Equitable machine learning counteracts ancestral bias in precision medicine, improving outcomes for all

Kiley Graim (kgraim@ufl.edu)
University of Florida https://orcid.org/0000-0002-4569-8444
Leslie Smith
University of Florida
James Cahill
University of Florida

Biological Sciences - Article

Keywords: equitable AI, genomics, cancer, ancestry, artificial intelligence

Posted Date: July 27th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3168446/v1

Additional Declarations: There is **NO** Competing Interest.

Equitable machine learning counteracts ancestral bias in precision medicine, improving outcomes for all

Leslie A Smith[1], James A Cahill[2], Kiley Graim[1*]

[1]Department of Computer & Information Science & Engineering, University of Florida, 432 Newell Dr, Gainesville, 32611, FL, USA.
[2]Environmental Engineering Sciences Department, University of Florida, 432 Newell Dr, Gainesville, 32611, FL, USA.

*Corresponding author(s). E-mail(s): kgraim@ufl.edu;
Contributing authors: leslie.smith1@ufl.edu; james.cahill@essie.ufl.edu;

Abstract

Gold standard genomic datasets severely under-represent non-European populations, leading to inequities and a limited understanding of human disease [1–8]. Therapeutics and outcomes remain hidden because we lack insights that we could gain from analyzing ancestry-unbiased genomic data. To address this significant gap, we present PhyloFrame, the first-ever machine learning method for equitable genomic precision medicine. PhyloFrame corrects for ancestral bias by integrating big data tissue-specific functional interaction networks, global population variation data, and disease-relevant transcriptomic data. Application of PhyloFrame to breast, thyroid, and uterine cancers shows marked improvements in predictive power across all ancestries, less model overfitting, and a higher likelihood of identifying known cancer-related genes. The ability to provide accurate predictions for underrepresented groups, in particular, is substantially increased. These results demonstrate how AI can mitigate ancestral bias in training data and contribute to equitable representation in medical research.

Keywords: equitable AI, genomics, cancer, ancestry, artificial intelligence

1 Introduction

Initiatives such as The Cancer Genome Atlas (TCGA) [9–12], All Of Us [13], and Therapeutically Applicable Research to Generate Effective Treatments (TARGET) [14] have generated a wealth of genomic resources for disease analysis, bringing about the field of precision medicine and revolutionizing cancer treatment [15]. However, these databases do not equally represent the diverse human ancestries that comprise the global population [1]. The GWAS Catalog [16], the largest genomic database that explicitly defines ancestry, is currently 95% European ancestry samples, despite Europeans being less than 9% of the global population (Fig. 1B). Numerous recent studies have observed substantial disparities in precision medicine effectiveness across ethnic groups [2–8], demonstrating a fundamental gap and inequity in some of the most cutting-edge approaches to precision medicine for cancer and other genetic disorders.

We address this need with the first ancestry-aware equitable machine learning framework for transcriptomics, PhyloFrame. By incorporating ancestry information, PhyloFrame creates equitable disease signatures that perform well globally, even when trained on non-diverse data. This approach offers an immediately actionable alternative to the mass sequencing that would be required to capture disease-driving genes across underrepresented populations in every cancer. Additionally, PhyloFrame's inclusion of population structure information inherent to human diversity [17] provides an essential and durable advantage over global bulk analyses regardless of the number of samples involved, allowing us to identify population specific variation more readily. We demonstrate that PhyloFrame provides improved predictive capacity in three diverse TCGA cancers (breast (BRCA), thyroid (THCA), and uterine (UCEC)), even in the (expected) scenario where few ancestries are represented in the training data. PhyloFrame exhibits marked improvements in predictive power for people of undersampled ancestry, illuminating a path towards more equitable and effective precision medicine.

2 Global population diversity impacts disease genomics

Despite overwhelming genomic similarity, human populations differ in ways that can substantially impact cancer treatment. African populations have the greatest genetic diversity [18], while non-African populations' genetic diversity is negatively correlated with distance from Africa due to population bottlenecks during the migration out of Africa [19] (Fig. 1A). In addition, as humans migrated throughout Africa and to other areas of the world (e.g., Asia, Europe, and the Americas), populations underwent selection for traits beneficial in their new environments [20–22]. To discern ancestral differences in disease presentation and susceptibility, we must first develop a better understanding of the genetic variation that exists among ancestries [23].

Ascertainment bias from unrepresentative sampling of ancestries is an acute unsolved challenge in major spheres of human disease genomics, including GWAS [24–26] and transcription models [27–29]. While big data genomic resources have revolutionized the study of cancer and other diseases, the systemic imbalance in cancer

genomic data collection across human populations is severe. Given that human ancestry has a substantial impact on gene expression (see Fig. 1C) in healthy and disease tissues [30–33], it is essential that we consider population-specific variation when modeling disease in order to improve cancer treatment for all populations. Access to diverse data is critical for creating robust precision medicine and answering scientific questions in the field [2].

3 PhyloFrame unifies ancestry-specific disease signatures

A recent analysis of cell line data estimated that only 5% of existing transcriptomic data is of African ancestry [34], while another study observed that TCGA cancers have a median of 83% samples of European (EUR) ancestry (range 49-100%) [9]. Recent studies in GWAS data have shown that model efficacy is inversely correlated with population sample size; populations with little or no representation in the data have larger disparities in the disease model performance and garner little benefit from the benchmark disease model [35]. Addressing ancestry imbalance to improve benchmark performance will require several years of dedicated large-scale sequencing efforts to address. Equitable AI approaches can help bridge the gap, and we demonstrate the feasibility of this approach. PhyloFrame, our equitable AI framework, adjusts disease signatures based on population-level genetic variation data (Fig. 2), resulting in ancestry-aware signatures that generalize to all populations, even those not represented in the disease data set. It mitigates the effects of ancestry imbalances in individual disease studies and improves outcomes for all ancestries (Fig. 3A-E).

3.1 Ancestry-specific signatures share pathway-level dysregulation

To distinguish ancestry-specific and disease-specific signatures, we first sought to identify differences when training a model using data from different ancestries. We selected the two largest TCGA BRCA populations, European (EUR; 665 samples, 107 basal and 558 luminal) and African (AFR; 90 samples, 37 basal and 53 luminal), and trained an elastic network to predict basal versus luminal breast cancer subtypes. Two models were trained, one using the EUR data and one using the AFR data. Fig. 1C shows the two resultant signatures of disease projected onto the HumanBase mammary epithelium functional interaction network. While there is little direct overlap in the signatures, the network projection highlights the shared pathway level dysregulation and the potential to use functional interaction networks to link together ancestry-specific disease signatures. This suggests that the ancestry-specific disease signatures are strongly interconnected, and that it is possible to create equitable signatures of disease without patient ancestry information. It potentially negates the need for sufficiently large training datasets from each human ancestry population across the globe, necessary to train ancestry-specific models. Instead, it suggests that equitable signatures of disease can be detected from models trained on unbalanced ancestral training data.

3.2 Identifying ancestry-specific genetic variants

To guide the network-based signature and eliminate bias toward over-represented ancestries, we identify ancestry specific variants to target during network propagation. We expect that genes with high variance among ancestries are more likely to underly ancestry specific variation, and important disease causing functions will be under-represented in existing European biased databases. To identify these loci, we define Enhanced Allele Frequency (EAF), a statistic to identify population-specific enriched variants relative to other human populations (see Methods). EAF captures population specific allelic enrichment in healthy tissue, so higher EAF means that individuals from a population are more likely to have a variant than individuals from all other ancestries. Because EAF is calculated from healthy tissue, this approach can integrate information from under-represented populations that are not present in a smaller disease specific databases, including TCGA. To confirm this, we compared EAF in COSMIC cancer-related genes to non-COSMIC genes and did not find an enrichment in EAF (Supplementary Fig. 4; t.test, p-value = 1), as expected, given EAF is calculated from healthy tissue. The average EAF across COSMIC genes is greatest in African ancestry $3.39e-4$, whereas in over-represented Europeans average EAF is only $-2.87e-5$, consistent with the greater genetic diversity found in African compared to other continents [18]. This highlights the importance of broader sampling, as most of the standing variation in humans is derived from under-sampled populations.

3.3 Integrating population and disease data

PhyloFrame combines EAF with tissue-specific big-data functional interaction networks to amplify disease-specific information from a user-provided data set (Fig. 2). Analyses in this study uses population-level variation data from gnomAD [36] and tissue-specific functional interaction networks from HumanBase [37]. However, the method accepts alternatives if provided by the user. For any disease, PhyloFrame needs solely the gene expression data and disease subtype information for each sample, allowing for easy application to many diseases. PhyloFrame does not require ancestry information for the training data samples, nor does it assume that individuals are of a single ancestry. Not needing to calculate ancestry is a significant boon; not only is it a computationally intensive task, the methods are recent and constantly evolving. We eliminate one potentially significant bias in equitable AI approaches by not needing to calculate ancestry on the training data and not using it in the model training.

PhyloFrame uses a logistic regression model with LASSO penalty to obtain an initial set of disease-relevant genes. It then projects this set of genes into the HumanBase functional interaction network most applicable to the diseases' tissue of origin, extending the network to include the first and second neighbors of each gene. These genes are then filtered by EAF, which is used to rank the set of genes identified during network propagation to determine a small set of equitable AI genes. Equitable AI genes identified by EAF are sorted by variation in the training data, and a subset of genes with high EAF and gene expression variability are selected to be added to the equitable AI signature. PhyloFrame then retrains the disease signature regression model and

forces the inclusion of the equitable genes using a ridge regression penalty. The resultant signature returned to the user now includes disease-relevant genes that generalize to all populations. We created a benchmark for comparative analyses using the exact LASSO implementation and signature size as PhyloFrame but without network propagation and EAF amplification. See Methods for complete method implementation details of PhyloFrame and the benchmark comparison models.

