Characterizing substructure via mixture modeling of genetic similarity in large-scale summary statistics

Authors: Hayley R Stoneman*[1,2], Adelle Price*[1,4], Nikole Scribner Trout[1,4], Riley Lamont [4], Souha Tifour[1,4], Nikita Pozdeyev[3,5], Kristy Crooks[3], Meng Lin[3], Nicholas Rafaels[3], Katie M Marker[1,2,3], Christopher R Gignoux[1,2,3], Audrey E Henrdicks[1,2,3,4]

Affiliations:
[1]Department of Biomedical Informatics, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, USA
[2]Human Medical Genetics and Genomics Program, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, USA.
[3]Colorado Center for Personalized Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, USA
[4]Mathematical and Statistical Sciences, University of Colorado Denver, Denver, CO 80204, USA
[5]Division of Endocrinology, Diabetes and Metabolism, Department of Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, USA
*These authors contributed equally to the manuscript

Corresponding Author: Audrey Hendricks

Abstract
Genetic summary data are both broadly accessible and highly useful including for risk prediction, causal inference, fine mapping, and the incorporation of external controls. Nevertheless, collapsing individual-level data into groups masks intra- and inter-sample heterogeneity (e.g., population structure), leading to confounding, reduced power, and bias. Unaccounted for substructure limits summary data usability, especially for understudied or admixed populations. Here, we present *Summix2*, a comprehensive set of methods and software based on a computationally efficient mixture model of genetic similarity to estimate and adjust for substructure in genetic summary data. In extensive simulations and application to public data, *Summix2* provides accurate and precise estimates detecting regions of selection, identifying fine-scale population structure, and disentangling latent genetic disease risk. *Summix2* increases the robust use of diverse publicly available summary data resulting in improved and more equitable research to address population structure, disease etiology, and risk prediction.

Introduction
Leveraging genomic summary data enables advances in fundamental research and clinical translation. Summary data, such as Allele frequencies (AFs) or Genome Wide Association Study (GWAS) test statistics, are used to prioritize putative causal genetic variants, and as external common controls, for causal inference via Mendelian randomization and for risk prediction using polygenic scores (PGS). Compared to individual level data, summary data is easier to access, store, and analyze while protecting person's data confidentiality.

This enables broad reuse and data-sharing and subsequently great statistical power for genetic discovery. New technologies, such as federated learning frameworks that leverage summary statistics or weights to update centralized AI models, hospital-based biobanks with genetic data, and an increase in

other genomic summary data types (e.g., methylation, expression) position the use of genomic summary data for exponential growth.

However, broad, rigorous, and representative use of summary data is limited by unaccounted for substructure within and between datasets. Indeed, issues with the use of summary data are often most pronounced for ancestrally admixed and diverse populations. While efforts are actively underway to collect more representative samples including those from admixed populations (e.g., African American, Latinx), the majority of genetic research thus far has been performed in groups of European ancestry and is not fully generalizable to diverse or admixed populations[1-5]. Additionally, existing admixed data are often underutilized due to barriers in completing high-quality research using summary data with genetic substructure[6,7]. The 2023 National Academy report on Population Descriptors in Genetics and Genomics Research recommends the use of genetic similarity (i.e., quantitative measures of shared genetic ancestry) to prevent confounding in most genetics and genomics research while also recognizing that estimating genetic similarity is often not possible with summary data[8]. Thus, researchers or clinicians working with admixed or underrepresented ancestral groups often face an unfair choice: choose the best (but poorly matched) public summary data and produce potentially biased results, or do not complete the study at all. The shear existence of this choice for only a subset of people exacerbates inequities in research and, ultimately, precision medicine and healthcare derived from the research. Importantly, a comprehensive understanding of fundamental biological processes is also constrained by the incomplete use of diverse and representative data[9-11].

There are numerous methods to detect global (i.e., average across the genome) and local (i.e., windows across the genome) ancestral substructure in admixed populations using individual level data. Global ancestry estimation methods include principal component analysis[12], ADMIXTURE[13], and identity by descent (IBD) clustering methods[14-16]. While both global and local ancestry estimation methods can be used to adjust for population structure, studies suggest that global ancestry adjustment is usually optimal to control type I error while retaining statistical power, and local ancestry adjustment is better at improving fine-mapping and accuracy of effect size estimation[17-20]. In addition, local ancestry methods, such as LAMP-LD[21], RFMix[22], ELAI[23] and Loter[24], can be used to identify signatures of post-admixture selection. Putative regions of selection have been identified in many studies, including the MHC locus, which is known to play a role in immune response and subsequently auto-immune diseases[23,25-32].

While many methods exist to detect and account for substructure in individual level data, few methods exist for summary level data. In 2021, we developed *Summix*, a computationally efficient mixture model employing AFs to estimate and adjust for continental level ancestry substructure in genetic summary data[33]. In 2022, Privé presented a similar method in the *bigsnpr* package that minimizes AF projections onto a principal component (PC) space[34]. Privé's method uses individual level reference data from UK Biobank data for 5,816,590 variants to create a reference PC space on which AFs can be projected, enabling identification of finer-scale ancestral substructure. Conversely, *Summix* does not require individual level data and can estimate substructure with far fewer variants (e.g., ~100). The ability to estimate substructure with a small subset of genetic variants opens the possibility of additional functionalities such as estimation of local ancestry and genetic predisposition to disease as presented here.

We present *Summix2*, a suite of methods to detect and leverage substructure in genetic summary data. Advances include estimating finer-scale substructure and local ancestry across the genome, an effective sample size for adjusted AF, generalizing the goodness-of-fit (GoF) metric for any array and sequencing technology, and estimating other genetically driven substructure such as the genetic predisposition to conditions. Substructure estimates are leveraged to produce adjusted AFs and to identify regions where local ancestry differs from average global ancestry (representing putative regions of selection). In addition, *Summix2* introduces features to improve the user experience by

simplifying inputs, performing input and output data filtering, and outputting warning and error messages to ease troubleshooting. *Summix2* is available as an R package on Bioconductor with open-source code and extensive examples to ensure reproducibility and enable others to use and build on our methods.

Results

Summix2 Overview

Summix[33], developed in 2021, employs an efficient mixture model to estimate subgroup proportions in genetic summary data by minimizing least-squares differences between a vector of observed sample AFs and a vector of AFs generated from a mixture of reference data groups (**Figure 1a**). *Summix* can then harmonize the observed AF subgroup proportions to match an individual or sample using the estimated mixing proportions of the reference data groups. *Summix2* has significantly expanded capabilities including a generalizable GoF metric across different data generation technologies, an effective sample size for adjusted AF, estimation of local-substructure and where it differs from global substructure, and generation of a finer-scale genetic similarity map for improved harmonization (**Figure 1a**). Additionally, we have assessed *Summix2*'s ability to estimate and harmonize across genetically driven substructures beyond ancestry.

Experiment Overview

We evaluated *Summix2* using AF data from gnomAD v3.1.2[35] as the observed summary data. This data includes whole genomes from 76,156 individuals and ten gnomAD groups (i.e., African/African American (AFR/AFRAM), Amish (AMI), Latino/Admixed American (AMR), Ashkenazi Jewish (ASJ), East Asian (EAS), Finnish (FIN), Non-Finnish European (NFE), Middle Eastern (MID), South Asian (SAS), and Other (OTH)) (**Extended Data Table 1**). For reference AF data, we used 73 finer-scale ancestry groups defined by the 1000 Genomes (1KG)[36] and Human Genetic Diversity Project (HGDP)[37] data that was harmonized and released with gnomAD v3.1.2 (**Extended Data Table 2**)[35]. We removed 21 admixed reference groups (i.e., containing more than one continental reference ancestry)[38], one group not in the five continental groups used (HGDP Kalash), and one with N=1 (HGDP Surui). Our final reference dataset contained 50 groups. Using gnomAD v3.1.2 grouping assignments, five continental ancestry reference groups (i.e., African (AFR), Non-Finnish European (EUR), East Asian (EAS), Indigenous American (IAM), and South Asian (SAS)) were created using a weighted average of the finer-scale ancestry groups (**Extended Data Table 2**).

We used simulated observed and reference AF data to assess the minimum accuracy and precision of substructure estimates, absolute and relative difference of AF adjustments, as well as type I error and power to detect regions where local substructure differs from global. Simulated AF data was generated using a multinomial distribution, with reference AFs as the input probabilities, and mixing proportions to generate observed data with substructure (e.g., admixture) with a default sample size of 10,000. Reference data was simulated without mixing proportions with a default sample size of 500 individuals per group for continental-like substructure and 100 for finer-scale substructure. For local substructure simulations, linkage disequilibrium (LD) filtered (r^2<0.2) chromosome 19 variants and 100 simulation replicates were used for each simulation scenario to assess minimum accuracy and precision across local substructure blocks over a variety of simulation parameters including mixing proportions, level of substructure (i.e., global vs local, broad vs. finer-scale), reference sample size, observed sample size, and window size (for local substructure estimation). For type I error and power to detect local substructure, we used 100 replicates where we simulated a region between 50kb and 1Mb on chromosome 19 with different substructure for power, and the remainder of the chromosome to match the global structure on average for type I error. We assessed type I error and power of *Summix-local*

across a variety of simulations in AFR/AFRAM-like and AMR-like samples varying the observed and reference sample sizes, magnitude and size of differing region, and window size.

We estimated the pairwise genetic similarity between the reference groups, then designed simulations investigating the minimum accuracy and precision of *Summix2* genetic similarity proportion estimates across references for varying degrees of similarity (as estimated by the Hudson estimator for pairwise F_{ST}) between finer-scale ancestries. For simulations of finer-scale substructure a random subset of 100,000 genome-wide variants was used across 100 simulation replicates. Full details are in **Online Methods**.

Generalized Goodness-of-Fit Metric

The GoF metric for the original Summix method was the sum of the least squares (LS) objective function per 1000 SNPs and was built and evaluated using gnomAD v2.1 whole-exome data[33]. A value below 0.5 indicated good fit, between 0.5 and 1.5 moderate fit, and above 1.5 poor fit. Through application to gnomAD v3.1.2 whole genome sequencing data[35], we identified that the original GoF metric is sensitive to minor allele frequency (MAF). Through simulations, we found that while the accuracy was stable over MAF bin, the original GoF metric (i.e., LS objective per 1000 SNPs) increased as MAF bin increased (**Extended Data Figure 1**). To remove dependency between the GoF metric and MAF distribution, we weighted the LS loss by MAF bin (bins: (0, 0.1], (0.1, 0.3], (0.3-0.5], weighted by 5, 1.5, 1, respectively). In simulations leaving out a reference ancestry group from the reference panel, we found that the new GoF measure was <0.5 when ≤10% of a reference ancestry group was missing, and was >1.5 when ≥20% reference ancestry was missing for some of the groups (AFR and some IAM). Thus, we propose to use the same thresholds from the original GoF metric: <0.5 good fit, 0.5-1.5 moderate fit, and >1.5 poor fit (**Extended Data Figure 2**). Additional details are in **Online Methods**.

Overview of Summix-local Method

By scanning regions of the genome, *Summix2* can be used to detect local substructure as well as regions where the local substructure differs from the average substructure across the genome (i.e., global). We employed two algorithms for scanning the genome: a classic sliding windows approach, and a fast catch-up algorithm which allows dynamic window sizes[39]. We developed a Wald test statistic to identify regions where local substructure differs from average substructure across the chromosome. Standard error was estimated as an average of the simulated within block standard error and observed between block standard deviation. (See **Online Methods** for details.)

