TITLE: Improving GWAS performance in underrepresented groups by appropriate modeling of genetics, environment, and sociocultural factors

AUTHORS: Chelsea C. Cataldo-Ramirez[1,2], Meng Lin[3,4], Aislinn Mcmahon[1], Christopher R. Gignoux[3,4], Timothy D. Weaver[1], Brenna M. Henn[1,5]

AFFILIATIONS:

[1]Department of Anthropology, University of California Davis, Davis, CA, 95616, USA

[2]Department of Population and Public Health Sciences, Center for Genetic Epidemiology, Keck School of Medicine, University of Southern California, CA 91001, USA

[3]Department of Biomedical Informatics, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, USA

[4]Colorado Center for Personalized Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, USA

[5]UC Davis Genome Center, University of California Davis, Davis, CA, 95616, USA

ABSTRACT:

Genome-wide association studies (GWAS) and polygenic score (PGS) development are typically constrained by the data available in biobank repositories in which European cohorts are vastly overrepresented. Here, we increase the utility of non-European participant data within the UK Biobank (UKB) by characterizing the genetic affinities of UKB participants who self-identify as Bangladeshi, Indian, Pakistani, “White and Asian” (WA), and “Any Other Asian” (AOA), towards creating a more robust South Asian sample size for future genetic analyses. We assess the relationships between genetic structure and self-selected ethnic identities resulting in consistent patterns of clustering used to train a support vector machine (SVM). The SVM model was utilized to reassign n = 1,853 AOA and WA participants at the subcontinental level, and increase the sample size of the UKB South Asian group by 1,381 additional participants. We then leverage these samples to assess GWAS performance and PGS development. We further include environmental covariates in the height GWAS by implementing a rigorous covariate selection procedure, and compare the outputs of two GWAS models: $\text{GWAS}_{\text{null}}$ and GWAS_{env}. We show that PGS performance derived from environmentally adjusted GWAS yields comparable prediction to PGS models developed with an order of magnitude larger training dataset (R^2=0.021 vs 0.026). Models with 7 - 8 environmental covariates double the variance explained by PGS alone. In summary, we demonstrate how GWAS performance can be improved by leveraging ambiguous ethnicity codes, ancestry matched imputation panels, and including environmental covariates.

KEYWORDS: Height; Genome-wide association study (GWAS); Polygenic score (PGS); UK Biobank; South Asia; Environmental covariates

MAIN TEXT

Background:

Unequal representation of diverse populations within biobanks has resulted in comparatively few insights into the genetic architecture of complex traits for many non-European-descent populations [1]. This Eurocentric emphasis within genetic research is often attributed to the current state of data availability, in which European-descent genomic data is most abundant. Consequently, the prevailing assumption is that smaller datasets will fail to produce results. This perspective acts to disincentivize prospective researchers from attempting to make the best use of the datasets currently available [2]. However, utilizing smaller samples need not be a futile undertaking. Genotype-phenotype research design can be optimized to increase the utility of smaller datasets. Here, we highlight how a detailed consideration of the intersection of genetic ancestry affinities and ethnic identifiers, in conjunction with incorporating environmental adjustments into GWAS of height can improve the quality of results output by smaller, often excluded non-European-descent samples.

PGS performance declines when applied to groups outside of the discovery GWAS training sample [3,4,5], and because European cohorts are vastly overrepresented within the available data, PGS applicability is much reduced in non-Europeans [6, 7]. Specifically, variation in frequencies and population-specificity of causal alleles, in conjunction with differences in linkage disequilibrium (LD) between populations, contribute to the limitations of out-of-sample PGS portability [8, 9]. Fine-mapping with exome data, improved imputation for GWAS, and including LD-informed analyses can ameliorate the effects of these factors [4, 10]. Methods that integrate discovery-sample GWAS summary statistics with LD structure from multi-ethnic cohorts to improve the predictive utility of European-derived PGS have been a primary focus of PGS research recently [12,13,14]. However, improving GWAS discovery by increasing the sample sizes of minority cohorts or including relevant environmental variables have not been explored as systematically as other approaches.

Using UK Biobank (UKB) South Asian participants as a case study, we aim to test how the discovery of significant genetic variants and accuracy of PGS prediction of complex traits may be improved by: 1) considering several levels of participant identity and genetic characteristics when establishing a sample dataset; and 2) adjusting for environmental confounders in the GWAS model. We focus on height as a canonical complex trait. The UKB is a vast resource for public health, epidemiology, and precision medicine research, containing genetic, phenotypic, and environmental data for over 500,000 adult UK residents [14]. While the majority of participants self-identify as “white” (defined by Data Field 21000 as including British, Irish, and “Any other white background”; *n* > 500,000 as of January 2024), there are over 28,000 individuals who do not identify as such. The Pan-UK Biobank and additional studies have identified tens of thousands of individuals with primarily non-European genetically inferred ancestry [3, 16]. The largest geographically-defined grouping of non-“white”-identifying participants is a conglomerate of South Asian identities, which include Bangladeshi, Indian, and Pakistani identifying individuals (Data Field 21000). However, ethnic identity does not necessarily reflect high genetic affinities with other individuals or populations occupying the same geographical region at the same point in time [17,18,19,20].

The use of ethnicity as a proxy, or at least as a starting point, for identifying a genetically homogenous population is further complicated when the ethnic labels participants are able to choose from are limited, discrete/categorical, follow several naming conventions, and/or are un-defined, as is the case with UKB participants who identify as “White and Asian” (WA) or “Any other Asian” (AOA). This, in conjunction with the limited UKB Data Fields associated with genetic affinity available to researchers and historical apprehension to including genetically “admixed” populations in GWAS research has resulted in reduced inclusion for AOA and WA participants [5, 21, 22]. To rectify this, we characterize in detail the genetic affinities of UKB

AOA and WA participants, and provide clarification on their genetic similarities to 1000 Genomes (1KG) reference populations (Table S1.6) and UKB Bangladeshi, Indian, and Pakistani (together referred to as UKB-BIP) participants (Table 1). This, accordingly, contributes to our first aim of increasing participant inclusion by integrating several descriptors of participant identity when building an associated genetic dataset.

Table 1. UK Biobank sample.