3.4 Ancestry-aware equitable AI disease models improve precision medicine for all

We applied these pan-population generalized signatures to subtype prediction and assessed the predictive power of models trained on different data sets. The breast cancer (BRCA) models were trained to predict Basal versus Luminal subtypes, the thyroid (THCA) models whether a tumor would metastasize (M0 versus MX), and the uterine (UCEC) models Endometrioid versus Serous subtypes. We divided samples from each cancer by ancestry (EUR, AFR, East Asian; EAS, or Admixed; ADMIX). Then we selected a training group size (14-48 samples) such that the smallest ancestry group would still be represented. Multiple models were trained on the over-represented populations to maintain training set sizes while incorporating all individuals. We also employed a 'Mixed' ancestry classification wherein each of the ancestry groups above was represented in proportion to their occurrence in the TCGA data. This resulted in many models for some populations, providing a basis for intra-population variation comparisons.

We calculated AUC for each training set under the PhyloFrame and benchmark models to assess predictive power (Fig. 3A-E, Supplementary Fig. 2). In thyroid and uterine cancer, PhyloFrame outperformed the benchmark across training data types. PhyloFrame's performance advantage is greatest in analyses of the Mixed models (Fig. 3A-B, Supplementary Fig. 2). Despite being trained on a diverse set of samples, the benchmark model has worse performance (mean AUC: THCA: benchmark = 0.63, PhyloFrame = 0.73; UCEC: benchmark = 0.95, PhyloFrame = 0.96) in AFR test data. PhyloFrame's performance advantage over the benchmark is least pronounced in thyroid cancer models trained on EUR samples (Fig. 3C). However, this effect is stratified relative to overall model performance; the benchmark performs slightly better when both models perform poorly (AUC <0.5), but as model performance improves the AUC differences between PhyloFrame and the benchmark increases, and PhyloFrame becomes more effective (Fig. 3C, mean AUC in >0.5 models PhyloFrame = 0.67, benchmark 0.64). Overall we observe a clear trend toward improved predictive performance in PhyloFrame relative to the benchmark across training sets and test data.

3.5 Equitable AI creates more stable disease signatures

Maximizing the amount of model input data is critical to optimizing model stability, minimizing overfitting and identifying a biologically-relevant signature of disease. Doing this requires representative information from all forms of the disease, or if this is not feasible, an understanding that the resultant model will be biased toward what

is represented in the data. The stability of disease signatures as the training data is changed is a key metric of model effectiveness. For example, while all of the BRCA models have high AUC (0.99-1), there is significantly less overlap in the disease signatures identified by the benchmark model compared to the PhyloFrame models trained on the same data (Fig. 3G; t-test, $p-value = 6.494e-06$). PhyloFrame models trained on EUR data have, on average, 47% overlap with other PhyloFrame model signatures, whereas benchmark models have 2% overlap with other benchmark model signatures. The greater signature stability between training sets observed in PhyloFrame relative to the benchmark suggests that PhyloFrame is less impacted by overfitting to the training data.

PhyloFrame models are also much more consistent in COSMIC gene identification, identifying 34 unique COSMIC genes (see heatmaps in Fig. 3H,I), compared to 145 COSMIC genes identified by benchmark models. These genes are shared at much higher rates between models in PhyloFrame; 76% COSMIC genes identified in PhyloFrame models are found in multiple signatures compared to only 21% COSMIC genes identified in more than one benchmark signature. Each PhyloFrame BRCA model signature identifies more COSMIC genes than its benchmark counterpart (13 vs 8 COSMIC genes on average; Supplementary Fig. 3, t.test, $p-value = 1.079e-06$). Additionally, the 5 COSMIC genes most frequently identified by benchmark models are more frequently identified by PhyloFrame models (Fig. 3I; 80% vs 50% of models). While the benchmark models most frequently identify canonical BRCA genes such as GATA3 and FOXA1, these are identified by the benchmark at lower rates than they are identified by PhyloFrame (Fig. 3I). COSMIC genes captured by PhyloFrame's BRCA model (Supplementary Fig. 3) consistently have a higher EAF in ancestries not found in the training data, most significantly in South Asians (t.test, $p = 1.126e-10$). That PhyloFrame also identifies other COSMIC genes at even higher rates than GATA3 and FOXA1 suggests that those genes are deserving of further enquiry for their role in breast cancer, particularly in non-European populations.

4 Acute Challenges: Severely underrepresented ancestries

4.1 Admixture illuminates precision medicine inequities

Admixed ancestry is widespread and likely to increase in our increasingly interconnected global society. We explored the impact of admixture on model performance in the TCGA BRCA dataset, limiting our analysis to African and European ancestries, (AFR, EUR and ADMIXED individuals with majority African or European Ancestry) as only these groups have a sufficient number of admixed individuals for meaningful analysis. Using models trained on EUR samples, we assessed model recall for PhyloFrame and the benchmark in relation to individual admixed ancestry (Fig. 4, Supplementary Fig. 6). Overall AUC for the benchmark and PhyloFrame is very high (>0.99) for all models and all ancestry test groups (Fig. 4A). Model performance is similarly high and stable across individuals with majority European ancestry, and PhyloFrame slightly outperforms the benchmark across admixture levels (Fig. 4B).

However, in individuals with majority African ancestry, PhyloFrame provides significantly higher recall than benchmark models when looking across ancestry levels (Fig. 4C). The performance of both PhyloFrame and the benchmark increase significantly as admixture levels increase. We hypothesize that the improvement in model performance with increased admixed ancestry is a product of ancestral bias in the training data. The preponderance of admixed ancestry in African American populations is European in origin [38], (Supplementary Fig. 1). As such, performance of the models trained on EUR data improves as the individuals' fraction of European ancestry increases. PhyloFrame does not entirely mitigate this inequity, but it performs substantially better than the benchmark in mitigating the inequalities observed in predicting disease state in individuals with majority African ancestry (Fig. 4C).

4.2 External Validation: Triple negative breast cancer in African populations

To externally validate the efficacy of PhyloFrame relative to the benchmark we analyzed triple negative breast cancer (TNBC) data from Martini et al [30], comprised of 9 African Americans, 6 Ghanaians and 11 Ethiopians, totaling 26 patients (Fig. 5A). We chose this dataset for external validation because it provides a unique opportunity to test the efficacy of our models in severely underrepresented populations. Individuals with ancestries that are not well represented in wealthy countries in particular are at greater risk of receiving lower quality precision medicine care under non-ancestry aware methods (e.g. the benchmark). Africa is the most severely impacted continent in this regard due to a combination of economic disadvantages and greater genetic diversity than the rest of the continents combined [18].

Ghanaian and Ethiopian populations are more genetically divergent than any pair of non-African populations [18]. Additionally, the majority of (already underrepresented) African ancestry individuals in genomic datasets are members of the African diaspora [9, 38]. The sizable majority of enslaved people taken to the United States during the transatlantic slave trade were taken from Atlantic Africa, a region including Ghana and several neighboring countries [38–40] (Supplementary Fig. 1). However, it is crucial to stress that Ghanaian and African American ancestry are not equivalent terms; African ancestry in African American populations is diverse and, beyond this diversity, an average of 15% of African American ancestry is European in origin [38]. Under-representation in disease genomics databases is even worse for populations from other regions of Africa, such as Ethiopians, who would be naively classified as African but are genetically very distinct from all of the TCGA training data populations [18]. Ethiopia and nearby countries contribute only 1-2% of the total African American ancestry pool. As such, this ancestry is not evenly distributed across African Americans, the vast majority of whom have little to no East African ancestry, while a small fraction have near 100% East African ancestry [38] (Supplementary Fig. 1).

We applied the trained PhyloFrame and the benchmark models to each validation population to predict TNBC status. Among African Americans, PhyloFrame and benchmark models' recall was similar when comparing models trained on datasets including individuals with African ancestry (AFR, ADMIX, MIXED). When comparing models trained with no African ancestry individuals, however, PhyloFrame models

trained on data that included individuals with African ancestry generally performed better than all other models (Fig. 5B). In African populations PhyloFrame consistently achieves high levels of recall whereas the benchmark produces with both high and very low recall (Fig. 5C-D). This effect is most pronounced in Ethiopians (median recall P 0.82, B 0.73), who are not well represented by any TCGA ancestry groups (Fig. 5D).

5 Towards a more equitable precision medicine future

Unbalanced population diversity within accessible genomic data has led to unintended bias in precision medicine models[2, 33, 41–44], resulting in disparate effectiveness in populations and sub-optimal treatment options[5, 7, 45, 46], and contributing to the inequality of medical resources[4, 6, 7, 47–50]. PhyloFrame mitigates these issues through big-data equitable AI, creating genomic signatures of disease that are equally effective in all populations regardless of training data available. This not only brings under-sampled populations to equality with better sampled populations but provides better performance to all populations, including over- and under-represented populations. Integrating population genomic data with disease oncology presents a unique opportunity for novel approaches in the examination of fundamental mechanisms of cancer.

We demonstrate PhyloFrame's ability to significantly reduce model overfitting caused by ancestral bias in the training data, and to create disease signatures that work better for all individuals. PhyloFrame signatures are more consistent across training sets (demonstrating higher biological relevance) and consistently detect known cancer-related genes. Unlike a comparable benchmark, PhyloFrame is effective regardless of whether an individual comes from a population represented in the training data. It performs well in individuals of divergent unsampled populations and varying levels of admixture.

PhyloFrame occupies a critical, and to this point unfulfilled, niche in the broader movement towards more equitable precision medicine. It serves as a catalyst to improve precision medicine equity, alongside improved genomic references that incorporate population structure and diversity (such as the Human Pangenome Reference Consortium [51, 52]) and sequencing efforts (such as the Nigerian 100K Genome Project [53]). PhyloFrame demonstrates the feasibility of equitable genomic AI approaches. Combined with continually improving genomic resources and data, it will further enable a future of equitable precision medicine, where all individuals can trust that models will accurately predict their genomic information. This represents a major step forward toward a future of equitable genomic AI.