Summix-local produces accurate and precise local substructure estimates

Summix-local produced local ancestry estimates with high accuracy and precision in simulated Latinx and African-American admixed populations. The minimum accuracy across blocks was high (>99.5%) across all simulation scenarios with moderate to large reference group sample sizes (N≥500), all observed sample sizes evaluated (N≥1000), and all window sizes evaluated (≥250 genetic variants). For simulations with small reference sample sizes (N=100 per group), the minimum accuracy decreased slightly (98%) (**Figure 2a**). Minimum precision across all simulations was high (>98%); and as expected, any reduction in precision was seen in simulations with smaller sample sizes and smaller window sizes (**Extended Data Figure 3**). We also assessed the impact of LD on local substructure estimates. LD did not impact accuracy but did result in a lower number of effectively independent single nucleotide polymorphisms (SNPs) per window, resulting in an artificially small standard error (**Extended Data Figure 4, Supplementary Note 1**).

Ability to detect regions where local genetic substructure differs from global

The type I error rate was far below the expected 5% rate for most simulation scenarios only approaching or exceeding 5% for small reference or observed sample sizes (N=100 and N=1000 respectively) in the AMR-like sample (AMR-like: **Figure 2**, AFR-like: **Extended Data Figure 5**, **Extended Data Table 3-4**). While both scanning algorithms had well controlled type I error and high power for the majority of simulation scenarios, the fast catchup algorithm had consistently lower type I error and higher power compared to the sliding windows algorithm.

Summix-local maintained high power (>80%) to detect regions with different local substructure proportions given sufficiently large sample sizes (observed N≥10000, reference N≥500), regions (≥ 400kb), and magnitude of difference (≥3%). Smaller reference sample sizes (N=100), regions (50Kb), or signals (2% difference in local substructure) reduced power to near zero (**Figure 2**). While also moderately reduced, power was more robust to smaller observed sample sizes (N=1000) staying above 60%.

Local-substructure scan in gnomAD

Using *Summix-local* with the fast-catchup algorithm and default parameters (**Online Methods**), we scanned gnomAD v3.1.2 AFR/AFRAM and AMR groups for regions where local ancestry differed from average chromosome global ancestry (**Figure 3, Table 1**). First, we assessed whether we found differential local structure for 16 and 29 regions of selection previously identified in AFR/AFRAM or AMR samples respectively [23,26,28,29,40-43]. HLA was previously identified in both groups. Significance was defined as passing a Bonferroni-corrected p-value calculated for AFR/AFRAM and AMR separately using 16 and 29 regions and two or three reference groups, respectively. In both groups we replicated the *HLA* locus on chromosome 6 (6p21-23) (AFR/AFRAM *P* = 6.50e-11; AMR *P* = 1.21e-5). The *HLA* gene cluster has been widely studied as a region known to have functional properties under selection in both ancient and recent human populations[44]. Additionally, in AMR, we replicated a signal on chromosome 1 (1p33) containing *CYP4A11* (*P* = 5.49e-4), which is a member of the cytochrome P450 superfamily of enzymes involved in drug metabolism[42]. Loci not replicated could be due to numerous factors including differences in ancestral populations or selection pressures.

In a genome-wide scan, we detected nine and eight significant blocks representing eight and seven regions in the AFR/AFRAM and AMR groups respectively. Genome-wide significance was defined by a Bonferroni-correction for 5945 blocks and the number of reference groups (two and three, respectively) resulting in a threshold of 4.21e-6 and 2.80e-6, respectively for AFR/AFRAM and AMR (**Table 1**). Two of the loci, 1q42.13 and 2p11.2 (AFR/AFRAM *P* = 2.18e-8, 1.19e-6; AMR *P* = 5.11e-8, 1.48e-6, respectively), were identified in both the AFR/AFRAM and AMR groups. The locus at 1q42.13 contains the gene *PSEN2*, which has been associated with Alzheimer's disease[45]. The locus at 2p11.2 contains the immunoglobulin kappa variable cluster (*IGKV*), which is involved in immune response and has been associated with breast cancer[46]. The 9p21.3 locus was identified in AMR (*P* = 6.77e-8) and contains a cluster of alpha interferon genes, *IFNA* and *MIR31HG*. These alpha interferons are a key part of the innate immune response. *MIR31HG* codes a long non-coding RNA that has been associated with cancers including pancreatic, lung, and bladder, as well as autoimmune disease[47-50]. The 11p11.2 locus was identified in AFR/AFRAM (*P* = 1.80e-9) and contains a block of *TRIM* genes, which are predicted to be involved in protein and ion binding and the innate immune response. There have been associations detected for SNPs in these genes with glomerular filtration rate and chronic kidney disease, and BMI[51,52]. When examining the Gene Ontology pathways of genes in the significant blocks, we found five regions in AFR/AFRAM and three in AMR with genes in pathways for immune response[53,54].

Detecting finer-scale genetic substructure

We estimated the genetic similarity of the 50 finer-scale and five continental ancestry groups using the Hudson estimator for pairwise F_{ST}[55] (**Figure 4**). Although pairwise F_{ST} estimates were generally greater, indicating less similarity, between continental groups (ranging from 0.026 to 1.31, **Extended Data Table 5**) compared to finer-scale groups (0.001 to 0.035), there were notable exceptions such as more similarity between EUR and SAS continental groups (F_{ST}=0.026) and less similarity within certain finer-scale ancestry groupings such as bantukenya and gwd (F_{ST}=0.032) within the AFR continental ancestry grouping (**Extended Data Table 6**).

We evaluated the ability of *Summix2* to accurately and precisely estimate finer-scale substructure proportions as the pairwise genetic similarity between simulated finer-scale ancestries increases, i.e., pairwise F_{ST} decreases. We simulated three mixture proportion scenarios (i.e., equal, varying, and one zero) across three pairwise F_{ST} scenarios from the EAS continental ancestry group (pairwise F_{ST}= 0.009, 0.007, 0.005) and estimated mixing proportions using the 50 finer-scale ancestry reference panel.

Minimum accuracy across all simulated reference groups decreased as simulated finer-scale ancestry groups similarity increased (i.e., F_{ST} decreased): 97% for F_{ST}=.009, 91% for F_{ST}=.007, and 87% for F_{ST}=.005 across the 'Varying ancestry proportions' simulation scenarios (**Figure 5a**). The minimum precision of *Summix2* estimates was consistently high, above 99%, for all scenarios (**Extended Data Table 7**). Importantly, in all simulation scenarios, bias in the *Summix2* mixture proportion estimates was captured by similar subgroups within the larger group of continental ancestry as can be observed by the genetic similarity map of the reference group *Summix2* proportion estimates (**Extended Data Figure 6**).

As a resource to the community, we also provide finer-scale substructure proportion estimates detected for seven gnomAD v3.1.2 groups with good or moderate GoF (**Extended Data Table 8**). Amish, Ashkenazi Jewish, Middle Eastern substructure estimations were excluded due to GoF (90% Confidence Interval (CI)) = 6.34 (6.16, 6.55), 2.15 (2.07,2.24), and 2.39 (2.32, 2.46) respectively.

Adjusting for finer-scale genetic substructure using genetic similarity mapping

We next investigated using the estimated finer-scale substructure proportions and genetic similarity map to adjust AFs in the observed sample to match a target sample, as might be done when harmonizing summary datasets. We simulated an AFR/AFRAM-like observed group and evaluated bias in the *Summix2* finer-scale mixture proportion estimates and the resulting genetic similarity mapping between the reference groups (**Figure 5b,c**). We then adjusted the AF for the simulated AFR/AFRAM-like observed group to a target simulated AFR-like group comparing unadjusted, continental level adjustment, and finer-scale level adjustment (simulated substructure in **Extended Data Table 9**). Even though the finer-scale substructure estimates were biased (**Figure 5b**), the finer-scale AF adjustment outperformed (i.e., lower bias) the continental level AF adjustment for genetic substructure. Notably, the bias in AF adjusted estimates was similar when using the finer-scale substructure estimates and known substructure proportion parameters indicating that bias in the substructure estimates was not propagated to the AF adjustment likely due to the bias being distributed to other similar reference groups as seen in the genetic similarity map (**Figure 5c,d**).

Improved data harmonization through adjusted allele frequencies

To demonstrate the utility of *Summix2* for data harmonization with varying levels of substructure we adjusted gnomAD AMR AFs to match the ancestry substructure for three target samples: 1KG Peruvian (PEL), a simulated Peruvian-like sample with continental-like substructure (10% AFR-like, 15% EUR-like, 75% IAM-like), and a simulated target with finer-scale-like substructure and PEL-like proportions (45% HGDP-maya-like, 10% HGDP-karitiana-like, 10% HGDP-pima-like, 10% HGDP-colombian-like, 5% 1KG-esn-like, 5% 1KG-yri-like, 8% 1KG-gbr-like, 7% 1KG-ibs-like), further details can be found in the **Online Methods**. We compared the unadjusted AF, four *Summix2* AF adjustments

(global continental, global finer-scale, local continental, and local finer-scale), and two adjustments using global and local ancestry proportions generated by gnomAD v3.1 using individual level data. To enable a comparison with gnomAD, we used 12, 513 variants on chromosome 19 with local-ancestry estimation from gnomAD and MAF > 1%.

All adjusted AFs had lower absolute and relative bias compared to the unadjusted AFs for all three targets (**Figure 6**, **Extended Data Figures 7-9**). For *Summix2*, AF harmonization was also available for a much larger set of 234,633 variants on chromosome 19 (**Extended Data Figure 10**). We observed a larger reduction in bias when adjusting to the simulated target data compared to adjusting to the real 1KG PEL data. This additional reduction in bias is likely due to all of the substructure being represented by the reference data in the simulated datasets as indicated by better GoF estimates for the simulated data (continental-like simulation: continental GoF=0.0995, finer-scale GoF=0.6728; finer-scale-like simulation: continental GoF=1.2226, finer-scale GoF=0.1883) compared to the real 1KG PEL data (continental GoF=5.4877, finer-scale GoF=5.0798). For the simulated target data, we also compared using the known simulated proportions to *Summix2* estimated proportions for the adjustment and found the results highly concordant (**Extended Data Figure 10**). When the underlying simulated structure was continental-like, using continental references for the adjustment performed best, producing the smallest differences between observed and target, and the highest concordance, measured by Lin's Concordance Correlation Coefficient(CCC)[56](**Figure 6**, **Extended Data Figure 8**). When the underlying structure was finer-scale-like, finer-scale AF adjustment performed best.

Whereas the local and global adjustments generally had similar performance, the local finer-scale adjustment had higher variability and bias compared to the global finer-scale adjustment for the finer-scale-like simulation only. The gnomAD adjusted AFs had slightly smaller bias when harmonizing to 1KG PEL, but the concordance with the target AFs was minimally improved (largest improvement: CCC = 0.9605 for gnomAD local adjusted vs CCC = 0.9417 for *Summix2* local continental adjusted (**Extended Data Figure 8**). Conversely, the gnomAD adjusted AFs had higher bias than the *Summix2* adjusted when adjusting to a simulated target, likely due to *Summix2* using the same references as were used to create the simulated target.