Ethnicity*	**Group Acronym****	***n***
Bangladeshi	BIP	212
Indian	BIP	5016
Pakistani	BIP	1528
Any other Asian	AOA	1614
White and Asian	WA	549

*The “Ethnicity” column reflects the participants’ response to UKB data field 21000 (“Ethnic background”).

**The “Group Acronym” column reflects the original grouping of participants used in this study.

Previous investigations into the genetic ancestries of UKB participants have aimed to broadly categorize participants at the subcontinental level, but in-depth assessments of population affinities within the identified non-European groupings have been lacking. Privé et al. (2022) defined nine subcontinental ancestry groups within the UKB cohort and demonstrated that PGS prediction significantly decreases when PGS models trained on Northwestern European individuals are applied to the non-European ancestry groups [3]. However, non-European ancestry groups were, in part, restricted to participants whose self-reported ethnicity (as assessed by a within-UKB PCA) matched their country of origin, resulting in PGS exclusion for many participants [3: Fig. 1, Table SA1] as individuals within diaspora communities were not retained.

Constantinescu et al. aimed to characterize UKB participants of “non-white British ancestry” for the explicit purpose of identifying relatively homogenous groups (2022: 2), to facilitate the inclusion of these participants in future health research [16]. We implement a similar approach to Constantinescu et al. to characterize a subset of the UKB participants of “non-white British ancestry” at a subcontinental resolution as well as at the level of 1KG reference groups [16]. Our subcontinental approach, however, differs at three primary levels from the approach taken by Constantinescu et al.: 1) we focus our analyses on a UKB South Asian metapopulation as opposed to their inclusion of all “non-white” individuals, therefore providing more detail into South Asian population substructure; 2) we utilize a subset of the 1KG-UKB genetic data intersection in our genetic affinity analyses chosen specifically to reduce the effects of ascertainment bias resulting from the use of the UKB SNP array; and 3) while our subcontinental classifications are also PCA based, we choose to utilize support vector machines (SVM) instead of K-means clustering to infer population assignments.

After creating a broader UKB South Asian dataset, our second aim is to assess how adjusting for environmental confounders in the GWAS model affects GWAS and PGS results. We implement an environmental variable selection procedure to identify a subset of covariates to control for confounding in our GWAS model (referred to here as $GWAS_{env}$). We then assess the downstream effects of $GWAS_{env}$ by developing a height PGS model and compare its performance to published population-matched PGS.

RESULTS

Genetic affinities of UKB South Asian participants

The genetic affinities of UKB participants who self-identify as Bangladeshi, Indian, Pakistani, “White and Asian” (WA), and “Any Other Asian” (AOA) were characterized via three main methods: principal components analysis (PCA), ADMIXTURE, and support vector machines (SVM). For each method, the UKB data were projected onto or trained on data from a globally distributed reference panel (1KG) to characterize genetically how UKB-BIP participants align with other South Asian groups and to identify underlying genetic patterns of participants with “ambiguously defined” ethnicities (WA and AOA). For WA and AOA participants, ascertaining genetic similarities with broader groups (of potentially differing ethnic labels) may be the only way that these participants are routinely included in future genetics research, since discretizing populations into ethnic groupings of small sample sizes typically results in their exclusion from genetic analyses [5, 16, 20]. Therefore, in addition to the PCA and ADMIXTURE assessments, an SVM model was trained to quantitatively integrate information from the previously calculated PC scores and infer both population-level affinities and subcontinental assignments for each AOA and WA participant (see Methods section and SI Section 2.2).

Ascertainment bias in the UKB SNP array

To conduct these analyses, the QC-ed UKB SNP array (see Methods) was merged with the 1KG data (n = 668,051 shared SNPs). The 1KG data were used to calculate 20 PCs and the UKB data were projected onto this PC space (FRAPOSA [23]). PC plots were visually assessed to identify large-scale patterns within and between the UKB groups.

During the initial visual inspections, however, PCA results indicated that Utah residents with Northern and Western European ancestry (CEU) and British in England and Scotland (GBR) samples explained most of the genetic variance within the 1KG reference data (prior to projection of the UKB data) (Fig. 1), prompting further investigation of the data qualities. Several assessments were employed to identify the root cause(s) driving the inflated European variance observed in the PC plots, ultimately indicating ascertainment bias of the UKB genotype array as the primary cause (see Methods, SI Section 1.2.3, and Figures S1.3 – S1.5). Because the array was designed to highlight European genetic variation, most of the rare alleles on the array are found at higher frequencies within European groups (Fig. S1.2), artificially inflating the magnitude of European genetic heterogeneity relative to other globally distributed populations. When the UKB and 1KG data are intersected for population genetic analyses, globally representative data (1KG) gets reduced to the variants that, within PCA results, fail to adequately capture variation among samples with non-European genetic ancestry (i.e., the formerly globally representative sample becomes primarily representative of genetic variation within the British Isles). Therefore, for all subsequent analyses (described below), the original intersection of QC-ed SNPs within both UKB and 1KG data was further reduced to a subset chosen to reflect broad patterns of expected global genetic diversity (*n* = 199,495) following more stringent MAF thresholds and expulsion of variants that failed F_{ST} assessments (See Methods section and SI Section 2.2.3 for more details).