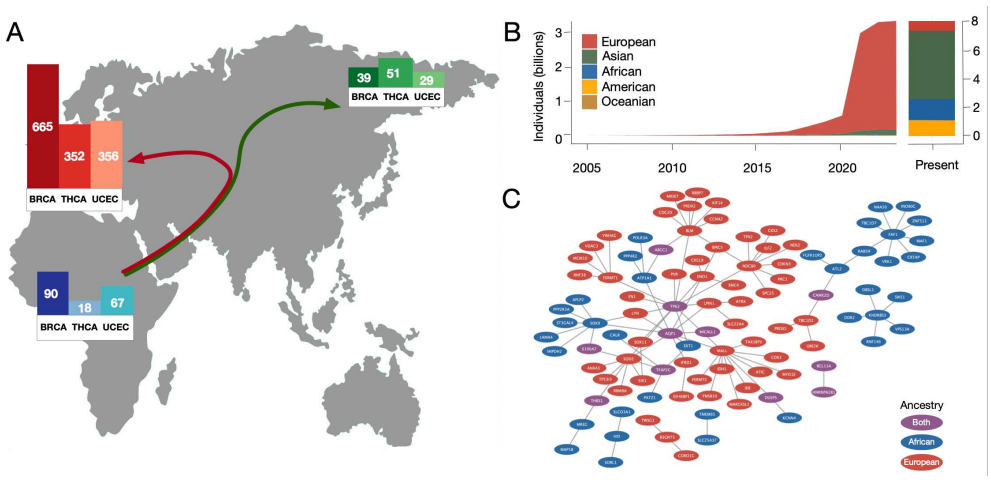


Fig. 1 Human diversity and cancer genomics. (**A**) Global map showing the number of samples and sample diversity within three of the most ancestrally diverse TCGA cancers (breast, thyroid, uterine). (**B**) Ancestry statistics for historical GWAS over time, using data from the GWAS Catalog [16]. (**C**) A projection of the EUR BRCA signature of disease (red) versus the AFR BRCA signature of disease (blue) onto a functional interaction network highlights the differences caused by ancestry bias in expression data, signature inter-relatedness, and how this impacts cancer-related precision medicine. EUR and AFR signature overlaps are shown in purple.

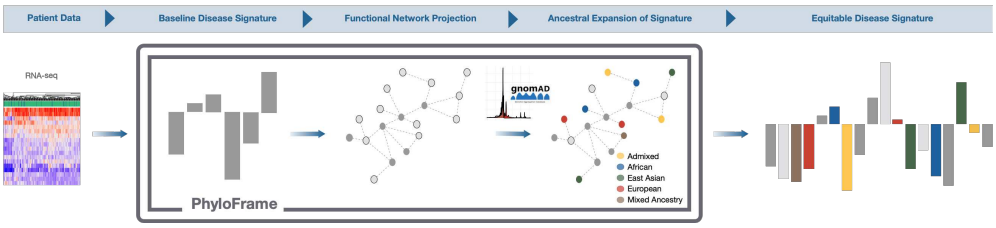


Fig. 2 PhyloFrame equitable AI approach. PhyloFrame method takes as input a gene expression matrix of patient disease data and patient outcome labels. Using this information, it calculates an original gene signature, projects that onto the functional interaction network, then identifies closely interacting genes that are enriched in human populations. Using this information, PhyloFrame calculates an updated equitable signature of disease, which is output to the user.

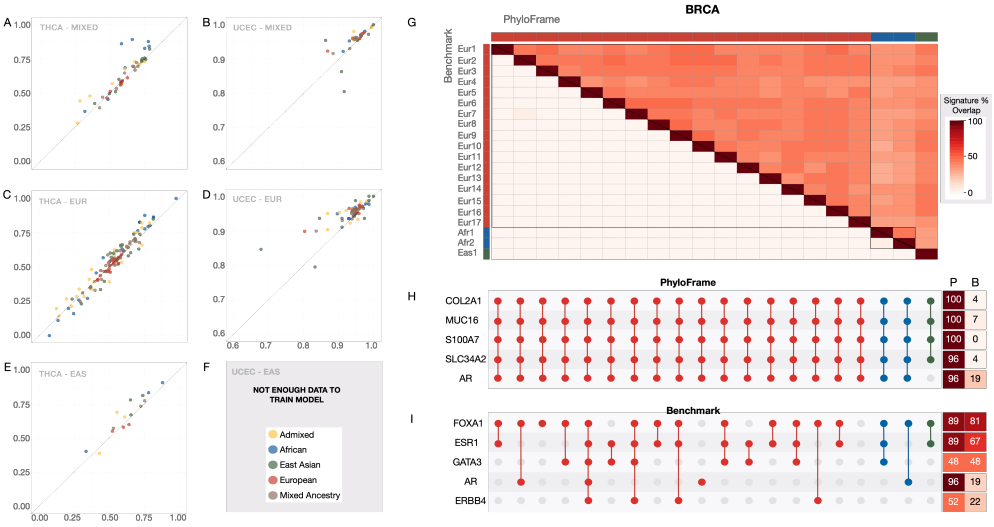


Fig. 3 Equitable AI effectiveness. AUC of the benchmark versus PhyloFrame models when training in (**A,C,E**) thyroid cancer (**B,D**) and uterine cancer when using different populations for the training and validation data and varying the ancestral population of the training data. Top row shows results when training using a representative global population, descending rows show training on each cancer using a single human ancestral population. All training sets are randomly sampled from the set of individuals from each ancestry so that all models within each cancer type are trained on the same number of samples. For populations with more samples, we randomly select samples repeatedly until all samples are used in at least one test, resulting in larger numbers of results in different populations. Each scatter plot has AUC of PhyloFrame versus benchmark when testing within an ancestral population. Color coding indicates the test data ancestry and plots are grouped by training data ancestry. (**G**) Signature-signature correlation of BRCA disease signatures when trained on different populations and (**H,I**) Overlap of COSMIC cancer genes in these signatures in the (**H**) PhyloFrame and (**I**) benchmark models. The heatmaps in **H,I** show the number of models in PhyloFrame (P) and benchmark (B) that include the given COSMIC gene.

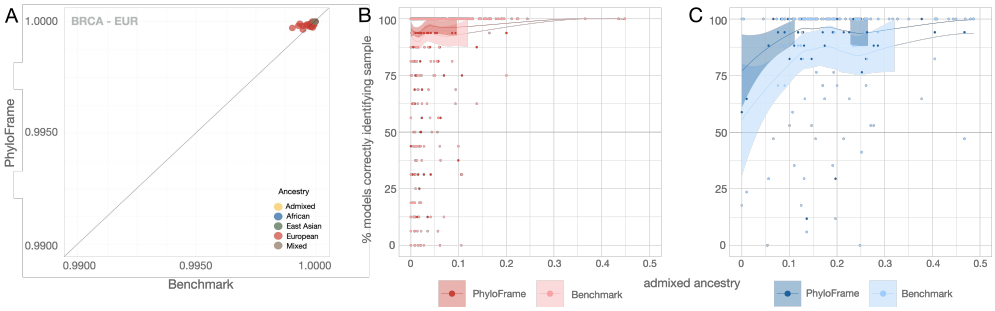


Fig. 4 Admixture affects model performance. (**A**) AUC scatter plot of PhyloFrame and benchmark performance in BRCA. (**B-C**) Comparison of models trained on EUR BRCA data and the percent of correctly predicted held-out validation set samples as admixture levels increase in (**B**) PhyloFrame and (**C**) the benchmark model.

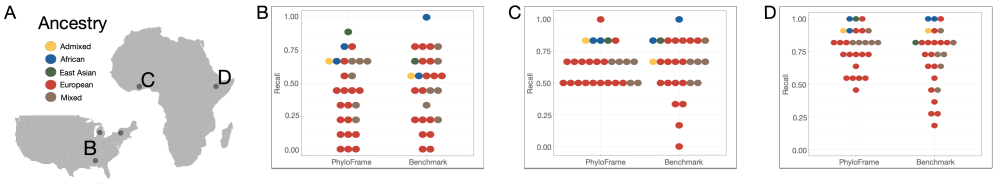


Fig. 5 Validation in a global model. ((**A**) A global map of the BRCA validation data sampling sites. PhyloFrame and benchmark model performance in the African breast cancer validation set in each of the sampling sites (((**B**) US, ((**C**) Ghana, ((**D**) Ethiopia).

6 Materials and Methods

PhyloFrame equitable AI framework

PhyloFrame provides a flexible platform for ancestry-aware analyses of gene expression data. Here we describe the core features and implementation of PhyloFrame. First, population variation data is downloaded from the selected repository (in this manuscript we use gnomAD v2.1.1, see below). Second, functional interaction network(s) are acquired; we use HumanBase tissue-specific functional interaction networks. These two big-data resources are internal to PhyloFrame. We provide processing scripts to create the correct formats (see Code Availability). PhyloFrame has the option to use user-specified information but these are not required. PhyloFrame defaults to using the HumanBase networks and gnomAD data that we provided in already processed formats. PhyloFrame's one required input is: gene expression data for a set of patients, and outcome labels with which to train a binary classification model. With this information, PhyloFrame creates a generic disease signature, then uses the population genomics data and functional interaction networks to modify that disease signature to be equally effective across all global populations. PhyloFrame returns this equitable disease signature, including weights describing the model importance for each gene. This process is described in detail below.

Population Genomics Compendium

PhyloFrame filters genes in each ancestry using sequence variation information from The Genome Aggregation Database (gnomAD) [36]. GnomAD compiles data from many extensive sequencing projects around the globe and is the largest reference population database currently available. The v2.1.1 data set used by PhyloFrame includes 125,748 exome sequences from diverse, unrelated individuals. PhyloFrame uses the the following 17 ancestries: African/African American, Latino/Admixed American, Ashkenazi Jewish, East Asian, European (Finnish), European (non-Finnish), South Asian, and Other. East Asian and European (non-Finnish) are broken down further into East Asian Korean, East Asian Japanese, East Asian Other East Asian, European (non-Finnish) Bulgarian, European (non-Finnish) Estonian, European (non-Finnish) North-Western European, European (non-Finnish) Southern European, European (non-Finnish) Swedish, and European (non-Finnish) Other.