Estimating genetic predisposition to disease

We investigated the ability of *Summix2* to detect genetic substructure other than ancestry, such as the genetic predisposition to a condition or disease. We used AFs for cases and controls for 138 variants from a 2018 prostate cancer PGS as reference groups to estimate the proportion of people in the Colorado Center for Personalized Medicine (CCPM) Biobank data genetically similar to the prostate cancer cases[57,58]. To match the European ancestry of the published PGS, we limited CCPM data to European individuals and subset the CCPM data by biological sex (i.e., female, male), males with diagnosed with prostate cancer, and age group (i.e., ≤40 years, 41-60 years, ≥61 years) (**Figure 7**). Here we had AF data for 10,092 males and 16,594 females, as defined by biological sex. Of the men, 1317 (13%) had clinical prostate cancer diagnoses, as defined by ICD-9 and ICD-10 codes. CCPM females, who do not have prostates, have 0% observed prostate cancer but still carry genetic risk for prostate cancer. We estimated between 4.4% to 11.6% proportion genetically similar to the prostate cancer case reference group across the age subsets. In CCPM males with prostate cancer we estimated 100% of the sample was genetically similar to reference cases in the subset of men aged 41-60 years. In men ≥61 years old with prostate cancer, the estimate was 86.5%, potentially indicating increased environmental contribution to prostate cancer risk in this group. Interestingly, for all males, the estimated proportion that is genetically similar to cases increased with age, from 1.8% for ≤40 years to 7.2% for 40-60 years, and 15.8% for ≥61 years, mirroring the increase in the observed prostate cancer diagnosis proportions as age increased. Since the proportion of people with a genetic predisposition to disease is not expected to increase with age, this increase is likely to reflect ascertainment bias, where older males are more

likely to have prostate cancer and subsequently come to the University of Colorado for healthcare, where they can be recruited into the CCPM Biobank.

Discussion

Here, we present *Summix2*, a Bioconductor[59] software package and suite of methods to estimate and adjust for substructure in summary-level genetic data. *Summix2* enables estimation of finer-scale genetic ancestry substructure, visualization of the genetic similarity mapping of substructure, estimation and testing of local-substructure, and adjusting AFs to harmonize and control for different genetic substructure between samples. Using prostate cancer as an example, we show that *Summix2* can detect genetically driven substructure beyond ancestry. *Summix2* also includes a new goodness-of-fit measure that is robust to the distribution of AFs in the observed data. These tools hold the promise of improving genetic summary data harmonization and use.

Summix2 introduces a new method to estimate local-ancestry proportions. Through simulations we showed that *Summix-local* can accurately and precisely estimate local genetic substructure proportions and can detect regions where local substructure differs from global substructure. For these simulations we used LD filtered data, due to the observation that failing to filter for LD falsely reduces the standard error of the estimates. We recommend using LD filtered data (e.g., $r^2<0.2$) or explicitly modeling the correlation structure to accurately estimate the effective number of independent SNPs[60]. Applying *Summix-local* to gnomAD v3.1.2 AFR/AFRAM and AMR groups, we replicated one and three out of 16 and 29 previously identified regions of selection and identified putative regions of selection, respectively. Loci may fail to replicate due to several reasons including different underlying ancestral populations or selection pressures. Indeed, the gnomAD AMR sample is a conglomeration of numerous groups with different ancestry, cultures, and histories – including likely different selection pressures. Any signal found here would be likely be strong and robustly observed across most of the subgroups.

We also discovered eight and seven putative novel regions, respectively, at a Bonferroni-corrected genome-wide significance level. Five out of eight in AFR/AFRAM and three out of seven in AMR regions identified contain genes involved in immune response and cancer susceptibility[45-54]. Since *Summix-local* compares expected global substructure to local substructure, it is highly similar to the selection statistic LAD, which can detect post-admixture selection and adaptive admixture[26,61]. Adaptive admixture is the process by which admixed populations acquire beneficial mutations from source populations which facilitates genetic adaptation to different environments. It has been well recorded that genes with important roles in survival and disease phenotypes can be under selection, including cancer susceptibility and immunity. Often the genes under selection affect traits that have different prevalence rates across populations[62]. Recently, several iconic signals of ongoing selection since admixture related to diseases and pathogens were confirmed[61]. While we suggest that these differences in local genetic substructure could be signatures of selection, local ancestry patterns are sensitive to migration, recombination, and random genetic drift, which could have influenced the signals found here[60]. The gnomAD AMR and AFR/AFRAM groups are each a larger grouping of many highly diverse samples from different locations, with 7,647 and 20,744 samples, respectively[35]. For example, the AMR sample has a mix of samples including, but not limited to Mexican, Costa Rican, Colombian, Karitiana, Maya, Puerto Rican, and Peruvian[35]. This results in a single group label being assigned to many groups that may have experienced different histories, admixture events, and environments and therefore different exposures and selective pressures. This also highlights the inadequacy of using Latinx as a single group label, which falsely suggests that all Latinx groups share similar genetic, social, and environmental backgrounds.

Although continental and finer-scale ancestry groupings have commonly been used to detect and control for population structure, genetic ancestry is not discrete[12,13,17-20]. Discrete groups are often

not representative of the true complex history of a population, and use of groups can further exacerbate existing inequities. A recent review by Lu *et al.* summarized the lack of consistent and appropriate use of race, ethnicity, and ancestry in genomics research[63]. They also show many examples where these group labels harmfully perpetuate beliefs that racial difference reflect biological differences. This can manifest in the differential treatment in the healthcare system due to internalization of racial stereotypes. The 2023 National Academies Report on Population Descriptors reccomends using genetic similarity to represent human population structure rather than discrete groupings[8]. When using individual level data, genetic similarity can be captured using principal components and variance component terms[12,13]. However, there has not been a standard method to model genetic similarity in summary data. Here, we enable genetic similarity mapping of genetic summary data using mixture modeling to a reference map of finer-scale population references. Many of the finer-scale references are highly similar genetically (F_{ST} < 0.01). Interestingly and importantly, while the genetic similarity of the finer-scale references results in biased mixture proportion estimates, the bias is contained locally to other similar reference populations within the genetic similarity map. Consequently, the bias does not propagate through adjustment and harmonization of the AFs. This shows *Summix2* is able to estimate and map the shared genetic information between similar references to accurately adjust AFs. Indeed, the locally contained bias in mixture estimates further highlights that discrete groupings for genetic ancestry is not correct.

Additionally, to highlight the ability of *Summix2* to identify other types of substructure, we estimated the proportion of people in the CCPM biobank who were genetically similar to prostate cancer cases. In CCPM males, we found the proportion of people genetically similar to prostate cancer cases increased with age. While this mirrors the increase in the observed proportion of men diagnosed with prostate cancer by age, barring evolutionary pressure or bias, we do not expect the proportion of people with a genetic predisposition to a condition to change with age. We do not anticipate evolutionary pressure given the short time frame and survival bias would result in a decrease, rather than the observed increase. Thus, the increase is likely due to ascertainment bias. The CCPM Biobank recruits individuals from the University of Colorado healthcare system, which likely results in over-sampling of individuals with health complications, including cancer diagnoses.

While the gnomAD adjustments using individual level data showed less bias than Summix for the real data adjustment to 1KG PEL, the difference was small. Additionally, *Summix2* adjustments were much improved to the unadjusted AF with the advantage of providing adjusted AFs for all variants in the dataset. Here, we find that a genetic similarity map of summary statistics can be used to adjust genetic summary data. We find that more granular or finer-scale subgroups is best for adjustment when the underlying genetic substructure is finer-scale, resulting in the greatest reduction in bias and best harmonization of summary data. When the underlying substructure is continental-like, the use of continental references performs better. These adjustments perform well even when the genetic similarity map contains bias, as long as that bias is distributed to highly similar reference groups. These substructure-adjusted AFs can then be used in secondary analyses.

It is important to note that *Summix2's* substructure estimates are dependent on the reference dataset. As such, adequate representation in the reference data is required for accurate substructure estimation and AF harmonization. The reference data used here are not a full representation of the genetic diversity within each of the continental groups, in particular for African, South Asian, and Indigenous American[64,65]. African populations are known to have the highest genetic diversity, followed by South Asian, amongst all humans and are only represented by 8 and 5 finer-scale groups, respectively, in reference data used here compared to 21 East Asian finer-scale groups and 12 European finer-scale groups[64-68]. Similarly, the Indigenous American group with only 47 samples in total compared to 704 in AFR, 787 in EAS, 741 in EUR, and 545 in SAS (**Extended Data Table 2**), is vastly underrepresented by sample size alone[69]. Many consortia, such as the African Diversity Project, All of Us, the GenomeAsia 100K Project, and the Population Architecture through Genomics and Environment

(PAGE) study, are actively pursuing efforts to improve the representation of genetic data from non-European groups[1-4]. While additional representation from diverse ancestries, cultures, and environments is vital to generalizable and impactful research, gathering and using data must be done with care including incorporating the values of data sovereignty, community engagement, and trust building; this is especially for historically marginalized groups that often see few rewards from genomic research, such as Indigenous Americans[70,71]. Additionally, reference data ancestry groups (e.g., 1KG, HGDP) are often labeled according to the geographic region from which the sample originated or was gathered, and likely do not represent the full genetic diversity of the region or people. Thus, care should be taken when interpreting group substructure proportions such as using the "-like" suffix (e.g., 1KG-PEL-like) to indicate similarity to the particular reference group rather than all people, as recommended by the National Academies Report[8].

While representation in the reference data is essential to providing harmonized AF with the lowest bias, homogeneous or highly dissimilar reference groups may not be necessary. The AF adjustment using highly similar, finer-scale ancestry substructure groups indicates that while there is bias in the subgroup proportion estimation, the bias is contained locally within the similarity network and does not propagate forward to the AF adjustments. Therefore, it is possible that heterogeneous subgroups (e.g., admixed) may be useful in creating a reference similarity map for summary data harmonization as long as the relevant genetic substructure is contained within the map. Representation and definition of reference samples has implications for substructure beyond ancestry. For example, when estimating the proportion of a sample with genetic similarity to cases and controls, the definition and ascertainment of cases and controls (e.g., all cases, early age of onset cases) may result in different genetic representation in the reference data and subsequently different mixture proportion estimates.

While the amount of publicly available and easily accessible genetic summary data continues to increase at impressive rates, not all of the genetic summary data is easily usable[72]. In an impact assessment of the GWAS catalog between 2020-2022, 17% of GWAS datasets and 50% of submissions were missing newly mandatory fields, with the effect AF being the most commonly omitted field[73]. Missing AF data reduces the ability to employ secondary data methods such as PGSs, mendelian randomization, and mixture modeling and harmonization methods (e.g, Summix). Additional gaps in meta-data relating to ascertainment, genetic ancestry, and environmental background also hinders the ability to robustly use summary genetic data. New standards for reporting of GWAS data have been proposed and include mandatory fields (e.g., effect AF, effect estimate, standard error), metadata reporting, and file standardization, which can be created using commonly used software such as PLINK 2.0[73,74]. While not required, reporting of AF by meaningful groups (e.g., environmental, case/control) would be especially valuable for harmonizing summary datasets, especially when the GWAS study used the grouping in the analysis.

Until recently, modeling substructure in genetic data has been limited to stratifying into homogeneous groups, which either ignores finer-scale substructure or cannot be applied to populations with substantial structure (e.g., genetically admixed populations) resulting in biased results or an underuse of data (often the very data that is also underrepresented in research)[75]. Here, we show that the methods contained in the *Summix2* software package can detect, adjust, and even harness substructure in summary data by estimating genetic similarity to reference groups using only summary level data. *Summix2* methods can estimate local-genetic substructure, detect putative regions of selection as well as estimate finer-scale substructure, and genetic substructure beyond ancestry laying the foundation for better use and harmonization for genetic summary level data. These methods are important in increasing equitable use of summary data, especially for data containing substantial substructure, such as from genetically admixed populations (e.g., African American and Latinx). These summary data are often both underrepresented and underused in research especially due to a lack of methods to robustly model the substructure within. Given the ability to estimate and use substructure,

the suite of methods in the *Summix2* software package enable robust and more equitable use of genetic summary data.

