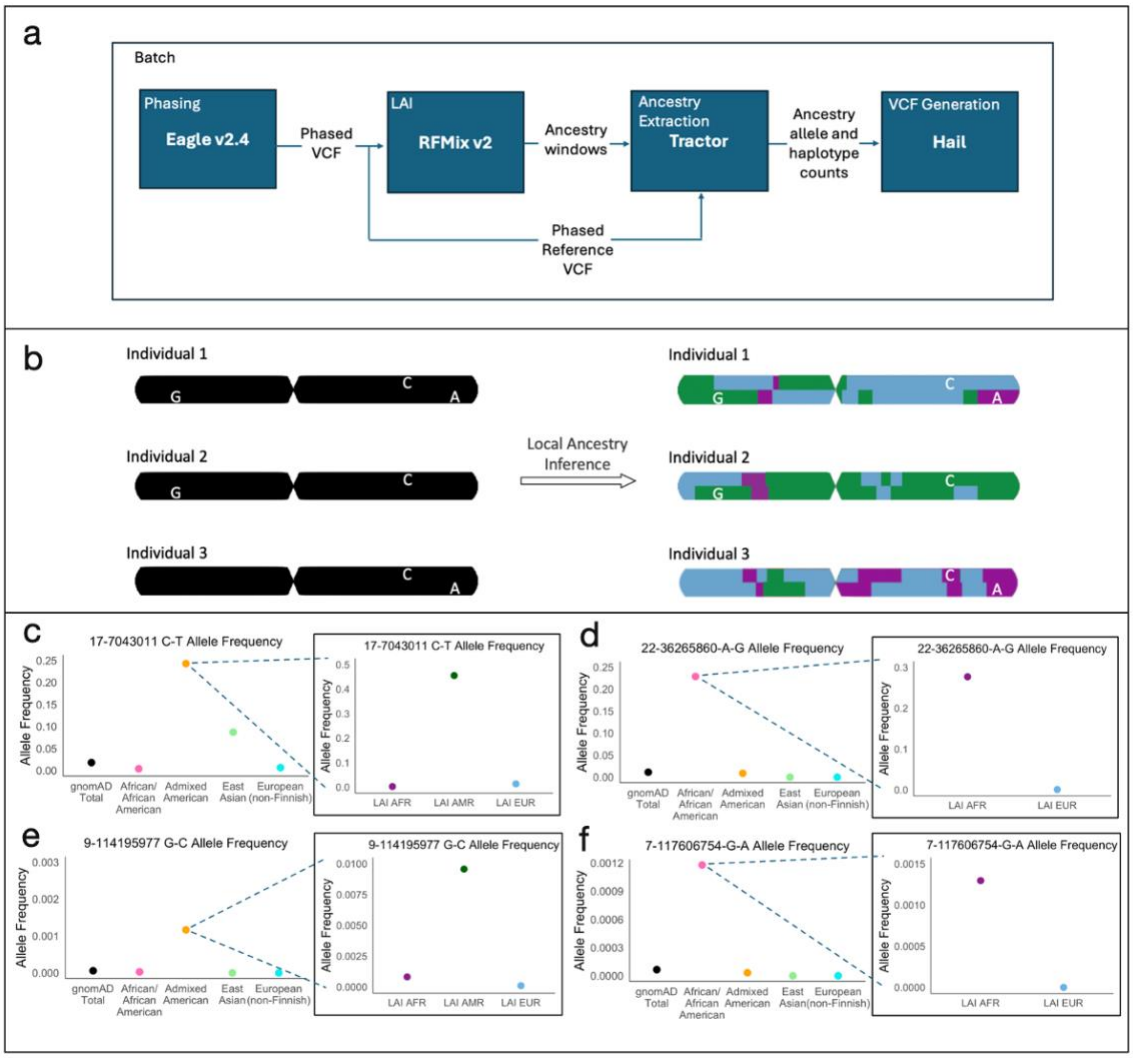


Figure 2: Local ancestry better resolves ancestral allele frequencies. a. Pipeline utilized for LAI frequency data generation**; b.** Schematic illustrating improved frequency resolution thanks to LAI. A single chromosome from three individuals is shown, with non-reference alleles at 3 SNPs (indicated 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; **c.** AF distribution of a *SLC16A11* variant linked to diabetes risk[11]; **d.** AF distribution of an *APOL1* variant associated with increased susceptibility to kidney disease[13]; **e.** AF distribution of a *COL27A1* variant implicated in Steel syndrome[14]; **f.** AF distribution of a *CFTR* variant linked to cystic fibrosis[15].