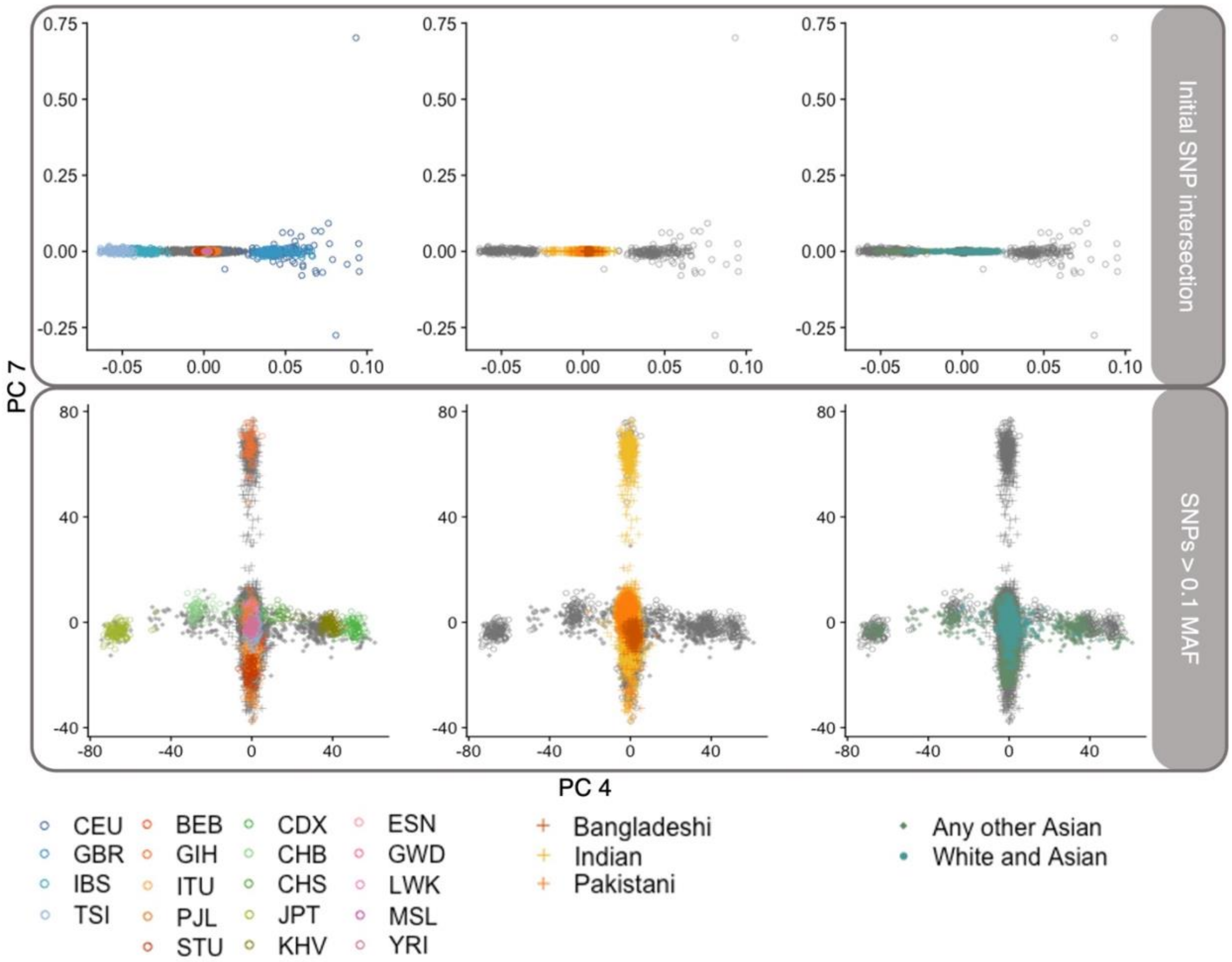

Figure 1. UKB BIP (middle column) and AOA and WA (right column) array data projected onto 1KG PC4 and PC7 space, when using standard QC thresholds for MAF (top row) and stringent MAF thresholds (bottom row). For each plot, 1KG data are represented by open circles, BIP are represented by plus signs, AOA are represented by filled diamonds, and WA are represented by filled circles. All data are plotted in the left column, with UKB data grayed. BIP and 1KG data are plotted in the middle column, with 1KG data grayed. AOA, WA, and 1KG data are plotted in the right column, with 1KG data grayed. Top: When using the full SNP intersection between UKB and 1KG data, PC4 and above are both driven by European genetic variation (See SI Section 1.2.3 for more information). Bottom: when more stringent MAF thresholds are used to filter out all but common SNPs (i.e., variants with MAF > 0.1 retained), PC4 is driven by East Asian genetic variation and PC7 is driven by South Asian genetic variation. Standard QC thresholds for filtering UKB array data will emphasize structure within European axes of variation and may only be appropriate for addressing within-sample structure if just the first three PCs are being considered. However, this may bias the interpretation of relative genetic variance within the sample data as well as patterns of global population structure (therefore also biasing attempts to characterize genetic ancestry).

UKB Bangladeshi, Indian, and Pakistani Principal Components Analysis (PCA)

UKB Bangladeshi, Indian, and Pakistani (BIP) participants generally cluster closer to one another and to 1KG South Asian reference groups than to other reference populations included in the global PCA. The UKB Indian group demonstrates the highest variance in scores across PCs, suggesting the highest levels of genetic diversity of the UKB BIP groups (Fig. 2; SI Section 1.3.1, Fig. S1.6-S1.8), although they also comprise the largest sub-sample, possibly confounding this interpretation. Across PCs, the Pakistani scores fall within the Indian ranges, making them largely indistinguishable from Indian-identifying participants. In 5 of the 20 calculated PCs, the Indian score distributions have more than one peak, suggesting population substructure that variably aligns with the Bangladeshi and Pakistani mean scores. Additionally, substantial population structure within the Indian group is visible on PC 7, which is driven by variation in South Asian reference populations. Along this PC, the Gujarati Indians sampled from Houston, TX (GIH) exhibit two peaks, one of which is far removed from other South Asian populations (this GIH substructure has previously been described by Sengupta et al. and Reich et al. [24, 25]). Clusters of UKB Indian participants can be found within both of the GIH peaks, as well as clinally distributed between them. Bangladeshi PC scores typically fall entirely within the Indian ranges, but with means closer to the East and Southeast Asian groups. This geographic trend (a West-East cline) is further reinforced where Pakistani and Bangladeshi mean scores diverge, as the Bangladeshi group tends to fall closer to the East and Southeast Asians while the Pakistani tend to fall closer to the Europeans (see Fig.S1.5 for examples).