GnomAD information is re-processed using the same pipelines as the gnomAD team, creating a harmonized repository of population-specific genomic information. The gnomAD v2.1.1 exome variant calling format (VCF) file for all chromosomes was downloaded via AWS CLI from gnomAD's Downloads page. We extract allele frequencies (AF) from gnomAD, for use in our statistical analysis (described below). For all of our analyses we use the full gnomAD dataset, not one of the provided subsets (e.g. pediatric). Variant information for each ancestry as well as the gene information for each Single Nucleotide Polymorphisms (SNPs) were parsed in C++ using the VCF operations library htslib and written to a new VCF file. For each SNP the new VCF contains the allele count, allele number, and allele frequency for each of the 17 ancestries as well as the chromosome, position, rs id, reference allele, alternate allele, gene, consequence, impact, and distance. Allele counts, numbers, and frequencies are

not sex specific within the ancestry. The PhyloFrame pipeline for population data is implemented in C++and R.

Enhanced Allele Frequency

To calculate ancestry enrichment for each SNP, we create a statistic called Enhanced Allele Frequency (EAF). Enhanced allele frequency (EAF) aims to identify Single Nucleotide Polymorphism (SNPs) in exons that vary between global populations. In the current study, EAF is calculated using GnomAD sequence variation data. For each SNP we calculate enhanced allele frequency for each ancestry (17 ancestries total) as follows:

$$EnhancedAFAncestry_1 = AFAncestry_1 - mean(AFAncestry_2, ..., AFAncestry_{17}).$$

Enhanced allele frequency shows which SNPs are enhanced, i.e. occur more frequently, in which ancestries and helps PhyloFrame target ancestry-specific genes that may affect the way an ancestry responds to given a disease. High EAF indicates that the given variant is more frequently altered in that ancestry, compared to other ancestries. Density plots of the resulting enhanced frequencies in each ancestry demonstrate that while most SNPs are negatively enhanced, there are ancestry unique peaks of enhanced SNPs (Supplementary Fig. 4).

EAF cutoffs were tested at several thresholds (EX: solely selecting genes with high EAF (0.5 - 1) or selecting genes with low EAF (0.00001 - 0.1)), however PhyloFrame did best with cutoffs between these extremes at 0.001 - 1. The higher threshold prevents highly ancestry specific mutations from being excluded while the lower threshold prevents filtering out major cancer genes, such as FOXA1, that have a relatively lower EAF across ancestries.

Functional Network Integration

For each disease, HumanBase tissue-specific functional interaction networks [37] were used to find genes likely to interact with the identified disease-associated genes. After the disease relevant tissue-specific network has been downloaded, Entrez IDs are mapped to Gene Symbols in R using Bioconductors' genome wide annotation for Humans (org.Hs.eg.db, Bioconductor version 3.17).

Genes from PhyloFrame's baseline elastic net run were used as start nodes in the search for disease relevant gene interactions in the tissue-specific network. First and second neighbors of the baseline signature with a connection between 0.2 - 0.5 are kept for further ancestry allele sorting. By projecting high profile disease associated genes into an interaction network, PhyloFrame is able to begin its ancestry search in genes with high involvement in the disease network.

We trained PhyloFrame on three diseases from the TCGA PanCancer Atlas [9–12]: breast cancer, thyroid cancer, and uterine cancer. These cancers were chosen due to their relatively diverse patient populations. PhyloFrame used the most relevant HumanBase tissue-specific functional interaction networks for each disease: mammary epithelium for breast cancer, thyroid gland for thyroid cancer, and uterine endometrium for uterine cancer. PhyloFrame network cutoffs were initially selected based on a grid search within each disease to find the optimal number of neighbors in the interaction network as well as the optimal edge weight for the interaction.

PhyloFrame was run on every combination of the grid N × E where N = {1, 2, 3} defining any gene within three neighbors of a top 20 mutated gene in the disease and E = {0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, .95, 1} is defining the minimum edge weight needed for a neighbor to be included. PhyloFrame was run on this grid search-for every ancestry in each disease, the combination with the highest overall average across all ancestral models was selected. Based on this search, we opted to included from the functional network the first and second degree neighbors of the original signature genes. We later updated the edge weights to keep nodes with a weight between 0.2-0.5, following analysis of the functional interaction between the European and African TCGA BRCA disease signatures.

Equitable loss function for machine learning

PhyloFrame builds ancestry unbiased disease signatures by minimizing loss in two domains: relevance to the given disease and enrichment in at least one of the seventeen ancestries described by gnomAD. To define the disease-related genes, we find genes that are functional network interaction partners with the top n most mutated genes related to the disease. Fig. 1C shows the related signatures for breast cancer subtype prediction when training in two ancestral populations. To create the PhyloFrame signatures for this task (results not shown), we created a signature of disease that does not incorporate ancestry, and instead simply calculates relationship within the training data samples to the disease outcomes. We use these genes as start nodes in the search for disease-relevant gene interactions in the tissue-specific HumanBase predicted gene interaction network. These genes are used as a base to put into a tissue-tissue interaction network. PhyloFrame performs a grid search to find fully connected networks within the undirected tissue network graph. Genes are selected if they are within y neighbors of the disease genes, narrowing the genomic landscape by selecting genes enhanced in each ancestry for genes relevant to the disease by looking at the enhanced allele frequency, then logistic regression with a ridge penalty to select critical genes.

Integrating ancestry and networks

For any disease, PhyloFrame requires the ancestry-specific EAFs for all genes, the tissue associated functional interaction network, patient gene expression data, and patient clinical data. Using the EAF information and functional interaction network, PhyloFrame first finds the neighborhood of the top enhanced genes by EAF, using node and edge parameters previously found in each diseases' grid search. The resulting neighborhood of genes are mapped to their associated single nucleotide polymorphisms using the previously calculated enhanced allele frequencies. We narrow down the disease's genomic landscape further by ordering the enhanced allele frequencies for each ancestry separately and targeting the unique allele frequency in each ancestry with the most enhanced SNPs. The genes included in this slice for each ancestry are assembled and prioritized in PhyloFrame's regression task. To do this, PhyloFrame first narrows the genomic landscape by selecting genes enhanced in each ancestry for genes relevant to disease in each ancestry by looking at the enhanced allele frequency, it then uses logistic regression with a ridge penalty to select highly important genes. It returns to the user an equitable gene signature of disease.

Data Description

All analyses for this study were conducted on published datasets. Population genomics data from gnomAD (gnomad.broadinstitute.org) was downloaded on July 22, 2022. We downloaded functional interaction networks for model generation from HumanBase (hb.flatironinstitute.org); the mammary network was downloaded on July 5th 2022 and uterine and thyroid networks were downloaded on December 21st 2022. Predictive model training (PhyloFrame and benchmark) and all analyses except the validation experiment (Fig. 5) use gene expression data and matching clinical data from the TCGA Pancancer Atlas for breast (BRCA), thyroid (THCA) and uterine (UCEC) cancers. Data for all TCGA samples was accessed from cBioPortal (www.cbioportal.org) on December 6, 2022 (BRCA), Dec 19, 2022 (THCA), and Dec, 22 2022 (UCEC). Ancestry information for the TCGA cancer data was obtained from Carrot-Zhang et al [33]. TNBC data for the BRCA model validation is from Martini et al [30]. COSMIC data was downloaded on Jan 11, 2023 from the COSMIC website (cancer.sanger.ac.uk/cosmic). It contains 736 genes and information on which cancers each gene has been associated with.

Data batching for model training

To account for the unequal sample sizes between ancestries, we split the data into batches. To prevent bias when training models due to the disproportionate amount of EUR samples, samples for each disease were split into batches. We also created batches of mixed ancestry, unbalanced ancestry, and single-population ancestry to highlight the effects of PhyloFrame on balanced (unbiased) training data. Number of batches and number of samples in each batch for a given ancestry were determined by the ancestry with the smallest sample size that could still successfully train a model in that disease. All sample sizes must be at least 9 samples and contain at least 4 samples in each subtype. The ancestry with the smallest sample size was used as the base to create other ancestry batches (see Supplementary Table 2 for number of samples per ancestry in each TCGA cancer). The number of samples in an ancestry was divided by the smallest sample size to find the number of batches the given ancestry would be split into. The subtypes in the ancestry were then separated and both were separately divided by the number of batches to find the number of samples from that subtype should be added to each batch. Admixed samples were grouped together regardless of primary ancestry. Overflow samples from each subtype and cancer were added one by one to each batch until all samples were assigned to a batch. For BRCA, a total of 842 basal and luminal samples divided using this approach resulted in 27 total batches; 17 EUR, 2 AFR, 1 EAS, 6 Mixed, and 1 Admixed of 38-48 samples per batch. THCA has data for 436 MX/M0 samples, which resulted in 37 total batches; 23 EUR, 1 AFR, 3 EAS, 9 Mixed, and 1 Admixed batch of 14-18 samples each. UCEC has data for 491 Serous and Endometrial samples total, which resulted in 22 total batches; 12 EUR, 2 AFR, 1 EAS, 6 Mixed, and 1 Admixed batch of 29-39 samples each. No other ancestries had enough samples to train a model, in any of the three cancers.

We used this approach to split the data into equal sized batches of samples, based on the smallest ancestry and cancer type. In total, there are 27 BRCA, 37 THCA,

and 22 UCEC training data batches. PhyloFrame and the benchmark models were trained on all sample batches, and applied to the remaining (held out) samples. For BRCA models we also tested on an external validation set [30] (described in *External Validation: Triple negative breast cancer in African populations*).

Model Validation Using External Data

To externally validate our model we chose to assess a dataset that both was outside of the training set used in the development of PhyloFrame and that provided an opportunity to assess the performance of PhyloFrame and the benchmark on ancestry groups not present in the training data. To meet these objectives we analyzed triple negative breast cancer (TNBC) data from Martini et al [30], comprised of 9 African Americans, 6 Ghanaians and 11 Ethiopians, totaling 26 TNBC patients (Fig. 5A).