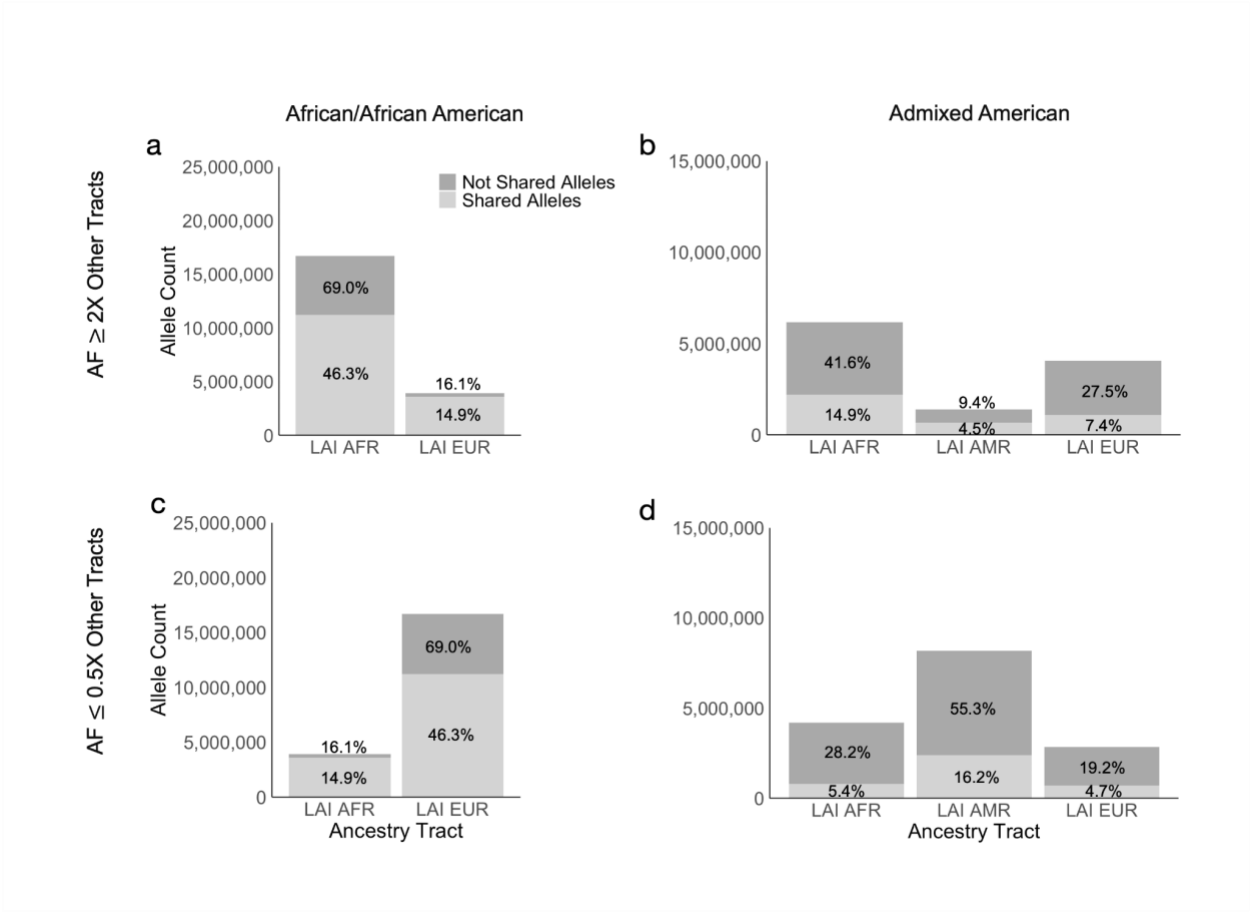


Figure 3: Distribution of variants with enriched or depleted ancestry tract allele frequencies across genetic ancestry groups. The total percentage for each local-ancestry background is indicated on the bars. **a,b.** Variants with ancestry-specific allele frequencies at least 2X higher than the other tracts within the African/African American and Admixed American groups, respectively. **c.d.** Variants with ancestry-specific allele frequencies at most 0.5X lower than the other tracts within the African/African American and Admixed American groups, respectively.