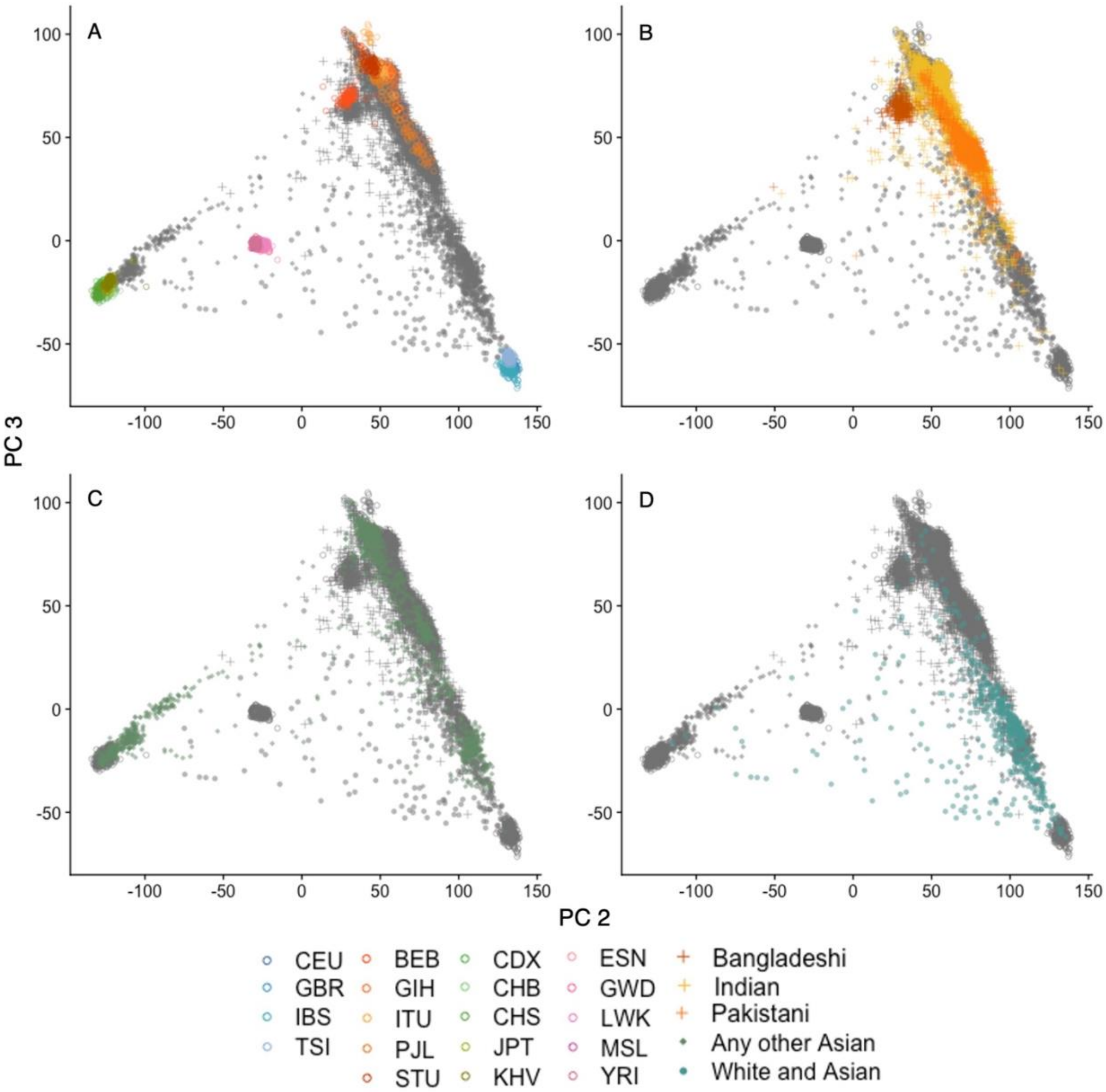

Figure 2. UKB self-selected ethnicity groups (UDI 21000) projected onto 1KG PCs 2 and 3. **A)** 1KG data used in the PCA, colored by reference population. **B - D)** UKB data projected onto 1KG PC space, with **B)** Bangladeshi, Indian, and Pakistani identifying participants colored; **C)** “Any other Asian” identifying participants colored; and **D)** “White and Asian” identifying participants colored.

UKB Bangladeshi, Indian, and Pakistani ADMIXTURE

Supervised ADMIXTURE analyses (informed by the unsupervised ADMIXTURE analyses of the 1KG references) were conducted for k = 3:9 and *pong* was utilized to visualize these results (Fig. 3; SI Section 1.3.3). Throughout lower and intermediate values of k (e.g., 4 - 6), the Indian and Pakistani groups are predominantly characterized by European and broadly South Asian components. At these *k*'s, Indian and Pakistani participants share similar proportions of these two clusters except for a small subset of Indian participants who have much higher South Asian affinities (Fig. S1.13, k = 4 - 6). However, at k = 7, the GIH reference group segregates as its own cluster, replicating the South Asian substructure demonstrated in the PCA. This GIH-associated component is found throughout BIP groups, though in the highest proportion in the subset of Indian participants previously characterized by their higher South Asian affinity (Fig. S1.13, k = 7 - 9). The Pakistani participants tend to fall within the ranges of cluster proportions that also characterize the Indian participants (generally, a combination of European, broad South Asian, and GIH-associated clusters). Throughout *k*'s, the Bangladeshi participants demonstrate higher affinities with the East Asian and Southeast Asian clusters than either of the other UKB-BIP groups, and generally align with the 1KG BEB reference group. Group average percentages for cluster affinities across *k*'s can be found in Table S1.10.