As with the previous analyses of breast cancer, PhyloFrame and the benchmark were tasked with classifying the samples as either basal or luminal. We applied the PhyloFrame and benchmark models trained using the same subdivisions of the TCGA BRCA data described above, resulting in 27 models (17 EUR, 2 AFR, 1 EAS, 1 ADMIXED and 6 MIXED). These trained models were applied to an external validation set, the Martini et al [30] TNBC data. Most basal breast cancers are also TNBCs, and the terms are often used interchangeably. Thus successful models should predict all of the validation set samples to be basal, as they are TNBCs. Because triple negative breast cancers are basal, accuracy functionally reduces to the proportion of the samples in each population that the models correctly identify as basal.

Supplementary information.

Supplementary Information

Model performance in three cancers

Model performance was compared using AUC, recall, and signature overlap. Each metric and associated results are described in the paragraphs below. We tested the efficacy of PhyloFrame against the benchmark for multiple cancer prediction tasks.

Subtype prediction in breast cancer (BRCA)

Cancer subtype prediction is a key component of directing the treatment of breast cancer. We trained 27 models on data from 842 individuals, divided into training batches of 38-48 samples each, to predict Basal vs Luminal subtype. The number of training sets correspond to the proportion of samples in the TCGA BRCA database resulting in 17 EUR, 2 AFR, 1 EAS, 1 ADMIXED and 6 MIXED models. Both PhyloFrame and the benchmark use the HumanBase mammary epithelium network for network construction. PhyloFrame uses population data from gnomAD.

For each model (PhyloFrame or benchmark) we predicted subtypes for entire populations and calculated AUC. If a model was trained on the same population being tested (eg. EUR trained model being tested in EUR), we excluded the samples used for training from the larger set. We plotted AUC results for the benchmark compared to PhyloFrame (Supplementary Fig. 2A-E).

Metastasis prediction in thyroid cancer (THCA)

Metastases have a substantial impact on patient prognosis, in this analysis we predicted whether patients would undergo metastasis (M0 versus MX). We trained 37 models on data from 436 individuals, divided into training batches of 14-18 samples each. The number of training sets correspond to the proportion of samples in the TCGA THCA database resulting in: 23 EUR, 1 AFR, 3 EAS, 1 ADMIXED and 9 MIXED models. Both PhyloFrame and the benchmark use the HumanBase thyroid gland network for network construction. PhyloFrame uses population data from gnomAD.

For each model (PhyloFrame or benchmark) we predicted subtypes for entire populations and calculated AUC. If a model was trained on the same population being tested (eg. EUR trained model being tested in EUR), we excluded the samples used for training from the larger set. We plotted AUC results for the benchmark compared to PhyloFrame (Supplementary Fig. 2F-J).

Subtype prediction in uterine cancer (UCEC)

Cancer subtype prediction is a key component of directing the treatment of uterine cancer. We trained 22 models on data from 491 individuals, divided into training batches of 14-18 samples each, to predict Endometrioid vs Serous subtype. The number of training sets correspond to the proportion of samples in the TCGA UCEC database

resulting in 12 EUR, 2 AFR, 1 EAS, 1 ADMIXED and 6 MIXED models. Both PhyloFrame and the benchmark use the HumanBase uterine endometrium network for network construction. PhyloFrame uses population data from gnomAD.

For each model (PhyloFrame or benchmark) we predicted subtypes for entire populations and calculated AUC. If a model was trained on the same population being tested (eg. EUR trained model being tested in EUR), we excluded the samples used for training from the larger set. We plotted AUC results for the benchmark compared to PhyloFrame (Supplementary Fig. 2K-O).

Model stability and consistency

Precision medicine models should be identifying biologically-relevant signatures of disease. A signature may be accurate on a set of training data, but it has limited utility if it does not generalize to other data or identify the biological drivers of disease. To quantify the amount of biological overlap between signatures, we calculate several factors. First, we identify how many known cancer-related genes are identified in each signature, and which cancers they have previously been associated with. This is done using COSMIC cancer genes (see Fig. 3H,I). We calculate both the number of COSMIC genes identified by each model, and calculate a t-test comparing these numbers in PhyloFrame versus benchmark models. Second, we calculate the overlap in disease signatures for each model. Higher signature overlap is an indication that, despite different training data, the models are identifying the same factors as driving the disease. To quantitatively compare this overlap, we calculate pairwise model signature correlations. Signature correlations are calculated based on presence/absence of each gene in a model signature; we do not consider model weights for the genes. This resulted in a matrix of signatures by signatures, filled in with pairwise signature-signature correlations. PhyloFrame signatures have significantly higher overlap than benchmark models (mean 47% vs 2% overlap, Fig. 3G). We then ran a t-test comparing all PhyloFrame-PhyloFrame model pairwise correlations against all benchmark-benchmark model pairwise correlations, and found there is statistically higher likelihood of signature overlap in PhyloFrame compared to benchmark models.

Sample-specific model performance

Some samples are far more difficult to predict than others, and while overall model AUC is important, it is also valuable to see how each model performs in the more difficult sample sets. To see this, we plot the per-sample differences in model performance (% of models that correctly predict each sample; Supplementary Fig. 5), and identify which samples are often misclassified, if any. Each point in these plots represents one sample. In the boxplots, samples are grouped by ancestry and by model type (PhyloFrame vs benchmark). The y-axis shows the percent of models that correctly predict each sample. Both PhyloFrame and the benchmark models struggle to correctly predict a small subset of samples. For most of the BRCA samples, all models correctly predict BRCA subtype. A small subset of samples are incorrectly predicted by all or most of the BRCA models (see Fig. 4 for a critical factor explaining of this effect. UCEC and THCA models have far more variability, as expected, given

the models have lower average AUC than the BRCA models. Metastasis is a harder prediction task than tumor subtypes, and so many of the THCA models have low performance. Of note, this per-sample performance is not shared across models; There is a wide range of success for each set of sample predictions. This suggests that there are factors relevant to the THCA models that are not being identified by the models. It is unclear if this is a tractable prediction task, given the small training data size. UCEC models similarly have varied performance for each sample, however most samples have a high % prediction success across models. The endometrial versus serous subtype UCEC model predictions for a subset of samples are unreliable. This prediction task is difficult due to the low number of serous samples available in the training data. Serous samples are only approximately 25% of the samples, and half of those samples are from individuals of European decent. For example, an East Asian model could not be trained in uterine cancer because there were only 3 serous samples. The samples most often misclassified in this prediction task are of the serous subtype.

COSMIC gene enrichment and presence

COSMIC currently includes 736 genes expert curated and validated as cancer-related based on previous studies. We used COSMIC in two sets of analyses. First, we used COSMIC genes to identify the EAF trends of disease-related genes, to assess the effectiveness of using EAF as an equity adjustment in AI methods. We found that there is no EAF enrichment in COSMIC versus non-COSMIC genes (Supplementary Fig. 4; t.test, p-value = 1), suggesting that the utility of EAFs in equitable AI is not limited to cancer studies. Second, we used COSMIC genes to determine the extent to which model signatures in this paper recapitulate known cancer processes and as a validation set of genes that ideally will be enriched in the disease signatures. As most of the models in this paper are trained on smaller sample sizes (due to ancestry bias in the data), we expect model overfitting. Identifying the number and variety of COSMIC cancer genes in each signature helps to determine how much of the signatures are cancer-related. COSMIC genes with high EAFs are more frequently enriched in African and East Asian but not European ancestries (Supplementary Fig. 4B; t-test, p-value ¡ 2.2e-16) . For example, FOXA1 is one of the five most frequently identified COSMIC genes by the benchmark BRCA models (Fig. 3H,I). While it is not one of the most frequently identified COSMIC genes by PhyloFrame models, more PhyloFrame (89%) than benchmark models (81%) include FOXA1 in their signatures (22 vs 24 of a total 27 models).

The impact of admixture

Continental and ethnic classifiers, including those used to group samples in this study, are flawed proxies for ancestral diversity. In an increasingly interconnected world, rates of admixed ancestry are likely to increase. Even in the present day, the extent and impact of admixture within human populations remains under-recognized, especially as it pertains to underrepresented groups in the United States. We sought to understand how admixture impacts the predictive power of PhyloFrame relative to the benchmark.

To explore the impact of admixture we examined predictive efficacy of models trained on EUR breast cancer data. We selected breast cancer because the overall high AUC across all models allows us to largely exclude noise generated by poor model performance, and instead identify which individuals the models' struggle to accurately predict. We used EUR training data because having more trained models (in this case, 54 models) enables us to precisely determine whether an individual is being stochastically or systematically incorrectly predicted. The current dramatic over-representation of Europeans in genomic databases [9, 34] provides additional value for this approach as it more closely mirrors expected real world studies across disease types.

We calculated the fraction of models that correctly predicted an individual's subtype and plotted those relative to their admixed ancestry proportion. Note that this includes individuals that are otherwise classified as Admixed (greater than 20% non-majority ancestry) being grouped according to their majority ancestry. Next, we measured the statistical variance of model prediction accuracy across populations (R geom_smoothing with LOESS model). Individuals with majority European ancestry show stable prediction across admixture levels (Fig. 4B), compared to PhyloFrame improvements on individuals from many ancestries and models, most notably the BRCA EUR models applied to Admixed and AFR samples (Supplementary Fig. 5). Majority African ancestry individuals show both significant increases in model performance with increasing admixed ancestry and significantly better performance in PhyloFrame than in the benchmark. Given that the vast majority of admixed ancestry in African Americans is European, including in the individuals in this study (Supplementary Table 2), this highlights a shortcoming of current predictive methods that is easily overlooked when grouping individuals by continental level ancestry or by overall model AUC and other performance metrics. Overwhelmingly European training data sets may appear to perform acceptably in African ancestry individuals, when in fact performance is not uniformly high across the group. This raises substantial concerns surrounding existing models' abilities to provide insightful and accurate precision medicine predictions for individuals of un-admixed African ancestry.

External validation of BRCA models

To externally validate our model we chose to assess a dataset that both was outside of the training set used in the development of PhyloFrame and that provided an opportunity to assess the performance of PhyloFrame and the benchmark on ancestry groups not present in the training data. To meet these objectives we analyzed triple negative breast cancer (TNBC) data from Martini et al [30], comprised of 9 African Americans, 6 Ghanaians and 11 Ethiopians, totaling 26 TNBC patients (Fig. 5A).