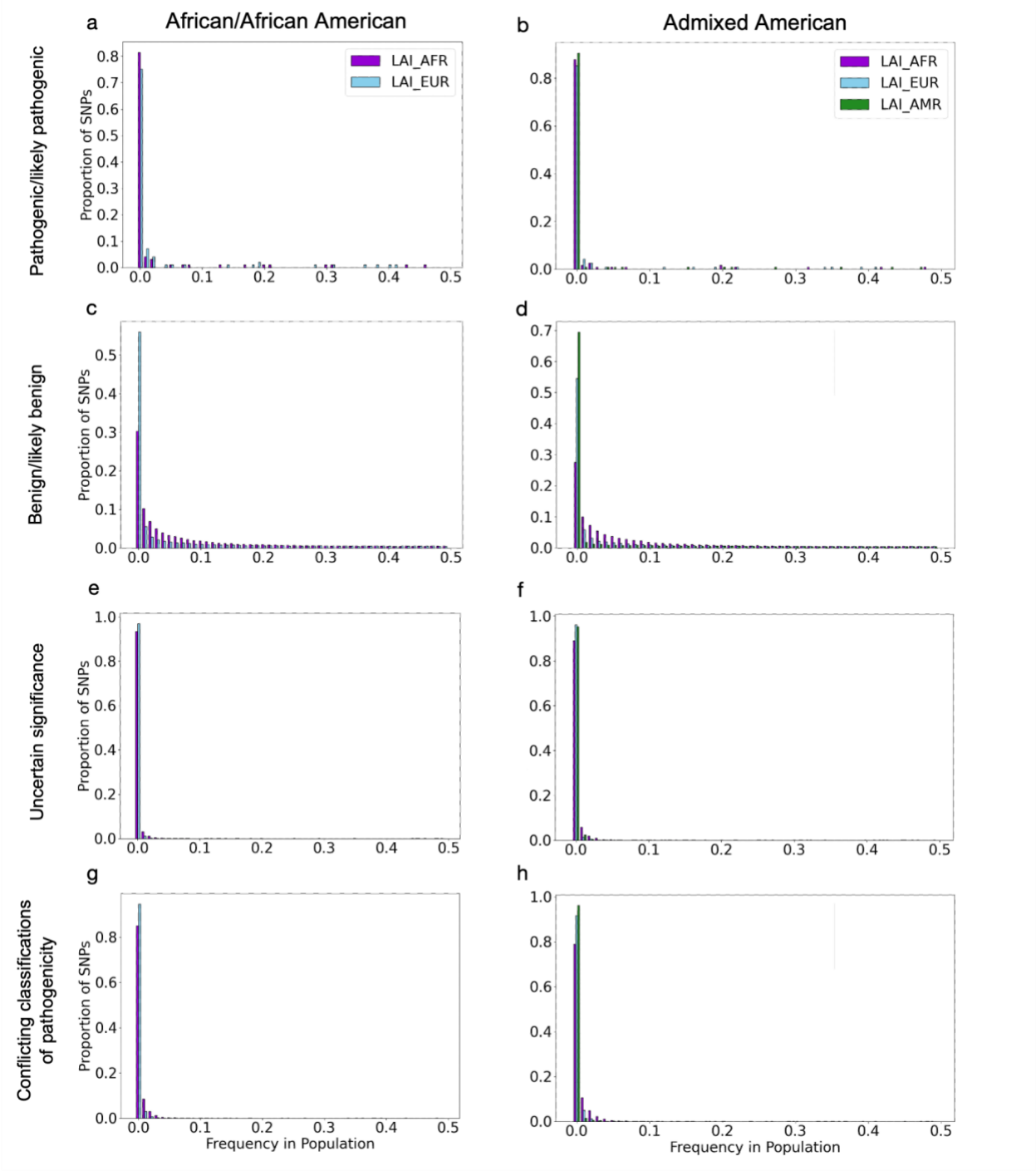


Figure 4: Site frequency spectrum for the inferred African/African American and Admixed American genetic ancestry groups based on ClinVar classifications. LAI-derived AFs for component ancestries are indicated with colors.