Online Methods/Methods

Ethics Statement

Here, we have primarily used publicly available resources from gnomAD[35]. We use the gnomAD, 1KG, and HGDP group labels[35]. As we recognize that group labels are somewhat arbitrary often based on geographic location of the individuals included, we refer to these groups using the assigned labels when directly referencing the public data. When we report estimated or simulated genetic substructure similarity proportions, we have adopted the -like nomenclature[8].

We use CCPM Biobank data for individuals classified as European ancestry using genotype information (for full details see publication[55]). We recognize that limiting research to individuals of European ancestry limits the generalizability of the research conclusions. In future research, we will evaluate estimating and harmonizing over multiple levels of genetic substructure (e.g., admixture and study ascertainment).

Data

All data was mapped to the GRCh38 build. AFs from gnomAD version 3.1.2 for 644,267,978 variants that passed gnomAD's QC filter for 76,156 whole genomes from ten groups (i.e., AFR/AFRAM, AMI, AMR, ASJ, EAS, FIN, NFE, MID, SAS, and OTH) was obtained in November 2021[35]. Reference data for 76 HGDP & 1KG groups was also obtained from gnomAD v3.1.2. After removing admixed, one group that was not classified in the five continental groupings, and a group with N=1, 50 finer-scale reference ancestry groups remained (**Experiment Overview** in **Results**; **Extended Data Table 2**). We merged the observed data and HGDP & 1KG reference data by chromosome, variant position, reference allele, and alternate allele. Variants with missing values in either dataset were removed. We then limited to variants with MAF > 1% in at least one of the 50 finer-scale reference groups resulting in 75,526,117 variants across the 22 autosomes. This data was used for the finer-scale evaluations. For comparisons using continental level reference groups, the data was additionally filtered to keep variants within MAF > 1% in at least one continental ancestry reference group. AF for continental level groups were calculated as a weighted sum of the finer-scale ancestry groups within each continental level grouping as released with the gnomAD v3.1.2 data (**Extended Data Table 2**).

AFs from the CCPM Biobank were obtained for individuals classified as EUR through genotype information (for full details on ancestry group assignment see publication[55]). AFs were calculated for individuals with and without prostate cancer, as defined by the presence or absence of prostate cancer ICD-9 and ICD-10 codes. The groups were further stratified by biological sex (male, female), and age (≤40 years, 41-60 years, ≥61 years). In CCPM we obtained AFs for 138 of the 148 SNPs from PGS00084[54].

Simulation AFs were generated assuming Hardy-Weinberg Equilibrium using the *rmultinom* function in base R with the probability defined using the AFs from the finer-scale or continental reference ancestry groups. Simulation code is available on accompanying GitHub site (see **Data Availability**) and more information is in the **Results** in ***Experiment Overview***.

Summix

Summix estimates subgroup mixing proportions (π_k) by minimizing the least-squares difference for a vector of N AFs in the observed sample (AF_{obs}) and the generated AF from a mixture of K reference groups ($AF_{ref,k}$) (equation (1)).

$$minimize: f(\boldsymbol{\pi}) = \sum_{i=1}^{N} (AF_{obs,i} - \sum_{k=1}^{K} (\pi_k * AF_{ref,i,k}))^2 \quad (1)$$

where $\pi_k \geq 0, k = 1, \ldots, K; \sum_{k=1}^{K} \pi_k = 1$, and where $i = 1, \ldots, N$ SNPs

We use sequential quadratic programming (SQP)[76] to efficiently estimate π_k. Full details on the original *Summix* method can be found in Arriaga-MacKenzie, et al.[33]

Goodness of fit

The GoF measure from Summix v1 was simply the LS objective function, $f(\boldsymbol{\pi})$ per 1000 SNPs. To identify the relationship between MAF and the GoF metric, we simulated 108,404 AFs from chr 22 for 85% AFR-like and 15% EUR-like mixing proportions. The variants were binned into MAF bins (0.0-0.01, 0.01-0.02, 0.02-0.05, 0.05-0.1, 0.1-0.3, 0.3-0.5) and variants in each of these bins were then used to estimate the substructure proportions subsequently obtaining a GoF value. This was repeated for 1000 replicates.

After observing that the original GoF metric increased by MAF bin, we developed a weighted GoF measure, equation (2). To determine the weights, w_l, we estimated the substructure proportions and obtained the original GoF for seven gnomAD groups, binning the SNPs by MAF as outlined above. For the reference data, we used the five continental level 1KG and HGDP groups rerunning the estimation excluding any reference groups with <2% estimated proportion. To normalize the GoF in each MAF bin to match the largest GoF value (i.e., in the 0.3-0.5 MAF bin), we estimated weights for each bin using equation (3). We observed similar bin weights across the gnomAD groups, and thus averaged over the gnomAD groups to estimate weights for the new GoF measure, equation (2).

$$GoF = \frac{1}{L} \sum_{l=1}^{L} \frac{1}{1000} w_l f(\widehat{\boldsymbol{\pi}})_{i \in G_l} \quad (2)$$

Where there are L MAF bins and $f(\widehat{\boldsymbol{\pi}}_i)$ is the LS objective function for SNPs within group, G_l. Here, we use L=3 MAF bins of (0, 0.1], (0.1, 0.3], (0.3-0.5].

$$w_l = \frac{f(\widehat{\boldsymbol{\pi}})_{i \in G_{MAF:\, 0.3-0.5}}}{f(\widehat{\boldsymbol{\pi}})_{i \in G_l}} \quad (3)$$

To identify GoF thresholds, we simulated a mixture of five continental reference ancestry groups using 1,000 replicates for each simulation scenario. The proportion of one reference ancestry was varied between 0.05 to 0.95 by 0.05 increments and was left out of the reference panel. The proportions for the other four groups were randomly assigned to sum to 1 and remained in the reference panel. The process was repeated so that each of the five groups was left out once.

Substructure adjusted allele frequencies

In the previous release of *Summix*, adjusted AFs were calculated by leaving out a single reference group, l, of the user's choosing. To ensure the adjusted AFs are less dependent on a single reference group, the default method for adjusted AFs was updated to iteratively leave out each reference group, and then average across K iterations (equation (4)).

$$x = \frac{1}{K}\sum_{l=1}^{K}\left(\frac{\pi_{target,l}}{\widehat{\pi_l}}\left(AF_{observed} - \sum_{k\neq l}\widehat{\pi_k}AF_{ref,k}\right) + \sum_{k\neq l}\pi_{target,k}AF_{ref,k}\right) \quad (4)$$

$$AF_{adj} = \begin{cases} removed, & x < \delta \\ 0, & \delta \leq x < 0 \\ x, & 0 \leq x \leq 1 \\ 1, & x > 1 \end{cases}$$

Where AF_{adj} is the adjusted AF and l is the index for the reference group left out for that iteration, and δ is the precision threshold for which adjusted AF are removed from the dataset or rounded to 0. We found a value of δ -0.005 worked well. We tested the performance of the previous allele frequency adjustment method compared to the updated allele frequency adjustment method and provide comparison results when adjusted from a simulated AMR-like observed sample to a simulated PEL-like target sample (**Supplementary Figure 1, Supplementary Note 2**).

As the adjusted AFs are a combination of the observed and reference datasets, the full sample size of the observed data is likely not fully used. Thus, *Summix2* now estimates an effective sample size, N_{eff}, for adjusted AF using equation (5). We also output a warning if the effective sample size is less than 50% of the sample size of the observed group.

$$N_{eff} = \left\lfloor \sum_{k} \min\left(\widehat{\pi_k}, \pi_{target,k}\right) * N_{observed} + \min\left(1, \pi_{target,k}K\right) * N_{reference,k} \right\rfloor \quad (5)$$

Where, $N_{observed}$ is the sample size of the observed group and $N_{reference,k}$ is the sample size of kth reference group.

Summix-local

We evaluated *Summix2*'s utility to estimate local substructure using two algorithms: sliding windows, and fast-catchup[39]. For the sliding window algorithm, the size and overlap of the windows is user defined by base pairs (bp) or number of SNPs. *Summix2* estimates the substructure proportions using the SNPs within each window. The substructure proportions within the overlapping sections are averaged. In the fast-catchup algorithm, the window size is variable, overcoming a limitation in which fixed window size may obscure the breakpoints between ancestral segments. The fast-catchup algorithm uses a start- and end-pointer. Beginning at the start-pointer, a window is created using a user-specified number of variants, or size in bp. During each iteration, the end-pointer moves forward one SNP, *Summix2* estimates the substructure proportions, which are compared to the substructure proportions from the previous iteration (the window with one fewer SNP included). The algorithm ends and a block is defined when a percent difference change in one of the reference groups is larger than the predefined threshold (default 2%). The start-pointer is then advanced to the current end-pointer SNP and the algorithm begins again, continuing until the end of the chromosome is reached. We refer to windows as the input to the Summix algorithm and blocks as the output. For the sliding window approach, block and windows are the same. For the fast catch-up method, blocks consist of windows with the same ancestry proportions.

We developed a test to identify local chromosomal regions with different substructure proportions compared to the global average, equation (6). We use a Wald test statistic where $\hat{\pi}_{w,k}$ is the substructure proportion estimate from Summix for block w and reference group k, $\overline{\pi_k}$ is the average substructure proportion over W blocks for the chromosome, excluding block w. P-values are estimated using the t-distribution with N – 1 degrees of freedom, where N is the number of SNPs in block w. A

Bonferroni-corrected threshold of 0.05 correcting for the total number of windows (*W*) and K-1 reference groups is used to adjust for multiple testing.

$$T_{w,k} = \frac{\hat{\pi}_{w,k} - \overline{\pi_k}}{SE\left(\hat{\pi}_{w,k}\right)} \tag{6}$$

To estimate the standard error, $SE\left(\hat{\pi}_{w,k}\right)$, we average estimates of within block and between block error for each reference group, *k*. We use simulations to estimate within block standard error. Briefly, we create 1000 replicates of simulated reference AFs using a multinomial distribution with parameters equal to the number of haplotypes and AFs in the real reference data. The simulated reference AFs are used to estimate the within block substructure proportions of the observed data. The within block standard error estimate is the standard deviation of the resulting sampling distribution. To account for between block variability, we estimate the standard deviation of the proportion estimates over the chromosome.

Local Ancestry Simulations

We filtered the gnomAD v3.1.2 chromosome 19 data for LD using PLINK, removing variants with a pairwise r^2>0.2. SNP genotypes were simulated using *rmultinom*, assuming Hardy-Weinberg equilibrium, and using reference AFs as the probabilities. We simulated both observed and reference samples. We simulated AFR/AFRAM-like samples with 85% AFR-like and 15% EUR-like proportions, and AMR-like samples with 10% AFR-like, 65% EUR-like, and 25% IAM-like proportions. We assessed observed sample sizes of 1000, 10,000, and 50,000 individuals, reference samples of 100, 500, and 1000 individuals per group, and window sizes of 200, 1,500, and 3,000 variants using *Summix-local*. The default parameters were 10,000 observed individuals, reference samples of 500 individuals per group, and a window size of 1,500 variants, then one parameter was changed at a time. Minimum accuracy and precision across the local substructure blocks of proportion estimates was assessed using equations (7) and (8).