Data and preprocessing

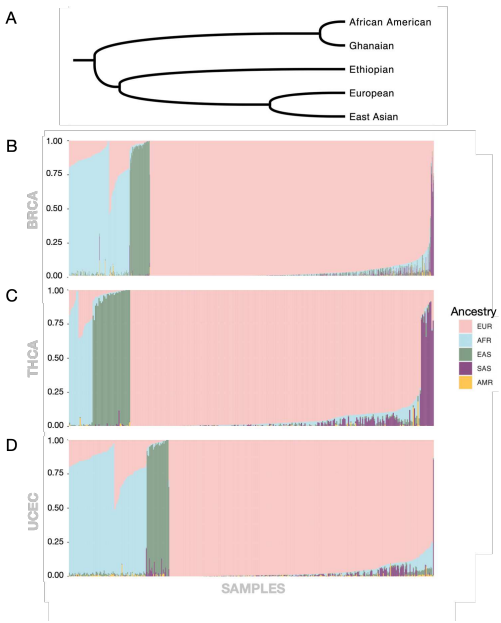
As with the previous analyses of breast cancer, PhyloFrame and the benchmark were tasked with classifying the samples as either basal or luminal. We applied the PhyloFrame and benchmark models trained using the same subdivisions of the TCGA BRCA data described above, resulting in 27 models (17 EUR, 2 AFR, 1 EAS, 1

ADMIXED and 6 MIXED). These trained models were applied to an external validation set, the Martini et al [30] TNBC data. Most basal breast cancers are also TNBCs, and the terms are often used interchangeably. Thus successful models should predict all of the validation set samples to be basal, as they are TNBCs. Because triple negative breast cancers are basal, accuracy functionally reduces to the proportion of the samples in each population that the models correctly identify as basal.

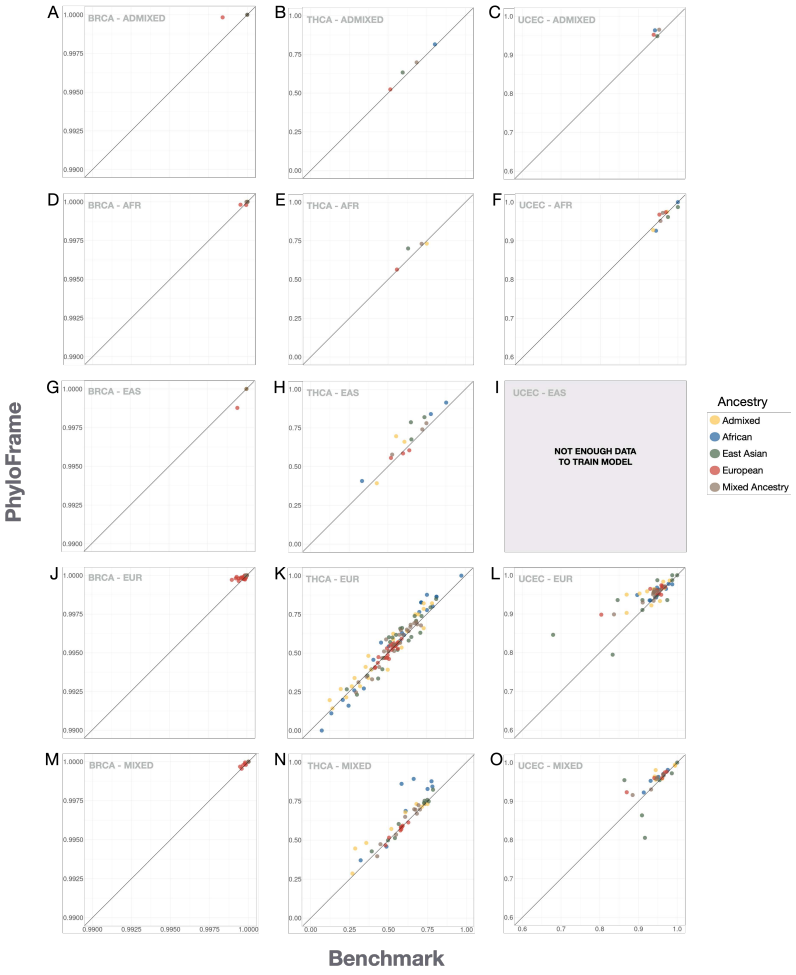
Model results

In African populations (Ghanaian and Ethiopian), all but one PhyloFrame model have performance greater than random chance accuracy (>50%). Mean PhyloFrame performance is higher than the benchmark model in both the Ghanaian validation set samples (mean recall PhyloFrame = 0.64 vs benchmark = 0.62) and the Ethiopian validation set samples (mean recall PhyloFrame 0.78 vs benchmark 0.70). Performance is more similar in the African American validation set samples (mean recall PhyloFrame = 0.42 vs benchmark 0.47).

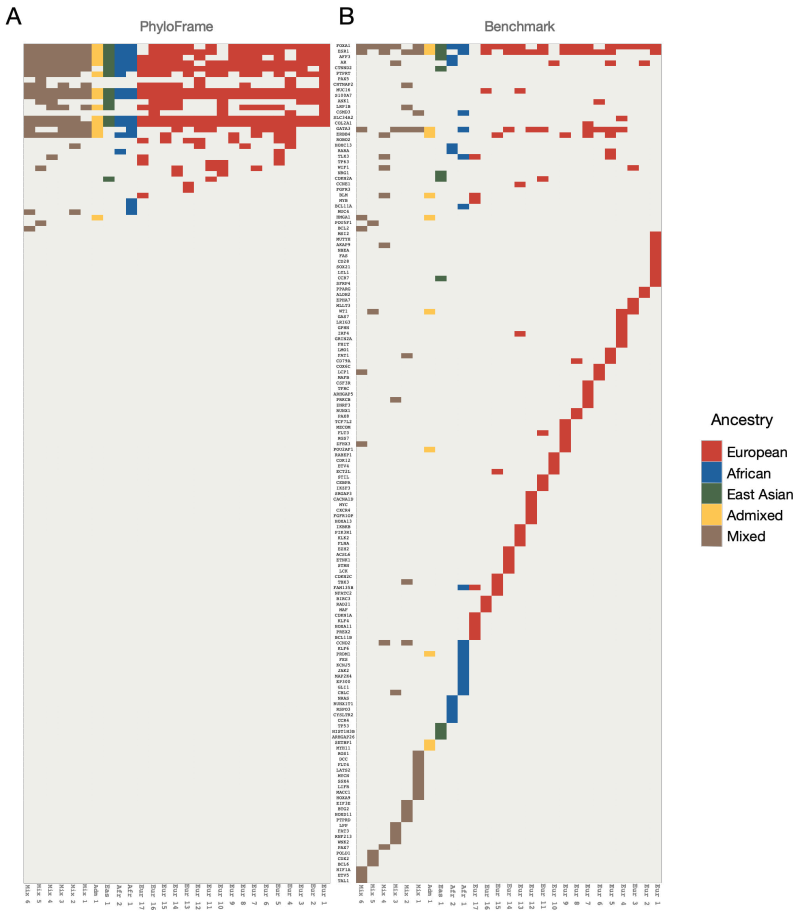
Martini et al samples from the US were collected from New York City (New York), Detroit (Michigan), Ann Arbor (Michigan), and Birmingham (Alabama). TCGA samples also were recruited from the US, and TCGA BRCA samples came from 42 tissue source sites, including several in New York City. Given that the training data includes samples from overlapping populations, it was expected that the benchmark model would perform well in the African American validation set samples, however this is not the case (median = 0.56). Both the benchmark and PhyloFrame models have highly variable performance in the validation set African American samples, suggesting that neither set of models are able to fully disentangle complexities of ancestry and breast cancer subtypes. Given that the Basal subtype is enriched in African Americans [30, 42, 54], this prediction task may be intrinsically connected to ancestry; It has been previously suggested that the Basal/Luminal subtypes are unintentionally linked to African ancestry. However, there are distinctions between the two datasets, even within African American samples, that may explain some of the variability. TCGA BRCA samples were diagnosed with BRCA from 1988-2003, compared to the BRCA validation set which began collection in 2006. Ghanaian samples average age 48 years, Ethiopian 41 years, and the African American samples 68 years [30], compared to 58 years for the TCGA BRCA samples. The age differences may account for some of the variance in the PhyloFrame and benchmark models' recall when applied to the African American samples.



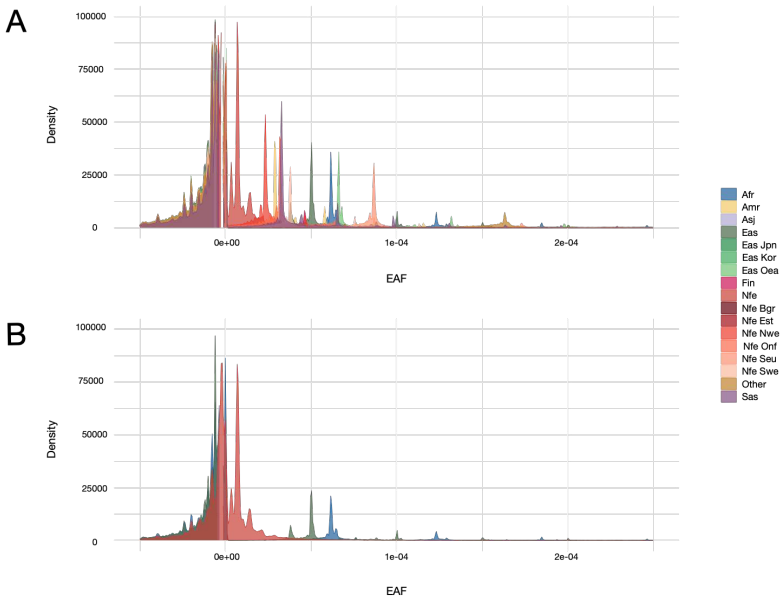
Supplementary Figure 1 African ancestry diversity. (**A**) Phylogenetic tree showing the general pattern of genetic relationships between the populations used for model training and disease state prediction in this study. (**B-D**) Estimated ancestry of each patient in the TCGA data for (**B**) BRCA, (**C**) THCA, and (**D**) UCEC. Genetic ancestry for each TCGA sample was computationally predicted by Carrot-Zhang et al [33], who used 5 ancestry-calling pipelines to generate ensemble predictions of each individual in the TCGA PanCancerAtlas. Ancestry sample counts across all TCGA cancers shown in Supplementary Table 2. Each column represents a single sample in the cohort, and the colored bar represents the estimated amounts of ancestry in each of the major global populations. Shown populations are EAS (green), SAS (purple), AFR (blue), EUR (pink), and AMR (yellow).