510 **METHODS**

511 **Sample selection and quality control**

512 Genetic ancestry group labels were assigned to samples as previously described[1] - briefly,
513 principal components (PCs) computed from hail's hwe_normalized_pca[2] were used as features
514 when training a random forest classifier using Thousand Genomes Project/Human Genome
515 Diversity Sample reference samples with known continental origin (**Fig. 1a**). A continental genetic
516 ancestry group was assigned to all samples for which the probability of that continental genetic
517 ancestry group was > 90% according to the random forest model. Samples were subset based on
518 inferred genetic ancestry groups, limiting to bi-allelic variants with low genotype missingness (i.e.,
519 call rate > 0.9) and an allele frequency greater than 0.1% within the Admixed American and
520 African/African American gnomAD genetic ancestry groups. Additionally, an allele count (AC)
521 threshold of AC > 7 was applied for each genetic ancestry group. The 7,612 Admixed American
522 and 20,250 African/African American whole genome sequencing samples from gnomAD v3.1
523 used here include all individuals in the database except those from the 1000 Genomes Project
524 (1kGP) and Human Genome Diversity Project (HGDP) joint callset[3], which were used to curate
525 reference panels for LAI in each genetic ancestry group.

526

527 **Global ancestry deconvolution**

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

533 superpopulations). For the African/African American genetic ancestry group, we set k=2 (using
534 just continental reference samples from the AFR and EUR superpopulations). In this work, we
535 focus on two admixed genetic ancestry groups represented in gnomAD: the African/African
536 American and Admixed American groups. The underlying composition of the genomes of
537 individuals from these groups contain complex mosaics of ancestries shaped by recent admixture
538 events involving African, European, and Indigenous American sources[5–7]. African/African
539 American individuals typically carry a majority of West African ancestry, with variable
540 proportions of European and, to a lesser extent, Indigenous American ancestry. Admixed
541 American individuals generally derive their genomes from a mixture of Indigenous American,
542 European (primarily Iberian), and African ancestries, with substantial regional and individual-level
543 variation. These groups have historically been underrepresented in genomic studies, and their
544 complex ancestry patterns make them particularly well-suited for local ancestry-informed
545 analyses.
546
547 **Local ancestry inference pipeline**
548 To infer local ancestry data in the gnomAD Admixed American and African/African American
549 group, we built an open-source LAI pipeline. **Fig. 2a** and **2b** shows a schematic of the software
550 pipeline, which includes four primary steps: phasing, LAI, ancestry extraction/dosage calculation,
551 and LAI-informed allele frequency generation. We used Eagle[8] for separately phasing the
552 Admixed American reference panel and for joint phasing of the African/African American group,
553 as it is computationally tractable at scale[8]. We used RFmixV2[9] for LAI, which requires a reference
554 panel representative of the continental ancestries present within the samples. Our reference panel
555 was composed of the harmonized HGDP and 1kGP3 "AMR" samples to capture the