To assess type I error and power to detect regions where the local substructure differs from the chromosomal global substructure, we replaced a segment of the simulated observed data with simulated data comprised of different ancestry proportions using a magnitude of change of 2%, 3%, or 5% and region size of 50,000bp, 400,000bp, or 1,000,000bp. The default parameter dataset was simulated to have a region of 400,000bp with a 3% difference in IAM-like local ancestry proportions for the AMR-like sample and 2% difference in AFR-like local genetic similarity proportions in the AFR/AFRAM sample compared to the global proportions. We defined type I error as the proportion of simulations out of the 100 replicates with at least one significant block outside of the true signal. Power was similarly defined as the proportion of simulation replicates with at least one significant block within the true signal. The type I error and power were then calculated using Bonferroni-corrected p-values < 0.05, corrected for K-1 ancestries and W, the total number of blocks.

$$minAcc(P) = 1 - \max\{|Bias\left(\hat{\pi}_p\right)|\} \tag{7}$$

Where $Bias\left(\hat{\pi}_p\right)$ refers to the mean difference between the true simulated proportion and the estimated proportion of genetic substructure and *P* refers to the number of blocks (*W*).

$$minPrec(P) = 1 - \max\{Std\ Dev(\pi_p - \hat{\pi}_p)\} \tag{8}$$

Where $Std\ Dev(\pi_p - \hat{\pi}_p)$ refers to the standard deviation of the difference between the true simulated proportion and the Summix estimate of the proportion of genetic substructure across P subgroups, here the number of blocks (*W*).

Local-substructure scan in gnomAD

We applied *summix-local* to gnomAD AMR and AFR/AFRAM using the fast-catchup algorithm and the default parameters. For computational efficiency, we assessed each chromosome separately, using the continental reference MAF filtered gnomAD data and the 1KG and HGDP continental reference data. We removed blocks with poor fit (GoF > 1.5) from further analysis. Multiple testing correction was accounted for by a Bonferroni-correction for 5945 genome-wide blocks and the number of reference groups (two and three, respectively) resulting in a threshold of 4.21e-6 and 2.80e-6 for AFR/AFRAM and AMR.

We then selected the blocks that overlapped by bp position with previously identified regions of selection. To determine replication, we performed Bonferroni correction for the number of reference groups and the number of replication signals being tested (16 and 29 for AFR/AFRAM and AMR, respectively).

Fixation Index

We used the Hudson pairwise Fixation Index (F_{ST})[55] to estimate similarity between reference groups, equation (9).

$$\hat{F}_{ST}^{Hudson} = \frac{(\tilde{p}_1 - \tilde{p}_2)^2 - \frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1 - 1} - \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2 - 1}}{\tilde{p}_1(1-\tilde{p}_2) + \tilde{p}_2(1-\tilde{p}_1)} \quad (9)$$

Where n_i is the sample size and $\tilde{p}_i$ is the sample AF for group i for $i \in \{1, 2\}$.

The Hudson pairwise F_{ST} estimator measures population similarity by averaging the Weir and Hill F_{ST}[77] estimates for two groups. The Weir and Hill F_{ST} estimator calculates the correlation between alleles from a given population relative to its most recent ancestral population. Unlike other F_{ST} estimators, estimates made by $\hat{F}_{ST}^{Hudson}$ are not sensitive to the ratio of samples sizes in each population[55].

We estimated the F_{ST} between each pair of the 50 finer-scale reference groups from 1KG and HGDP using individual level data. We kept variants with MAF ≥ 0.01 and r^2 < 0.2 was estimated using *bcftools* with the command *bcftools +prune*. Individuals labeled as related in gnomAD v3.1.2 metadata [35] were removed. We calculated pairwise F_{ST} using the $\hat{F}_{ST}^{Hudson}$ function available through *PLINK*[74].

Finer-scale Ancestry Simulations

We performed simulations using the HGDP and 1KG finer-scale groups as reference data to assess the minimum accuracy and precision of using *Summix2* to estimate finer-scale substructure. AFs were simulated independently for each SNP using a multinomial distribution (the *rmultinom* function in base R), assuming Hardy-Weinberg equilibrium, and using reference AFs as the probabilities. We simulated reference ancestry sample sizes of 100 ($N_{ref,k}$) for each group and observed group sample sizes of 10,000 (N_{obs}). Each simulation scenario was replicated 100 times sampling 100,000 genome-wide SNPs in each replication. Minimum accuracy and precision were defined in equation (7) and equation (8), respectively, where p is the number of finer-scale references (K).

To evaluate how similarity in substructure affected the accuracy and precision of the proportion estimates, we simulated observed data from reference data groups with varying levels of similarity as measured by pairwise F_{ST}. Finer-scale EAS ancestries with pairwise F_{ST} of .005, .007, and .009 were chosen to represent high, medium, and low fine-scale group similarity between reference groups, respectively. Two pairs of fine-scale ancestries with the given pairwise F_{ST} were used as reference ancestries in each simulation scenario. Oroqen and daur (pairwise F_{ST}=.009) and yizu and naxi (pairwise F_{ST}=.009) were selected for pairwise F_{ST} grouping of .009, tujia and han (pairwise F_{ST}=.007) and xibo and mongola (pairwise F_{ST}=.007) were selected for pairwise F_{ST} grouping of .007, chb and khv (pairwise F_{ST}=.005) and tu and mongola (pairwise F_{ST}=.005) were selected for pairwise F_{ST} grouping of .005. See **Extended Data Table 7** for simulated ancestry proportions.

Using the 50 HGDP and 1KG finer-scale reference groups, we evaluated finer-scale substructure estimation in real data using gnomAD v3.1.2's ten groupings: African/African American, Amish, Latino/Admixed American, Ashkenazi Jewish, East Asian, Finnish, Non-Finnish European, Middle Eastern, South Asian, Other. Ninety-percent CIs for the proportion estimates were calculated by randomly sampling 100,000 genome-wide SNPs across 100 replicates (**Extended Data Table 8**).

Adjusted Allele Frequencies

On chromosome 19, we assessed adjusting AFs by adjusting the observed gnomAD AMR AFs to three different target samples: 1) 1KG PEL, 2) a simulated target with 1KG PEL-like continental substructure (10% AFR-like, 15% EUR-like, 75% IAM-like), and 3) a simulated target with 1KG PEL-like finer-scale substructure (45% HGDP-maya-like, 10% HGDP-karitiana-like, 10% HGDP-pima-like, 10% HGDP-colombian-like, 5% 1KG-esn-like, 5% 1KG-yri-like, 8% 1KG-gbr-like, 7% 1KG-ibs-like). The reference group proportions to simulate the continental and finer-scale target samples were estimated from the 1KG-PEL-like sample using Summix. We estimated the observed reference group proportions using *Summix2* using either continental or finer-scale reference data and at either the global or local level for four estimation groups: 1) continental global, 2) continental local, 3) finer-scale global, 4) finer-scale local. We compared these AF adjustments to adjustments made using our adjusted AF method and the local ancestry proportion estimates from gnomAD v3.1 for the AMR group using continental reference groups (AFR, EUR, and IAM) and RFMix on individual level data. We limited our comparison to SNPs with MAF>0.01 that had gnomAD local proportion estimates available (i.e., 12,513 out of 234,633 SNPs on chr 19). We assessed the absolute and relative difference between the observed (adjusted or unadjusted) AF to the target AF across MAF bins ([0.01, 0.02), [0.02, 0.05), [0.05, 0.10), [0.10, 0.30), [0.30, 0.50]). To assess if the absolute or relative difference differed between adjustment groups, we performed paired t-tests. We also assess Lin's CCC between the unadjusted AFs, adjusted AFs and target AFs.

Estimating the proportion of a sample with genetic similarity to case-status

We investigated the ability of *Summix2* to detect genetic substructure using prostate cancer cases and controls from a 2018 prostate cancer GWAS used to create a PGS000084. AFs for the prostate cancer cases and controls were used as the reference data to estimate the proportion of people in the Colorado Center for Personalized Medicine (CCPM) Biobank data that were genetically similar to the prostate cancer cases[57,58]. Out of the 147 SNPs included in the published prostate cancer PGS, 138 existed in the CCPM Biobank v2 and were used for further analysis. We subset the CCPM Biobank data by biological sex (i.e., female, male), males with prostate cancer, and age group (i.e., < 40 years, 40-60 years, >60 years). We estimated jackknife 95% CIs.

Data Availability

The gnomAD, 1KG and HGDP data is available for download at https://gnomad.broadinstitute.org/downloads. A sample simulated AF dataset, as well as code used to generate simulated datasets is available on github at https://github.com/hendriau/Summix2_manuscript. The provided simulated dataset contains the 234,633 LD 0.2 filtered variants from chromosome 19 with observed finer-scale-like simulated substructure proportions of 45% HGDP-maya-like, 10% HGDP-karitiana-like, 10% HGDP-pima-like, 10% HGDP-colombian-like, 5% 1KG-esn-like, 5% 1KG-yri-like, 8% 1KG-gbr-like, 7% 1KG-ibs-like. The dataset also has the 50 simulated finer-scale references and 5 continental simulated references with n = 200 and 500 respectively. This dataset was simulated from the dataset filtered for MAF > 1% in at least one continental reference group.

Code Availability

Summix2 software (R package) is available at https://bioconductor.org/packages/devel/bioc/html/Summix.html. R software is available at https://www.r-project.org/. PLINK 1.9 software is available at www.cog-genomics.org/plink/1.9/.

FIGURES

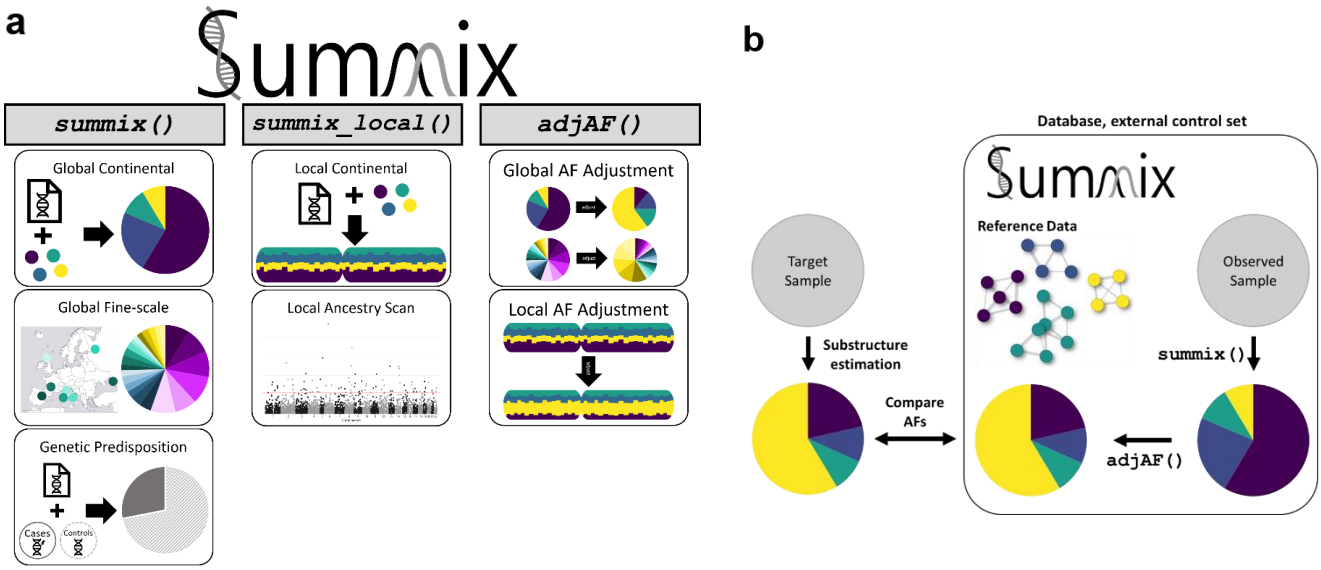
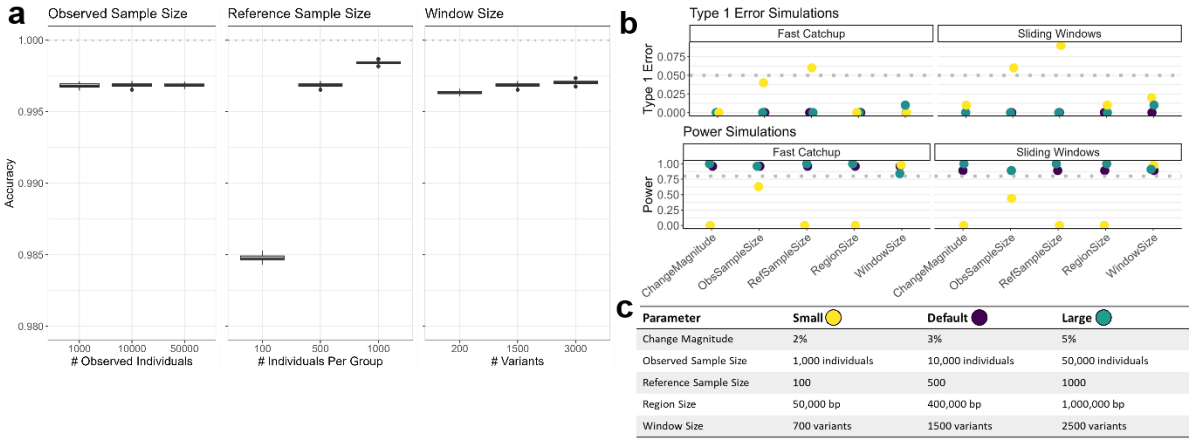