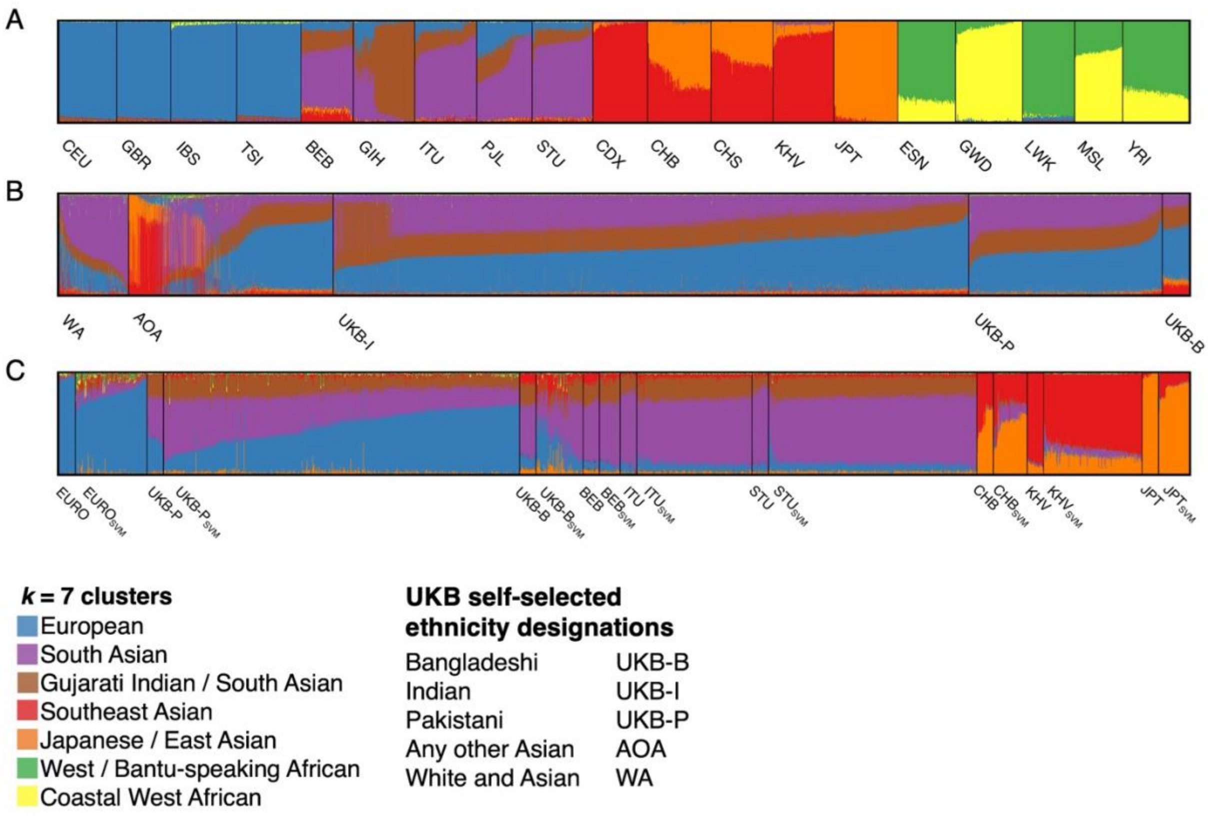

Figure 3. Supervised ADMIXTURE output trained on 1KG data, for k = 7. **A)** 1KG populations used in the unsupervised analyses. **B)** Supervised ADMIXTURE output grouped by the original UKB self-selected ethnicity labels (UDI 21000). **C)** Supervised ADMIXTURE output grouped by SVM training group (n = 30; left) and respective predicted group classifications (right) for UKB "White and Asian" and "Any other Asian" identifying participants. Only groups with 30 or more SVM classifications are included in this visualization. UKB Indian-identifying participants were not included in the SVM analyses and are therefore not included as a classification option here.

UKB "Any Other Asian"

Results of PCA, ADMIXTURE, and SVM classifications consistently demonstrate that UKB participants who self-identify as "Any Other Asian" (AOA) (n = 1,614) do not comprise a genetically homogeneous group (as the ethnicity label may imply). PCA results indicate that the AOA group mostly consists of participants who genetically align with Asian reference populations. Across PC's, score distributions highlight substructure within this grouping, especially along PC's driven by Asian genetic variance (e.g., PC's 4, 7, and 8; Fig. 1; SI Section

1.2.1). At least four clusters can be visualized in the PCA plots, consistently aligning with Japanese sampled in Tokyo (JPT), Chinese Dai sampled in Xishuangbanna (CHB), and the conglomerate of South Asian clusters, as well as a cluster of participants that tend to fall near or between the Kinh Vietnamese sampled in Ho Chi Minh City (KHV) and Southern Han Chinese (CHS) samples. However, not all participants can be characterized solely by Asian substructure. One participant consistently aligns with 1KG African reference groups, and European ancestry is also possible for a subset of the participants (Fig. 2). ADMIXTURE analyses further replicate these results, again indicating substantial genetic substructure within this ethnic label across all values of *k* (Fig. 3B) and demonstrating that a small portion of AOA participants are more genetically similar to European reference populations as opposed to Asian reference populations (Table S1.11).

SVM classifications reiterate the genetic structure within this ethnic label, inferring affinities with reference groups across the globe (classifications at ≥ 0, ≥ 0.75, and ≥ 0.9 probabilities are provided in Table S1.8). Using four subcontinental probabilities (with summed classification probabilities ≥ 0.7), we are able to classify 1,408 AOA participants. Specifically, 2.8% of the total AOA sample classify as European, 20.6% as East Asian, and 64.2% as South Asian (Table S1.9). These subcontinental classifications are derived from grouping the population-level assignments, for example the ‘East Asian’ classification is created by summing CDX, CHB, CHS, KHV, and JPT classification probabilities. The majority of AOA participants were classified as Bangladeshi or Pakistani, as UKB Indians were not incorporated into the training data; (see SI Section 1.3.2 for more information), and Indian Telugu (ITU) or Sri Lankan Tamil (STU) at lower classification probabilities. Of these participants, 159/208 that were classified as ITU and 309/373 classified as STU (classification probability > 0) report Sri Lanka as their region of origin. In total, 500 AOA participants report Sri Lanka as their region of origin, followed by Mauritius (*n* = 85), Afghanistan (*n* = 72), Kenya (*n* = 59), Iran (*n* = 33), and Malaysia (*n* = 31) (see Fig.S1.11 for further details). Notably, no participants were classified as Punjabi from Pakistan (PJL) at any classification probability threshold. Of the 1,613 AOA participants included in these analyses, 1,036 align at the subcontinental level with South Asian samples.

UKB "White and Asian"

While AOA results indicate the presence of multiple structured populations within a single ethnic label, "White and Asian" (WA) analyses suggest a clinal genetic pattern between European and Asian populations (Fig. 2D). ADMIXTURE analyses also indicate high levels of European affinity throughout values of *k* (Fig. 3; Fig.S1.14; Table S1.11). SVM classifications predominantly place WA participants as European or UKB Pakistani, and to a lesser extent UKB Bangladeshi (Table S1.8; Fig. 4). Of participants with an SVM classification of UKB Pakistani (classification probability > 0), 54/93 report India as their region of origin and 11/93 report Pakistan as their region of origin. India and Pakistan reflect the most common region of origin for WA participants, followed by Myanmar (*n* = 7), The Guianas (*n* = 5), Germany (*n* = 4), and Sri Lanka (*n* = 4) (see Fig.S1.10 for further details). Probabilities summed at the subcontinental level classified 16.8% of WA participants as European, 1.5% as East Asian, and 62.8% as South Asian (Table S1.9). Of the 549 WA participants included in these analyses, 345 align at the subcontinental level with South Asian samples.