556 Amerindigenous genetic components in the Admixed American group, "EUR" samples for the
557 European components, and "AFR" samples to capture continental African genetic ancestries.
558 RFmixV2[9] was run with one EM iteration for the Admixed American samples to account for the
559 high amount of admixture present in the three-way reference. For the curated reference for the
560 African/African American group, we excluded the Americans of African Ancestry in Southwest
561 USA (ASW) and African Caribbeans in Barbados (ACB) due to their higher level of admixture.
562 Given the substantially larger sample size of the African/African American genetic ancestry group
563 in gnomAD, reducing admixture within the reference panel and therefore removing the need for
564 an EM iteration dramatically reduces the required computing time.
565
566 To extract ancestral segments and calculate ancestry-specific minor allele dosages, we used the
567 tool Tractor[10]. Due to computational constraints, samples from the inferred African/African
568 American genetic ancestry group were run in batches of 6,750 samples for chromosomes 1 and 3-
569 12, and batches of 3,375 samples for chromosome 2 (**Supplementary Table 4**) for LAI. Since the
570 same reference panel was used across batches, this is not expected to impact results. Finally, we
571 tabulated local ancestry-informed allele frequencies using a custom Python script that utilizes the
572 Hail library[2]. To ensure consistency with the most up-to-date dataset available to users, we report
573 total allele frequencies and genetic ancestry group AFs from gnomAD v4.1.0 for all example
574 variants. While LAI calls remain unchanged between gnomAD versions, the aggregate allele
575 frequencies (both total and stratified by genetic ancestry group) have been updated in v4.1.0 due
576 to the inclusion of additional diverse samples. To evaluate the pipeline's sensitivity, we applied it
577 to a simulated admixed American truth dataset and computed the true positive rates of LAI[11]
578 (**Supplementary Fig. 8**). We also validated our pipeline by comparing global ancestry proportions

579 estimated by ADMIXTURE to those from our LAI pipeline results (**Supplementary Fig. 9**). The
580 entire pipeline is unified using the Hail Batch Python module for scalability and is available at
581 [github.com/broadinstitute/gnomad_local_ancestry].
582
583 **DATA AVAILABILITY**
584 Data can be downloaded from or browsed interactively at [gnomad.broadinstitute.org].
585
586 **CODE AVAILABILITY**
587 Code is freely available at [github.com/broadinstitute/gnomad_local_ancestry], with a detailed
588 README outlining the LAI pipeline and step-by-step instructions for running it.
589
590 **METHODS REFERENCES**
591 1. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in
592 141,456 humans. *Nature 2020 581:7809* **581**, 434–443 (2020).
593 2. Hail | Genetics. https://hail.is/docs/0.2/methods/genetics.html.
594 3. Koenig, Z. *et al.* A harmonized public resource of deeply sequenced diverse human
595 genomes. *Genome Res* **34**, 796–809 (2024).
596 4. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in
597 unrelated individuals. *Genome Res* **19**, 1655 (2009).
598 5. Gravel, S. *et al.* Reconstructing Native American Migrations from Whole-Genome and
599 Whole-Exome Data. *PLoS Genet* **9**, e1004023 (2013).
600 6. Ongaro, L. *et al.* The Genomic Impact of European Colonization of the Americas.
601 *Current Biology* **29**, 3974-3986.e4 (2019).

602 7. Homburger, J. R. *et al.* Genomic Insights into the Ancestry and Demographic History of
603 South America. *PLoS Genet* **11**, e1005602 (2015).
604 8. Loh, P. R., Palamara, P. F. & Price, A. L. Fast and accurate long-range phasing in a UK
605 Biobank cohort. *Nature Genetics 2016 48:7* **48**, 811–816 (2016).
606 9. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: A Discriminative
607 Modeling Approach for Rapid and Robust Local-Ancestry Inference. *Am J Hum Genet* **93**, 278
608 (2013).
609 10. Atkinson, E. G. *et al.* Tractor uses local ancestry to enable the inclusion of admixed
610 individuals in GWAS and to boost power. *Nat Genet* **53**, 195–204 (2021).
611 11. Honorato-Mauer, J. *et al.* Characterizing features affecting local ancestry inference
612 performance in admixed populations. *The American Journal of Human Genetics* **112**, 224–234
613 (2025).
614

615 **ACKNOWLEDGMENTS**

616 We wish to thank Alicia Byrne, Anna Lewis, and Alham Saadat for literature review; Jackie
617 Goldstein, Dan King, and Ben Weisburd for pipeline scalability assistance; Marcos Santoro for
618 providing LAI reference panel guidance; Julia Goodrich for pipeline code review. This work was
619 funded by the National Institutes of Health (K01 MH121659 and R01 HG012869 to E.G.A.; P.K.
620 was supported by T32 GM139534-03 and T32 GM136554-05).
621

622 **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.