Figure 1 | *Summix2* suite of methods overview of functions and uses. a, Three primary methods to detect and leverage substructure in genetic summary data and **b,** an example of its utility.



Parameter	Small	Default	Large
Change Magnitude	2%	3%	5%
Observed Sample Size	1,000 individuals	10,000 individuals	50,000 individuals
Reference Sample Size	100	500	1000
Region Size	50,000 bp	400,000 bp	1,000,000 bp
Window Size	700 variants	1500 variants	2500 variants

Figure 2 | *Summix-local* accurately estimates local substructure proportions and detects regions where local substructure differs from global in a Latinx sample. a, *Summix-local* accuracy (y-axis) using default simulation parameters (1000 observed individuals, 500 individuals per reference group, window size 1500 variants) except when varying by: observed sample size (left), reference sample size (middle), and window size defined by number of variants (right). 100 replicates per simulation scenario. **b**, Type I error (top) and power (bottom) for the *Summix-local* local substructure scan for fast-catchup (left) and sliding windows (right) algorithms across simulation scenarios as shown in **c**.

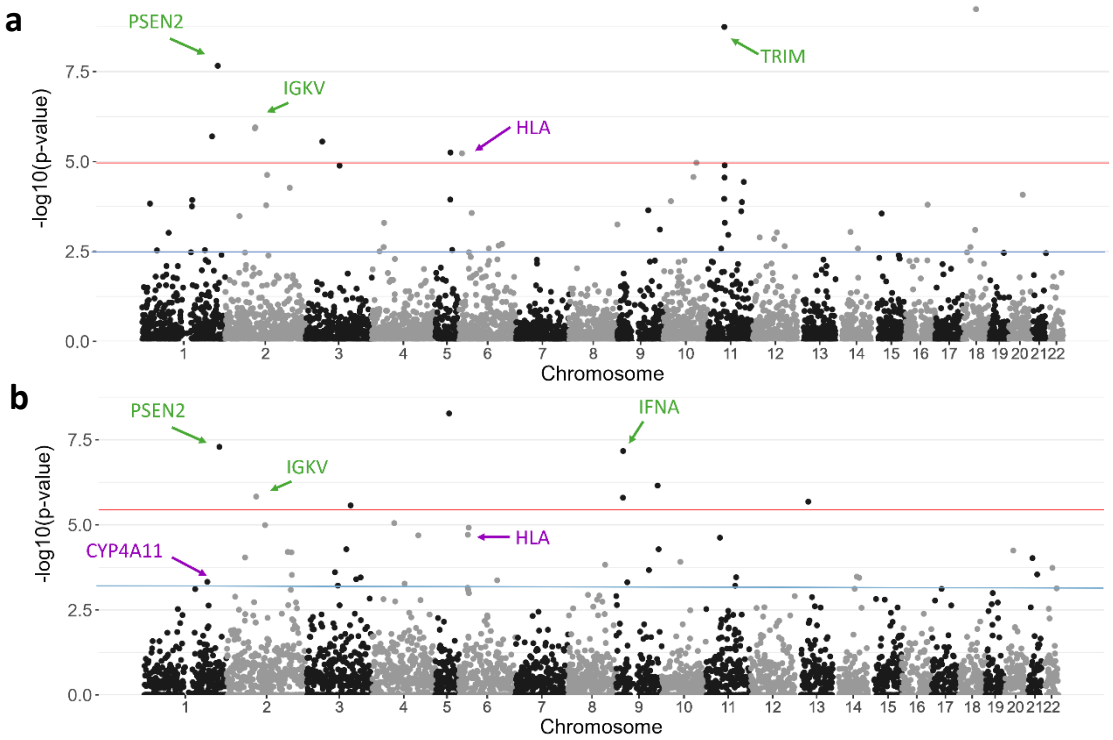


Figure 3 | Local ancestry scan in gnomAD African/African American and Latinx/Admixed American identifies potential regions of selection. *Summix-local* applied to gnomAD v3.1.2 **a,** African/African American (AFR/AFRAM) and **b,** Latinx/Admixed American (AMR). Red line indicates Bonferroni-corrected genome-wide significance ($P < 4.21e\text{-}6$ AFR, 2.80e-6 AMR); blue indicates Bonferroni-correction for the replication signals ($P < 1.56e\text{-}3$ AFR, 5.75e-4 AMR).

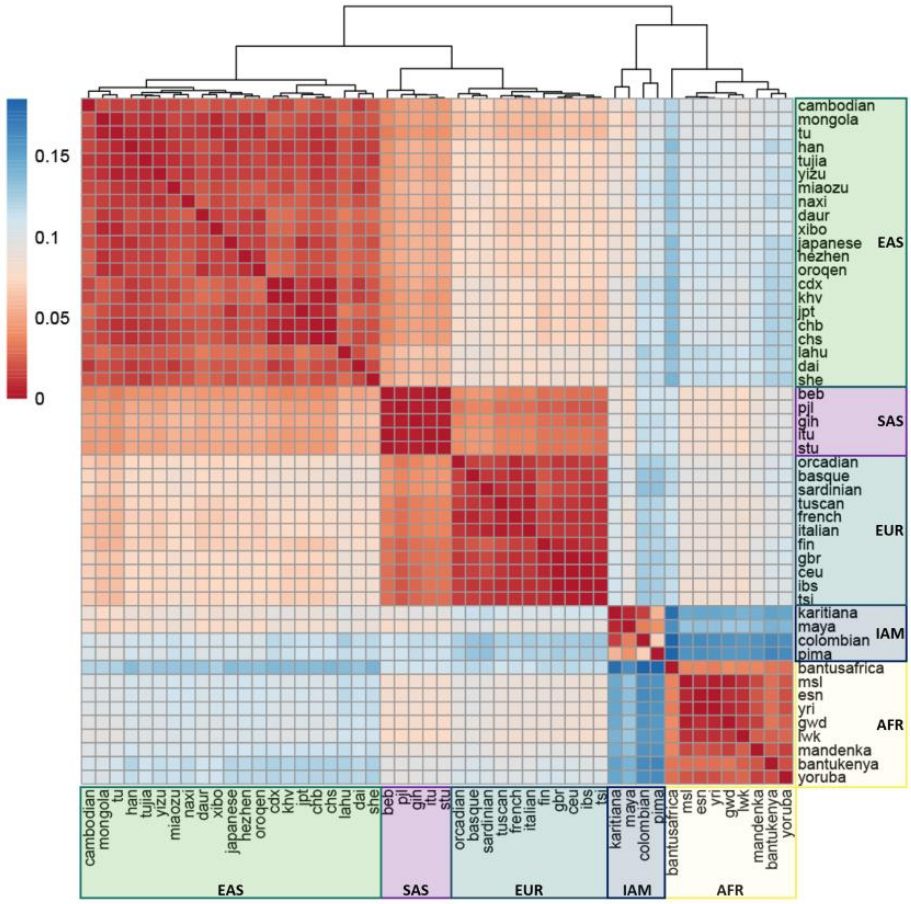


Figure 4 | Comparison of genetic similarity within continental and finer-scale ancestry groups. Pairwise F_{ST} ranging from .001(red, indicating high similarity) to .035 (blue, indicating low similarity). Fifty finer-scale groups within continental ancestry groups as defined by 1KG and HGDP: EAS (East Asian, green), SAS (South Asian, purple), EUR (European, teal), IAM (Indigenous American, blue), AFR (African, yellow).