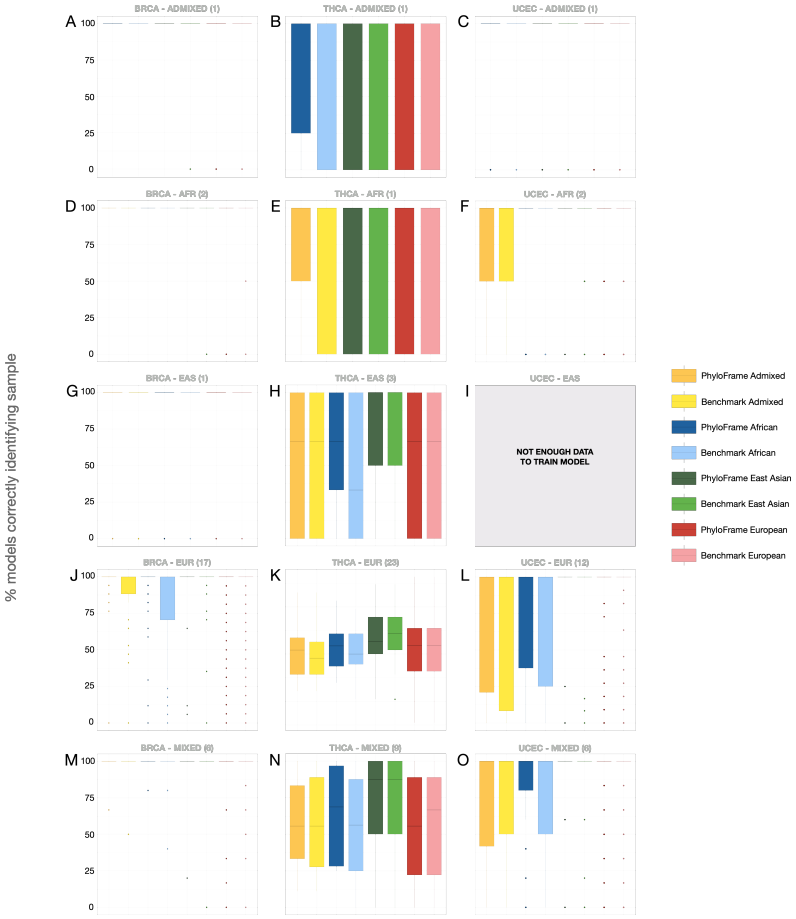
Supplementary Figure 2 Equitable AI effectiveness. AUC of the benchmark versus PhyloFrame models when training in (**A-E**) BRCA, (**F-J**) THCA, and (**K-O**) UCEC using different populations for the training and validation data and varying the ancestral population of the training data. Rows correspond to training data used (ADMIXED, AFR, EAS, EUR, MIXED). MIXED indicates that the training data ancestry diversity matches that of the TCGA data; it is not representative of the global population distributions. Each dot coesponds to a combination of training and test data, color coded by test data.



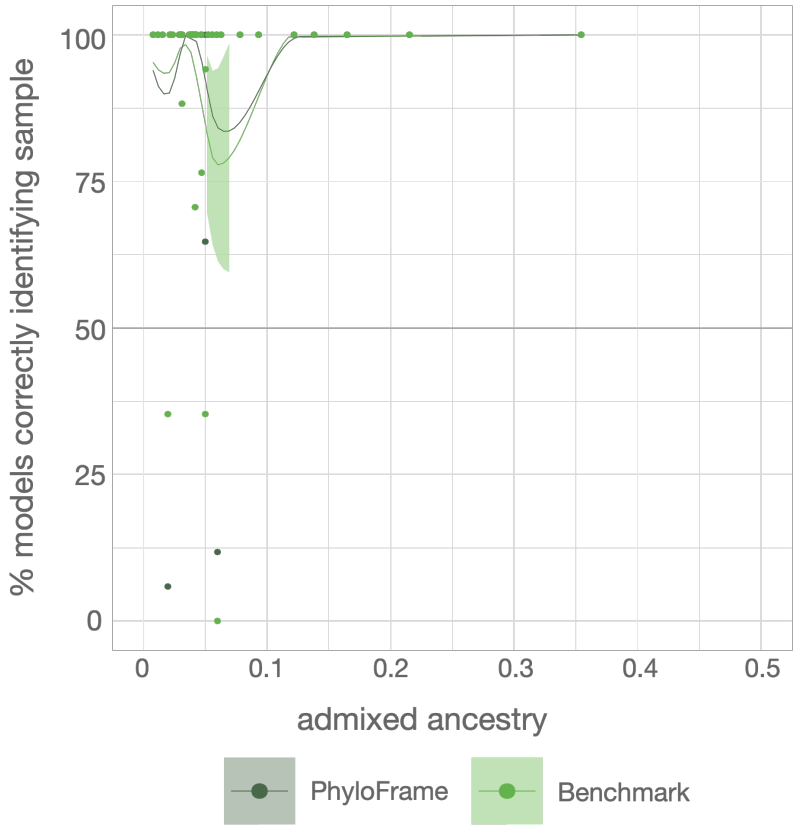
Supplementary Figure 3 COSMIC gene enrichment in model signatures An indicator plot showing (**A**) PhyloFrame and (**B**) benchmark model signatures and which COSMIC genes are included in each. Each column represents one trained model and each row is a COSMIC gene. Indicator marks are colored by the ancestry of the training data for the model.



Supplementary Figure 4 Transcriptome-wide EAF enrichment EAF density plots for (**A**) all genes and (**B**) COSMIC cancer genes, grouped by ancestry. **A** shows all 17 gnomAD ancestries and **B** shows the EUR, EAS, and AFR ancestries used to group the ancestry-specific training data sets for the AI models. Peaks demonstrate unique EAF across ancestries.



Supplementary Figure 5 Sample-specific model performance To ascertain which, if any, samples are harder to predict, we calculated performance of all models for each sample. Boxplots show the percent of models that correctly subtype each sample in (**A,D,G,J,M**) BRCA, (**B,E,H,K,N**) THCA, (**C,F,I,L,O**) UCEC. Samples are grouped by genetic ancestry and model type (PhyloFrame or benchmark). The EAS UCEC plot is greyed out as there are not enough samples to train the models. In each plot, each dot is a single sample and y-axis shows percent of models that correctly classify that sample.



Supplementary Figure 6 Effect of admixture on EUR-trained model performance. A comparison of models trained on EUR BRCA data and the percent of correctly predicted held-out EAS BRCA samples as admixture levels increase in PhyloFrame (dark green) and benchmark (light green) models.

TumorType	admix	AFR	AFR_Admix	AMR	EAS	EAS_Admix	EUR	EUR_Admix	SAS	SAS_A
ACC	0	0	2	0	2	0	81	1	0	0
BLCA	0	14	7	1	43	0	328	1	1	1
BRCA	1	125	56	5	56	1	822	3	4	4
CESC	0	19	12	1	21	1	194	15	0	0
CHOL	0	2	0	0	2	0	29	1	0	0
COAD	0	46	14	0	12	0	380	2	0	0
DLBC	0	1	0	0	15	1	30	0	1	0
ESCA	0	3	1	0	44	0	123	4	0	0
GBM	1	24	21	0	6	0	450	5	0	3
HNSC	1	38	12	6	6	1	437	4	3	2
KICH	0	3	1	0	1	0	56	0	0	1
KIRC	0	32	23	2	7	1	432	6	0	1
KIRP	0	43	20	0	6	0	208	1	1	0
LAML	0	16	0	0	2	0	178	0	0	0
LGG	0	15	8	4	10	0	449	7	1	2
LIHC	1	14	4	0	163	0	177	3	0	1
LUAD	0	42	18	1	9	0	502	0	0	0
LUSC	0	15	15	0	11	0	455	3	0	0
MESO	0	0	0	0	0	0	82	0	0	1
OV	0	23	11	1	15	1	504	1	2	4
PAAD	1	6	3	0	11	0	158	0	0	0
PCPG	0	8	12	0	3	0	146	0	4	0
PRAD	1	38	21	0	9	0	409	2	0	2
READ	0	1	5	0	1	0	155	1	0	0
SARC	2	12	6	0	6	0	220	1	0	0
SKCM	0	1	0	2	12	0	448	0	0	0
STAD	0	8	7	0	90	0	294	0	0	0
TGCT	0	1	3	0	4	0	121	1	0	0
THCA	3	21	12	11	53	0	364	1	9	2
THYM	0	4	4	0	12	0	92	4	0	0
UCEC	1	72	40	7	34	1	389	1	1	0
UCS	0	4	5	0	3	0	43	0	0	0
UVM	0	0	0	0	0	0	80	0	0	0

Supplementary Table 2 TCGA Cancer samples per ancestry. Summary of samples per ancestry across all TCGA can
BRCA has by far the most samples and is the most diverse cancer in the TCGA dataset. AMR, SAS, and admix individuals ar
severely underrepresented with only 41 AMR, 27 SAS, and 12 admix individuals across all cancers.

Acknowledgments. This work was supported by funding from (NIH/NCI) R01 CA259396-01 (awarded to K. Graim and J.A. Cahill). L.A. Smith is supported by a UF Dean's Fellowship Award and by the UF GenNext program. This project makes use of data from gnomAD, HumanBase, and TCGA.