Cumulatively, the dispersed PCA scores, high European affinity demonstrated by ADMIXTURE, and low population-level SVM classification probabilities suggest that WA participants either have high levels of both European and Asian ancestry (aligning with an "admixed" interpretation of the label "White and Asian") or that these participants share genetic affinities with populations not represented in the reference data (e.g., Central and/or West Asian groups). Future analyses of haplotype data or the inclusion of more representative reference data will be necessary to disentangle these two hypotheses.

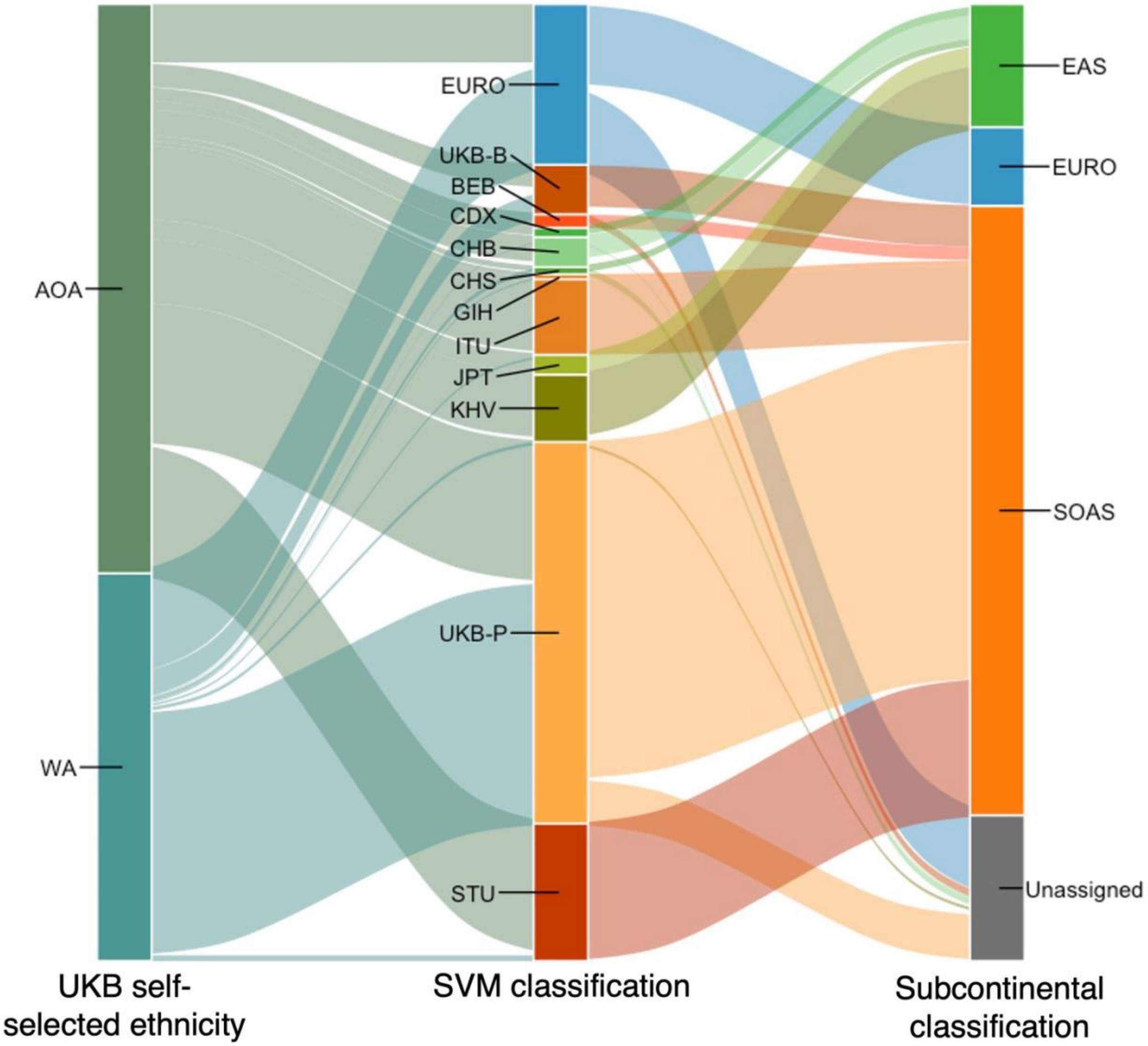

Figure 4. Relationships between group labels and genetic affinities throughout the SVM pipeline, for UKB participants who selected “Any other Asian” (AOA) or “White and Asian” (WA) as their ethnicity. Participants with self-selected AOA or WA ethnic backgrounds (UKB data-field 21000) (left column) were first assigned a group membership to one of the 13 reference populations used to train the SVM model (AFR not pictured), based on the highest classification probability assigned (middle column). Classification probabilities for each participant were grouped by subcontinental region and summed. Those with a summed subcontinental classification probability ≥ 0.7 were assigned to that group (right column). AOA and WA participants assigned to the South Asian (SOAS) group were added to the UKB-BIP sample to create a new UKB South Asian sample.

Residual Population Stratification

While comparative analyses collectively suggest that UKB-BIP participants are genetically most similar to non-UK South Asian groups (i.e., the 1KG South Asian reference populations), ADMIXTURE results, specifically, indicate slightly higher European affinities within all UKB groups assessed than the 1KG South Asian samples (Fig. 3) and there are visible (and quantifiable) discrepancies between UKB WA and AOA ADMIXTURE proportions compared to the 1KG SVM associated training group (Fig. 3; Fig. S1.12 – S1.14; Tables S1.10 and S1.11). It’s unclear if this is an artifact of the genotype array or a true signal of higher shared genetic ancestry between UKB participants and other Europeans.