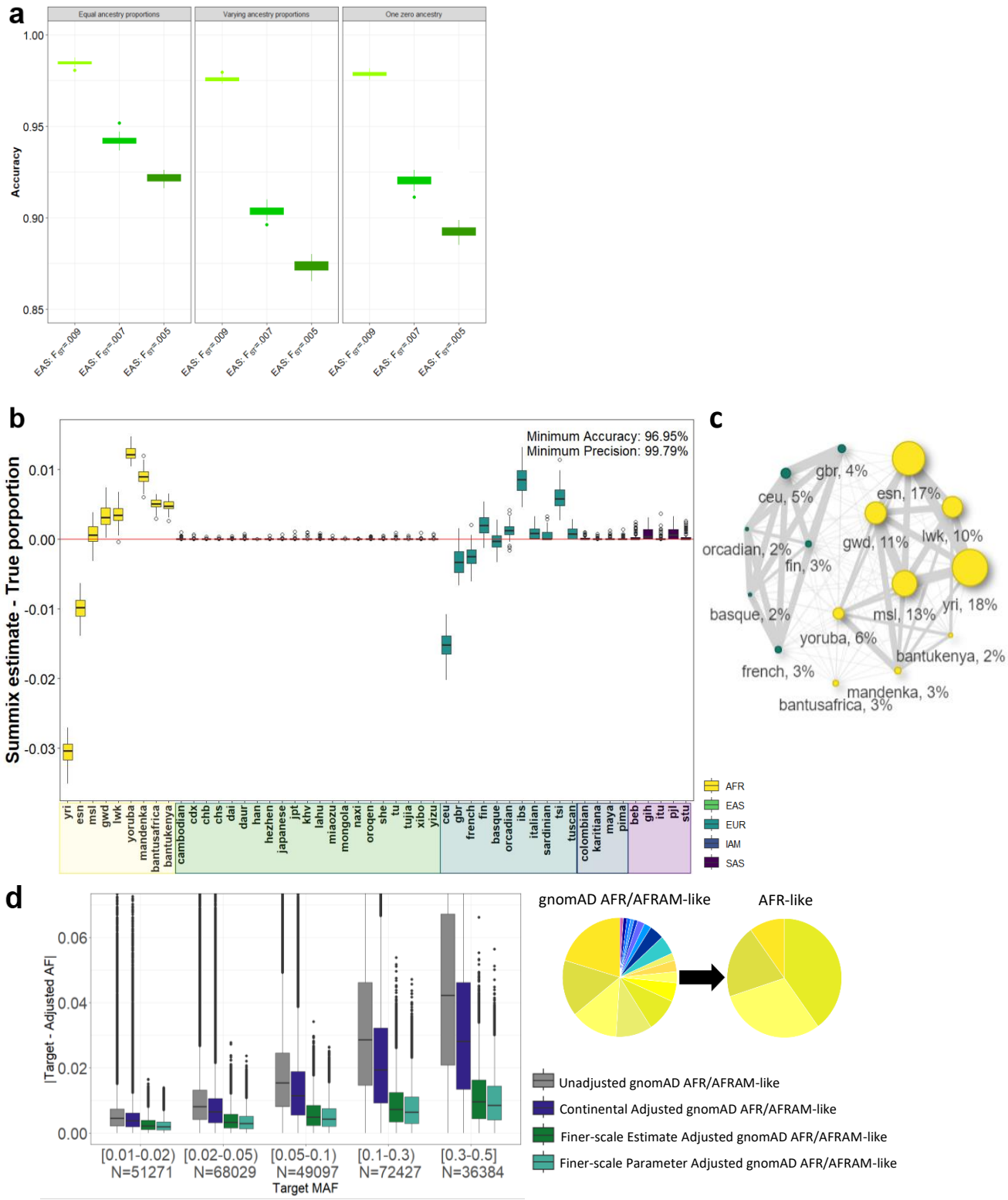


Figure 5 | *Summix2* detects and adjusts for fine-scale genetic similarity proportions. a, Accuracy of *Summix2* estimates for three simulation scenarios: Equal Ancestry Proportions (.25, .25, .25, and .25), Varying Ancestry Proportions (.1, .2, .3, and .4), and One Zero ancestry (0, .33, .33, .33). **b**, Accuracy and precision for gnomAD AFR/AFRAM-like simulation scenario (Note: "-like" nomenclature omitted from x-axis plot labels for simplicity). **c**, Genetic substructure similarity map of *Summix2* estimates for the gnomAD AFR/AFRAM-like population (Note: "-like" nomenclature omitted from x-axis plot labels for simplicity). **d**, Absolute difference between target AFR-like population AFs and unadjusted or adjusted gnomAD AFR/AFRAM-like AFs grouped by MAF bin.

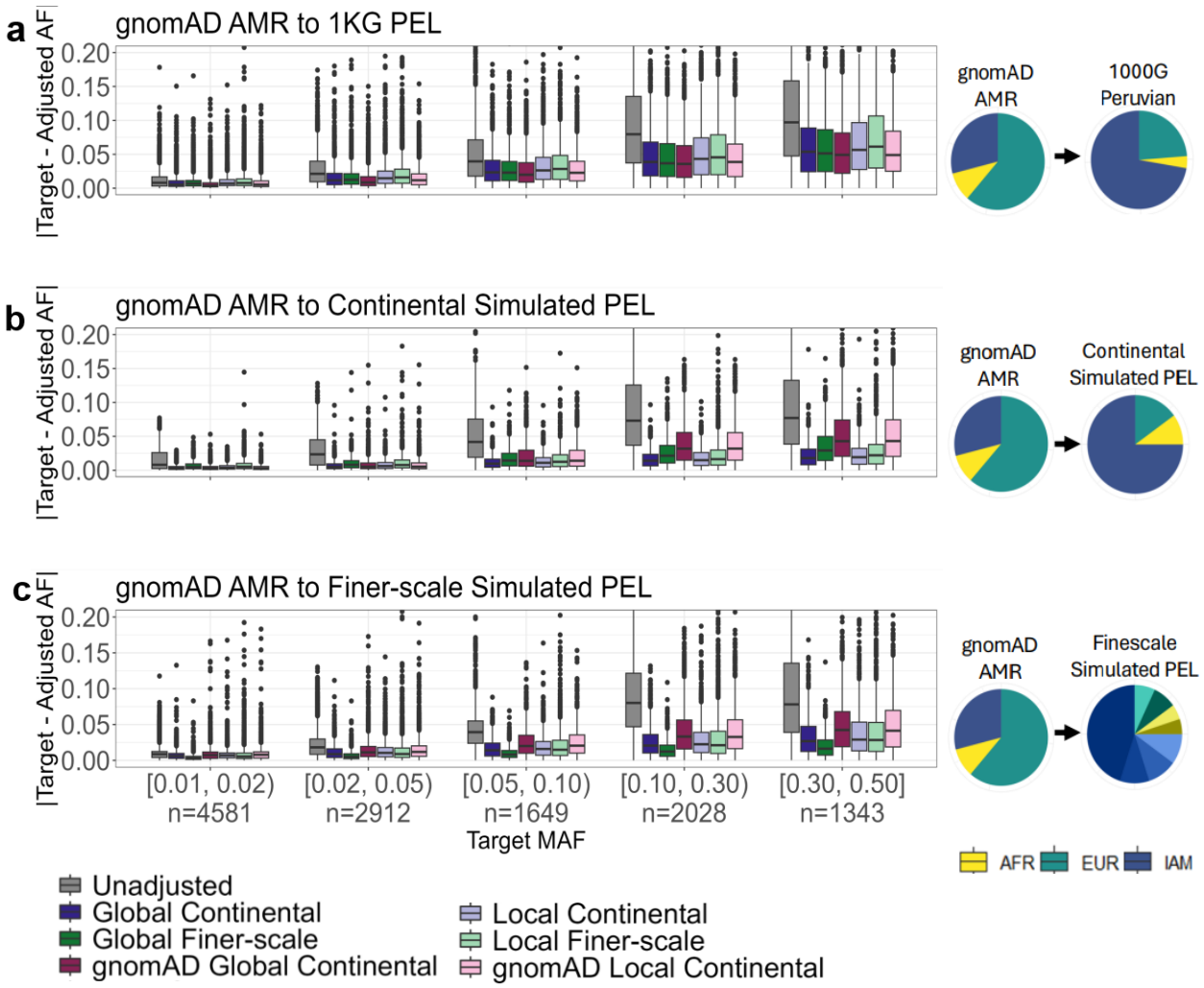


Figure 6 | Allele frequency adjustment improves data harmonization. Absolute difference after adjusting gnomAD AMR to three different target samples **a**, 1000 Genomes Peruvian**. b**, Simulated Peruvian-like with continental-level substructure. **c**, Simulated Peruvian-like with fine-scale substructure, by minor allele frequency bin. Showing results for the 12,513 variants on chromosome 19 with gnomAD local ancestry deconvolution results, and filtering for LD ($r^2 > 0.2$) and MAF (>1%). The unadjusted gnomAD AMR (grey) is compared to adjustments using gnomAD and *Summix2* using continental and finer-scale simulated references, as well as global and local-level substructure. The global substructure proportions of the observed and targets is shown in the pie charts.

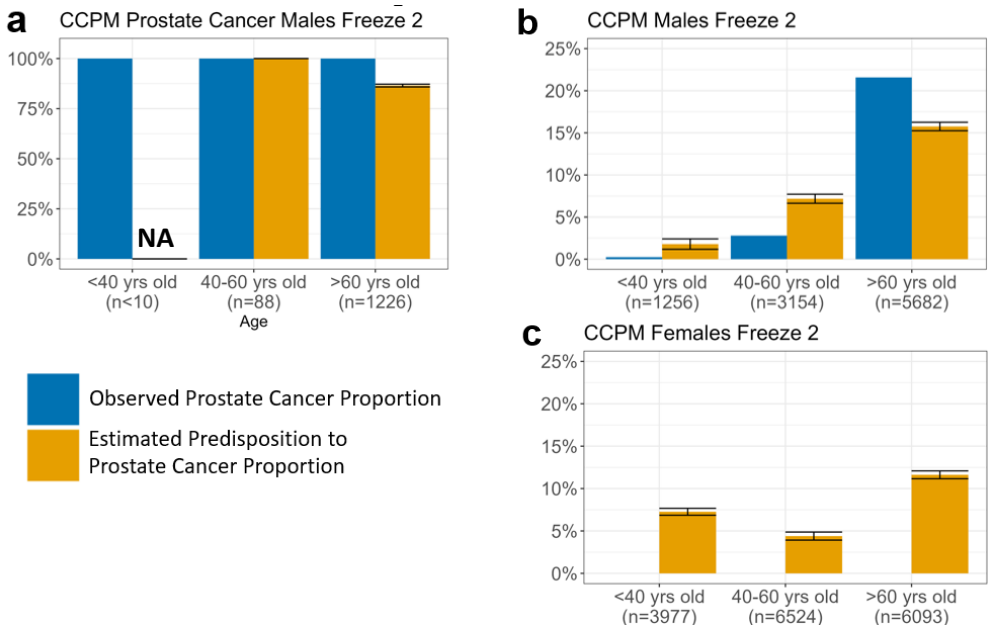


Figure 7 | Detecting genetic predisposition to prostate cancer in the CCPM Biobank. Genetic substructure estimates using *Summix2* in CCPM Biobank using prostate cancer cases and controls as reference data for 138 variants from a 2018 prostate cancer PGS stratified by **a**, Males with prostate cancer (n = 1318)**, b**, Males (with and without prostate cancer, n = 10,092), and **c**, Females (n = 16,594), as well as age. The observed proportion of individuals with prostate cancer (blue) is compared to the estimated proportion similar to reference prostate cancer cases (orange) with jackknife 95% confidence intervals shown. There were too few men with prostate cancer under the age of 40 years (<10) to estimate substructure proportions.

Table 1 | Replicated and genome-wide significant local ancestry blocks in gnomAD

Locus	Coordinates	Sample	P-value	Gene(s)
Replicated Signals				
6p21-22**	28749843-33594193	AFR	6.50e-11	*HLA*
1p33*	48977431-49882955	AMR	5.49e-4	*CYP4A11*
6p21-22*	27410678-32584612	AMR	1.21e-5	*HLA*
Genome-Wide Scan				
1q32.2	209710004-210350566	AFR	1.99e-6	*LAMB3, HSD11B1, IRF6, SYT14*
1q42.13	226810857-227371817	AFR	2.18e-8	*PSEN2, COQ8A*
2p11.2	89085547-91505258	AFR	1.19e-6	*IGKV*
3p21.31	49230341-50009909	AFR	2.78e-6	*RHOA, DAG1, TRAIP, MST1R*
10q24.2	98531740-99215582	AFR	1.08e-5	*HPSE2*
11p11.2	48713155-49141143	AFR	1.80e-9	*TRIM*
18q12.2	39086699-39784890	AFR	5.79e-10	*MIR924HG*
1q42.13	226810857-227371817	AMR	5.11e-8	*PSEN2, COQ8A*
2p11.2	89085547-91505258	AMR	1.48e-6	*IGKV*
3q22.1	132492441-133141527	AMR	2.69e-6	*UBA5, NPHP3*
5p13.1	41972605-42576421	AMR	5.32e-9	*GHR*
9p21.3	20543715-21637268	AMR	6.77e-8	*IFNA, MIR31HG*
9q33.3	125309605-125966124	AMR	7.03e-7	*MAPKAP, PBX3*
13q14.11	40546196-41169051	AMR	2.09e-6	*FOXO1, SLC25A15*