Declarations

- Funding
  This work was supported by funding from (NIH/NCI) R01 CA259396-01 (awarded to K. Graim and J.A. Cahill).
- Competing interests
  Authors declare that they have no competing interests.
- Ethics approval
  Not applicable
- Consent to participate
  Not applicable
- Consent for publication
  Not applicable
- Availability of data and materials
  This project makes use of data from gnomAD, HumanBase, and TCGA.
- Code availability
  All code for the PhyloFrame equitable AI model and comparable benchmark model can be accessed at github.com/leslie-smith1112/phyloFrame.
- Authors' contributions
  KG, LAS and JAC designed the study conceptualized the method and planned experiments. LAS and KG developed software and conducted experiments. KG, LAS and JC analyzed and interpreted results. KG and JAC aquired funding and supervized the study. KG, LAS and JAC Worte the manuscript, visualized data and approved the final manuscript. Please address correspondence to Kiley Graim at kgraim@ufl.edu

If any of the sections are not relevant to your manuscript, please include the heading and write 'Not applicable' for that section.

Appendix A Extended Data

References

[1] Bentley, A. R., Callier, S. L. & Rotimi, C. N. Evaluating the promise of inclusion of african ancestry populations in genomics. *NPJ genomic medicine* **5**, 5 (2020).

[2] Pereira, L., Mutesa, L., Tindana, P. & Ramsay, M. African genetic diversity and adaptation inform a precision medicine agenda. *Nature Reviews Genetics* **22**, 284–306 (2021).

[3] Dovey, Z. S., Nair, S. S., Chakravarty, D. & Tewari, A. K. Racial disparity in prostate cancer in the african american population with actionable ideas and novel immunotherapies. *Cancer Reports* **4**, e1340 (2021).

[4] Popejoy, A. B. *et al.* The clinical imperative for inclusivity: race, ethnicity, and ancestry (rea) in genomics. *Human mutation* **39**, 1713–1720 (2018).

[5] Stopsack, K. H. *et al.* Differences in prostate cancer genomes by self-reported race: Contributions of genetic ancestry, modifiable cancer risk factors, and clinical factorsracial differences in prostate cancer genomes. *Clinical Cancer Research* **28**, 318–326 (2022).

[6] Petrovski, S. & Goldstein, D. B. Unequal representation of genetic variation across ancestry groups creates healthcare inequality in the application of precision medicine. *Genome biology* **17**, 1–3 (2016).

[7] Kessler, M. D. *et al.* Challenges and disparities in the application of personalized genomic medicine to populations with african ancestry. *Nature communications* **7**, 1–8 (2016).

[8] Moore, E., Allen, J. B., Mulligan, C. J. & Wayne, E. C. Ancestry of cells must be considered in bioengineering. *Nature Reviews Materials* **7**, 2–4 (2022).

[9] Yuan, J. *et al.* Integrated analysis of genetic ancestry and genomic alterations across cancers. *Cancer cell* **34**, 549–560 (2018).

[10] Hoadley, K. A. *et al.* Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell* **173**, 291–304 (2018).

[11] Ding, L. *et al.* Perspective on oncogenic processes at the end of the beginning of cancer genomics. *Cell* **173**, 305–320 (2018).

[12] Sanchez-Vega, F. *et al.* Oncogenic signaling pathways in the cancer genome atlas. *Cell* **173**, 321–337 (2018).

[13] of Us Research Program Investigators, A. The "all of us" research program. *New England Journal of Medicine* **381**, 668–676 (2019).

[14] Ma, X. *et al.* Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. *Nature* **555**, 371–376 (2018).

[15] Mardis, E. R. The translation of cancer genomics: time for a revolution in clinical cancer care. *Genome medicine* **6**, 1–6 (2014).

[16] Sollis, E. *et al.* The nhgri-ebi gwas catalog: knowledgebase and deposition resource. *Nucleic Acids Research* **51**, D977–D985 (2023).

[17] of Sciences Engineering, N. A., Medicine *et al. Using Population Descriptors in Genetics and Genomics Research: A New Framework for an Evolving Field* (The National Academies Press, 2023).

[18] Duda, P. & Zrzavỳ, J. Human population history revealed by a supertree approach. *Scientific Reports* **6**, 1–10 (2016).

[19] Betti, L., Balloux, F., Amos, W., Hanihara, T. & Manica, A. Distance from africa, not climate, explains within-population phenotypic diversity in humans. *Proceedings of the Royal Society B: Biological Sciences* **276**, 809–814 (2009).

[20] Williams, T. N. & Thein, S. L. Sickle cell anemia and its phenotypes. *Annual review of genomics and human genetics* **19**, 113–147 (2018).

[21] Swallow, D. M. Genetics of lactase persistence and lactose intolerance. *Annual review of genetics* **37**, 197–219 (2003).

[22] Yang, J. *et al.* Genetic signatures of high-altitude adaptation in tibetans. *Proceedings of the National Academy of Sciences* **114**, 4189–4194 (2017).

[23] Lewis, A. C. *et al.* Getting genetic ancestry right for science and society. *Science* **376**, 250–252 (2022).

[24] Ju, D., Hui, D., Hammond, D. A., Wonkam, A. & Tishkoff, S. A. Importance of including non-european populations in large human genetic studies to enhance precision medicine. *Annual Review of Biomedical Data Science* **5**, 321–339 (2022).

[25] Ding, Y. *et al.* Polygenic scoring accuracy varies across the genetic ancestry continuum. *Nature* 1–8 (2023).

[26] Conti, D. V. *et al.* Trans-ancestry genome-wide association meta-analysis of prostate cancer identifies new susceptibility loci and informs genetic risk prediction. *Nature genetics* **53**, 65–75 (2021).

[27] Fan, S. *et al.* Whole-genome sequencing reveals a complex african population demographic history and signatures of local adaptation. *Cell* **186**, 923–939 (2023).

[28] Bisogno, L. S. *et al.* Ancestry-dependent gene expression correlates with reprogramming to pluripotency and multiple dynamic biological processes. *Science*

advances **6**, eabc3851 (2020).

[29] Harwood, M. P. *et al.* Recombination affects allele-specific expression of deleterious variants in human populations. *Science Advances* **8**, eabl3819 (2022).

[30] Martini, R. *et al.* African ancestry–associated gene expression profiles in triple-negative breast cancer underlie altered tumor biology and clinical outcome in women of african descent. *Cancer Discovery* **12**, 2530–2551 (2022).

[31] 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. *The American Journal of Human Genetics* **104**, 1097–1115 (2019).

[32] Gay, N. R. *et al.* Impact of admixture and ancestry on eqtl analysis and gwas colocalization in gtex. *Genome biology* **21**, 1–20 (2020).

[33] Carrot-Zhang, J. *et al.* Comprehensive analysis of genetic ancestry and its molecular correlates in cancer. *Cancer Cell* **37**, 639–654 (2020).

[34] Dutil, J., Chen, Z., Monteiro, A. N., Teer, J. K. & Eschrich, S. A. An interactive resource to probe genetic diversity and estimated ancestry in cancer cell linesestimated cancer cell line ancestry. *Cancer research* **79**, 1263–1273 (2019).

[35] Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature genetics* **51**, 584–591 (2019).

[36] Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

[37] Greene, C. S. *et al.* Understanding multicellular function and disease with human tissue-specific networks. *Nature genetics* **47**, 569–576 (2015).

[38] O'Connell, J. *et al.* A population-specific reference panel for improved genotype imputation in african americans. *Communications Biology* **4**, 1269 (2021).

[39] Micheletti, S. J. *et al.* Genetic consequences of the transatlantic slave trade in the americas. *The American Journal of Human Genetics* **107**, 265–277 (2020).

[40] Eltis, D. The volume and structure of the transatlantic slave trade: A reassessment. *The William and Mary Quarterly* **58**, 17–46 (2001).

[41] Kim, M. S. *et al.* Testing the generalizability of ancestry-specific polygenic risk scores to predict prostate cancer in sub-saharan africa. *Genome biology* **23**, 1–16 (2022).

[42] Roelands, J. *et al.* Ancestry-associated transcriptomic profiles of breast cancer in patients of african, arab, and european ancestry. *NPJ breast cancer* **7**, 1–14

(2021).

[43] Arriaga-MacKenzie, I. S. *et al.* Summix: a method for detecting and adjusting for population structure in genetic summary data. *The American Journal of Human Genetics* **108**, 1270–1282 (2021).

[44] Zhao, D. *et al.* Exosomal mir-1304-3p promotes breast cancer progression in african americans by activating cancer-associated adipocytes. *Nature Communications* **13**, 1–15 (2022).

[45] Barral-Arca, R., Pardo-Seco, J., Bello, X., Martinon-Torres, F. & Salas, A. Ancestry patterns inferred from massive rna-seq data. *RNA* **25**, 857–868 (2019).

[46] Luo, Y. *et al.* Estimating heritability and its enrichment in tissue-specific gene sets in admixed populations. *Human molecular genetics* **30**, 1521–1534 (2021).

[47] Krainc, T. & Fuentes, A. Genetic ancestry in precision medicine is reshaping the race debate. *Proceedings of the National Academy of Sciences* **119**, e2203033119 (2022).

[48] Atkinson, E. G. *et al.* Cross-ancestry genomic research: time to close the gap. *Neuropsychopharmacology* **47**, 1737–1738 (2022).

[49] Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538**, 161–164 (2016).

[50] Sirugo, G., Williams, S. M. & Tishkoff, S. A. The missing diversity in human genetic studies. *Cell* **177**, 26–31 (2019).

[51] Liao, W.-W. *et al.* A draft human pangenome reference. *Nature* **617**, 312–324 (2023).

[52] Wang, T. *et al.* The human pangenome project: a global resource to map genomic diversity. *Nature* **604**, 437–446 (2022).

[53] Fatumo, S. *et al.* Promoting the genomic revolution in africa through the nigerian 100k genome project. *Nature Genetics* **54**, 531–536 (2022).

[54] Huo, D. *et al.* Comparison of breast cancer molecular features and survival by african and european ancestry in the cancer genome atlas. *JAMA oncology* **3**, 1654–1662 (2017).