Disclaimer regarding genetic affinity analyses: It should be emphasized that human genetic diversity is not categorical, and the analyses utilized here leverage the incredibly small amount of genetic diversity that differs among populations [25, 26]. The co-opting of statistical methods designed to classify and discriminate between discrete data towards analyzing human genetic variation is employed here only to facilitate the incorporation of typically excluded participants in genetic studies. The terms used in the Methods and Results sections (e.g., “classification”) reflect the statistical procedures conducted, and are not meant to impose identities or group affiliations outside of this context.

Environmentally-adjusted GWAS and PGS

Environmental covariates

Bone mass and joint integrity (and consequently, height) peak at around 30 years of age [28, 29] before gradually breaking down over the next several decades of life. Non-genetic factors that influence an individual’s peak bone mass include primarily the environment experienced during ontogeny. To account for these effects, variables encompassing diet, general health, activity patterns, exposures, and socioeconomic status were requested from the data available in the UKB. After growth has ceased, systemic health and maintenance become the prominent factors concerning skeletal mass and morphology. Variables that affect bone mineralization, joint function, overall health, and the likelihood of incurring fractures were included to account for post-ontogenetic environmental influences.

Following assessments of the quality, missingness, and correlations among the requested environmental variables (Methods), nine data fields were selected for inclusion in the GWAS (Table 2; Figure S2.2). To test if these specific environmental variables are significantly associated with height, and to identify the variance in height explained by additive environmental components alone, an ordinary least squares (OLS) regression model was tested. The full UKB South Asian sample with un-imputed data (*n* = 7,331) was used for this assessment.

The environment-only model explains 6-8% of the variance in height in males and females, respectively (Table 3; Table S2.2). When applied to the combined sex sample, all variables meet significance ($p < 0.01$) except for ‘Number of live births’ (Table S2.1), though visualization of the effect does suggest a weak, negative association with height (Figures S2.4 - S2.5). Variables with the strongest effects on height include sex, ‘Birth location’ (whether or not the participant was born in the UK), and ‘Excludes dairy’ (whether or not the participant includes dairy in their diet). Dietary variables reflecting meat consumption, and ‘Health’ (a self-assessed overall health indicator) are associated with moderate effects on height, while ‘TDI’ and ‘YOB’ confer the weakest effects of all significant covariates included in the model.

Table 3. Summary statistics for environmental-only models.

Model covariates	**Sample**	**Adj. *R*^{2}**	***p***
Sex + TDI + YOB + Birth location + Eats pork + Eats beef + Excludes dairy + Health + Number of births	All (*n* = 7,331)	0.57	$< 2.2 \times 10^{-16}$
TDI + YOB + Birth location + Eats pork + Eats beef + Excludes dairy + Health + Number of births	F (*n* = 3,430)	0.08	$< 2.2 \times 10^{-16}$
TDI + YOB + Birth location + Eats pork + Eats beef + Excludes dairy + Health	M (*n* = 3,901)	0.06	$< 2.2 \times 10^{-16}$

Height GWAS models

To assess how environmental covariate inclusion may affect GWAS results, two GWAS models were run: GWAS$_{\text{null}}$, which includes typical covariates used in previously published height GWAS (e.g., age, sex, and genetic PCs); and GWAS$_{\text{env}}$, which incorporates environmental covariates hypothesized or previously demonstrated to influence growth and/or systemic health, in addition to the covariates in the “null” model. These covariates can be found in Table 2. The GWAS were run on a subset of the full sample (*n* = 6,954) as 1000 randomly selected participants were “left-out” for PGS assessments and some participants did not have standing height data. SNP significance and effect sizes output by both GWAS models were compared.

Table 2. Environmental variables included as covariates in GWAS$_{null}$ (italicized) and/or GWAS$_{env}$ (all variables listed here).

Variable (UKB data ID)	**Category**	**Missingness / % imputed**	**Description**
Sex (31)	*Demographic*	*0*	*Sex*
YOB (34)	*Demographic*	*0*	*Year of birth*
TDI (189)	Demographic	0.14	Townsend deprivation index at recruitment
Birth location (1647)	Demographic	2.63	Country of birth (binarized; UK/ Elsewhere)
Eats beef (1369)	Diet	2.53	Beef intake (binarized; eats beef/ doesn’t eat beef)
Eats pork (1389)	Diet	2.31	Pork intake (binarized; eats pork/ doesn’t eat pork)
Excludes dairy (6144)	Diet	2.13	Never eat eggs, dairy, wheat, sugar (binarized; excludes dairy/ doesn't exclude dairy)
Health (2178)	Health	1.72	Overall health rating (binarized; poor/ fair-excellent)
Number of births (2734)	Life history	14.23*	Number of live births

Variable descriptions are modified from the UKB data showcase listings.

* ‘Number of births’ was not imputed. Any missing values associated with female participants were recoded as “0”, as were all responses associated with male participants. The percent listed above reflects the number of “0” responses after recoding, within female-only data (*n* = 533/3745).

Variant significance and effect size comparisons

Comparisons of SNP significance between GWAS$_{null}$ and GWAS$_{env}$ outputs demonstrate increased detection of tag-SNPs in GWAS$_{env}$ (Table 4; Figure 5). This increase in significantly identified SNPs did not occur in a uniform manner, however, as the incorporation of environmental covariates decreased the significance of some associations relative to GWAS$_{null}$ while increasing others.