* Replication Bonferroni-corrected p-value < 1.56e-3 AFR, 5.75e-4 AMR

**Genome-wide Bonferroni-corrected p-value < 4.21e-6 AFR, 2.80e-6 AMR

References

1. Popejoy, A.B. & Fullerton, S.M. Genomics is failing on diversity. *Nature* **538**, 161-164 (2016).
2. All of Us Research Program, I. *et al.* The "All of Us" Research Program. *N Engl J Med* **381**, 668-676 (2019).
3. Wojcik, G.L. *et al.* Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**, 514-518 (2019).
4. GenomeAsia, K.C. The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature* **576**, 106-111 (2019).
5. Consortium, H.A. *et al.* Research capacity. Enabling the genomic revolution in Africa. *Science* **344**, 1346-8 (2014).
6. Martschenko, D.O., Wand, H., Young, J.L. & Wojcik, G.L. Including multiracial individuals is crucial for race, ethnicity and ancestry frameworks in genetics and genomics. *Nat Genet* **55**, 895-900 (2023).
7. Wojcik, G.L. *et al.* Opportunities and challenges for the use of common controls in sequencing studies. *Nat Rev Genet* **23**, 665-679 (2022).
8. National Academies of Sciences, E. & Medicine. *Using Population Descriptors in Genetics and Genomics Research: A New Framework for an Evolving Field*, 240 (The National Academies Press, Washington, DC, 2023).
9. Martin, A.R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am J Hum Genet* **107**, 788-789 (2020).
10. Petrovski, S. & Goldstein, D.B. Unequal representation of genetic variation across ancestry groups creates healthcare inequality in the application of precision medicine. *Genome Biol* **17**, 157 (2016).
11. Manrai, A.K. *et al.* Genetic Misdiagnoses and the Potential for Health Disparities. *N Engl J Med* **375**, 655-65 (2016).
12. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-9 (2006).
13. Alexander, D.H. & Lange, K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics* **12**, 246 (2011).
14. Nait Saada, J. *et al.* Identity-by-descent detection across 487,409 British samples reveals fine scale population structure and ultra-rare variant associations. *Nat Commun* **11**, 6130 (2020).
15. Shemirani, R. *et al.* Rapid detection of identity-by-descent tracts for mega-scale datasets. *Nat Commun* **12**, 3546 (2021).
16. Belbin, G.M. *et al.* Toward a fine-scale population health monitoring system. *Cell* **184**, 2068-2083 e11 (2021).
17. Liu, J., Lewinger, J.P., Gilliland, F.D., Gauderman, W.J. & Conti, D.V. Confounding and heterogeneity in genetic association studies with admixed populations. *Am J Epidemiol* **177**, 351-60 (2013).
18. Martin, E.R. *et al.* Properties of global- and local-ancestry adjustments in genetic association tests in admixed populations. *Genet Epidemiol* **42**, 214-229 (2018).
19. Wang, X. *et al.* Adjustment for local ancestry in genetic association analysis of admixed populations. *Bioinformatics* **27**, 670-7 (2011).
20. Zhang, J. & Stram, D.O. The role of local ancestry adjustment in association studies using admixed populations. *Genet Epidemiol* **38**, 502-15 (2014).
21. Pasaniuc, B. *et al.* Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nat Genet* **44**, 631-5 (2012).

22. 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**, 278-88 (2013).
23. Guan, Y. Detecting structure of haplotypes and local ancestry. *Genetics* **196**, 625-42 (2014).
24. Dias-Alves, T., Mairal, J. & Blum, M.G.B. Loter: A Software Package to Infer Local Ancestry for a Wide Range of Species. *Mol Biol Evol* **35**, 2318-2326 (2018).
25. Matzaraki, V., Kumar, V., Wijmenga, C. & Zhernakova, A. The MHC locus and genetic susceptibility to autoimmune and infectious diseases. *Genome Biol* **18**, 76 (2017).
26. Tang, H. *et al.* Recent genetic selection in the ancestral admixture of Puerto Ricans. *Am J Hum Genet* **81**, 626-33 (2007).
27. Zhong, Y., Perera, M.A. & Gamazon, E.R. On Using Local Ancestry to Characterize the Genetic Architecture of Human Traits: Genetic Regulation of Gene Expression in Multiethnic or Admixed Populations. *Am J Hum Genet* **104**, 1097-1115 (2019).
28. Basu, A. *et al.* Genome-wide distribution of ancestry in Mexican Americans. *Hum Genet* **124**, 207-14 (2008).
29. Mendoza-Revilla, J. *et al.* Disentangling Signatures of Selection Before and After European Colonization in Latin Americans. *Mol Biol Evol* **39**(2022).
30. Zhou, Q., Zhao, L. & Guan, Y. Strong Selection at MHC in Mexicans since Admixture. *PLoS Genet* **12**, e1005847 (2016).
31. Secolin, R. *et al.* Distribution of local ancestry and evidence of adaptation in admixed populations. *Sci Rep* **9**, 13900 (2019).
32. Salm, M.P. *et al.* The origin, global distribution, and functional impact of the human 8p23 inversion polymorphism. *Genome Res* **22**, 1144-53 (2012).
33. Arriaga-MacKenzie, I.S. *et al.* Summix: A method for detecting and adjusting for population structure in genetic summary data. *Am J Hum Genet* **108**, 1270-1282 (2021).
34. Prive, F. Using the UK Biobank as a global reference of worldwide populations: application to measuring ancestry diversity from GWAS summary statistics. *Bioinformatics* **38**, 3477-3480 (2022).
35. Chen, S. *et al.* A genome-wide mutational constraint map quantified from variation in 76,156 human genomes. *bioRxiv*, 2022.03.20.485034 (2022).
36. Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
37. Cavalli-Sforza, L.L. The Human Genome Diversity Project: past, present and future. *Nat Rev Genet* **6**, 333-40 (2005).
38. Lopez Herraez, D. *et al.* Genetic variation and recent positive selection in worldwide human populations: evidence from nearly 1 million SNPs. *PLoS One* **4**, e7888 (2009).
39. Piterman, S. How to Solve Sliding Window Problems: An Intro To Dynamic Programming. (2018).
40. Jin, W. *et al.* Genome-wide detection of natural selection in African Americans pre- and post-admixture. *Genome Res* **22**, 519-27 (2012).
41. Bryc, K. *et al.* Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proc Natl Acad Sci U S A* **107**, 786-91 (2010).
42. Deng, L., Ruiz-Linares, A., Xu, S. & Wang, S. Ancestry variation and footprints of natural selection along the genome in Latin American populations. *Sci Rep* **6**, 21766 (2016).
43. Lea, A.J. *et al.* Natural selection of immune and metabolic genes associated with health in two lowland Bolivian populations. *Proc Natl Acad Sci U S A* **120**, e2207544120 (2023).
44. Meyer, D., VR, C.A., Bitarello, B.D., DY, C.B. & Nunes, K. A genomic perspective on HLA evolution. *Immunogenetics* **70**, 5-27 (2018).
45. Cai, Y., An, S.S. & Kim, S. Mutations in presenilin 2 and its implications in Alzheimer's disease and other dementia-associated disorders. *Clin Interv Aging* **10**, 1163-72 (2015).

46. Lv, M. *et al.* LncRNAs as new biomarkers to differentiate triple negative breast cancer from non-triple negative breast cancer. *Oncotarget* **7**, 13047-59 (2016).
47. He, A., Chen, Z., Mei, H. & Liu, Y. Decreased expression of LncRNA MIR31HG in human bladder cancer. *Cancer Biomark* **17**, 231-6 (2016).
48. Saevarsdottir, S. *et al.* FLT3 stop mutation increases FLT3 ligand level and risk of autoimmune thyroid disease. *Nature* **584**, 619-623 (2020).
49. Liu, B., Chen, Y. & Yang, J. LncRNAs are altered in lung squamous cell carcinoma and lung adenocarcinoma. *Oncotarget* **8**, 24275-24291 (2017).
50. Yang, H. *et al.* Long noncoding RNA MIR31HG exhibits oncogenic property in pancreatic ductal adenocarcinoma and is negatively regulated by miR-193b. *Oncogene* **35**, 3647-57 (2016).
51. Wuttke, M. *et al.* A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nat Genet* **51**, 957-972 (2019).
52. Zhu, Z. *et al.* Shared genetic and experimental links between obesity-related traits and asthma subtypes in UK Biobank. *J Allergy Clin Immunol* **145**, 537-549 (2020).
53. Gene Ontology, C. *et al.* The Gene Ontology knowledgebase in 2023. *Genetics* **224**(2023).
54. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-9 (2000).
55. Bhatia, G., Patterson, N., Sankararaman, S. & Price, A.L. Estimating and interpreting FST: the impact of rare variants. *Genome Res* **23**, 1514-21 (2013).
56. Lin, L.I. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* **45**, 255-68 (1989).
57. Schumacher, F.R. *et al.* Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nat Genet* **50**, 928-936 (2018).
58. Wiley, L.K. *et al.* Building a vertically integrated genomic learning health system: The biobank at the Colorado Center for Personalized Medicine. *Am J Hum Genet* **111**, 11-23 (2024).
59. Gentleman, R.C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**, R80 (2004).
60. Hendricks, A.E., Dupuis, J., Logue, M.W., Myers, R.H. & Lunetta, K.L. Correction for multiple testing in a gene region. *Eur J Hum Genet* **22**, 414-8 (2014).
61. Cuadros-Espinoza, S., Laval, G., Quintana-Murci, L. & Patin, E. The genomic signatures of natural selection in admixed human populations. *Am J Hum Genet* **109**, 710-726 (2022).
62. Barreiro, L.B., Laval, G., Quach, H., Patin, E. & Quintana-Murci, L. Natural selection has driven population differentiation in modern humans. *Nat Genet* **40**, 340-5 (2008).
63. Lu, C., Ahmed, R., Lamri, A. & Anand, S.S. Use of race, ethnicity, and ancestry data in health research. *PLOS Glob Public Health* **2**, e0001060 (2022).
64. Bergstrom, A. *et al.* Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367**(2020).
65. Metspalu, M. *et al.* Shared and unique components of human population structure and genome-wide signals of positive selection in South Asia. *Am J Hum Genet* **89**, 731-44 (2011).
66. Gomez, F., Hirbo, J. & Tishkoff, S.A. Genetic variation and adaptation in Africa: implications for human evolution and disease. *Cold Spring Harb Perspect Biol* **6**, a008524 (2014).
67. Xing, J. *et al.* Genetic diversity in India and the inference of Eurasian population expansion. *Genome Biol* **11**, R113 (2010).
68. Campbell, M.C. & Tishkoff, S.A. African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. *Annu Rev Genomics Hum Genet* **9**, 403-33 (2008).
69. Garrison, N.A. *et al.* Genomic Research Through an Indigenous Lens: Understanding the Expectations. *Annu Rev Genomics Hum Genet* **20**, 495-517 (2019).

70. Tindana, P. & de Vries, J. Broad Consent for Genomic Research and Biobanking: Perspectives from Low- and Middle-Income Countries. *Annu Rev Genomics Hum Genet* **17**, 375-93 (2016).
71. Tsosie, K.S., Yracheta, J.M. & Dickenson, D. Overvaluing individual consent ignores risks to tribal participants. *Nat Rev Genet* **20**, 497-498 (2019).
72. Sollis, E. *et al.* The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res* **51**, D977-D985 (2023).
73. Hayhurst, J. *et al.* A community driven GWAS summary statistics standard. *bioRxiv*, 2022.07.15.500230 (2022).
74. Chang, C.C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
75. Gouveia, M.H. *et al.* Unappreciated subcontinental admixture in Europeans and European Americans and implications for genetic epidemiology studies. *Nat Commun* **14**, 6802 (2023).
76. Bonnans J.-F., G.J.C., Lemarechal C., Sagastizábal C.A. *Numerical Optimization: Theoretical and Practical Aspects*, (Springer Science & Business Media, 2006).
77. Weir, B.S., Cardon, L.R., Anderson, A.D., Nielsen, D.M. & Hill, W.G. Measures of human population structure show heterogeneity among genomic regions. *Genome Res* **15**, 1468-76 (2005).