Table 4. Number of SNPs meeting different significance thresholds between GWAS models.

p <	10^{-5}	10^{-6}	10^{-7}	5^{-8}
$GWAS_{null}$	530	248	11	8
$GWAS_{env}$	558	294	17	11

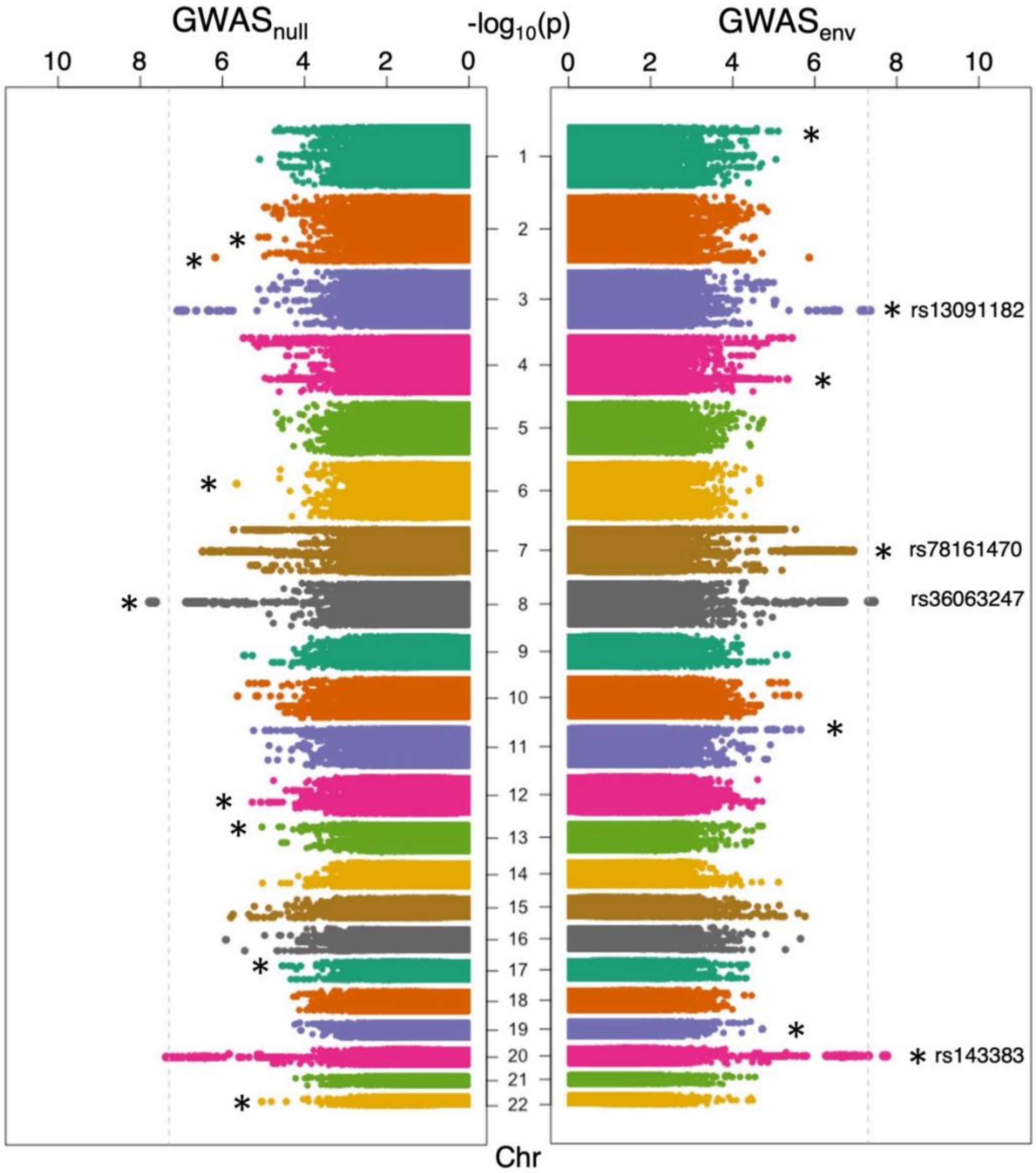

Figure 5. Manhattan plots of significantly associated SNPs resulting from $GWAS_{null}$ (left panel) and $GWAS_{env}$ (right panel). A significance threshold of $p < 5^{-8}$ is designated by the dashed lines. The rsID for the variants with the lowest p-value per peak is listed in the $GWAS_{env}$ panel. Regions with inflated

significance in $\mathrm{GWAS_{null}}$ that are reduced in $\mathrm{GWAS_{env}}$ are designated by asterisks in the $\mathrm{GWAS_{null}}$ panel. Alternatively, regions with increased significance in $\mathrm{GWAS_{env}}$ in comparison to $\mathrm{GWAS_{null}}$ are designated by asterisks in the $\mathrm{GWAS_{env}}$ panel.

Changes in SNP effect size between GWAS outputs were also investigated, demonstrating both reductions and increases in the effects of significant variants. Comparisons between SNP effects output by $\mathrm{GWAS_{env}}$ (β_{env}) and $\mathrm{GWAS_{null}}$ (β_{null}) were restricted to variants with $p < 10^{-6}$ output by $\mathrm{GWAS_{env}}$ (n = 294), as these reasonably represent the most reliable results of the variants assessed. As such, no sign flips (i.e., changes in effect directionality) between β_{null} and β_{env} were detected for SNPs meeting this significance threshold. However, of these 294 variants, 131 associated effect sizes were strengthened (i.e., 2 positive values increased, and 129 negative values decreased), 163 effects were attenuated (i.e., 98 positive values were reduced, and 65 negative values were increased), and all β standard errors decreased in $\mathrm{GWAS_{env}}$ relative to $\mathrm{GWAS_{null}}$ (Table S2.4). The distribution of effect sizes differs as well, primarily with regards to SNPs with large, positive effects (Fig. S2.9). Within chromosomes, the largest shifts in effect size distributions can be found in chromosomes 7 and 8 (Fig. S2.9). Interestingly, none of the chromosome 7 SNPs meeting the significance threshold for these assessments are published in the GWAS catalog for associations with height.

Polygenic score (PGS) construction and assessments

Following the GWAS, we developed a height PGS model using $\mathrm{GWAS_{env}}$ summary statistics and LDpred2-auto [30, 31]. Only $\mathrm{GWAS_{env}}$ summary statistics were used to develop a PGS model, as the aforementioned GWAS comparisons suggest $\mathrm{GWAS_{env}}$ to be more reflective of "true" genetic effects than $\mathrm{GWAS_{null}}$ (e.g., $\mathrm{GWAS_{env}}$ resulted in a larger number of significant variants and produced smaller β standard errors per variant). We tested the performance of $GWAS_{env}$ by calculating the adjusted R^2 between PGS and height for UKB South Asian participants held out from the sample prior to GWAS analyses (n = 996). With $GWAS_{env}$ PGS scores alone (Table 5), 3.8% - 5.6% of the variance in height can be explained in males and females, respectively. When the sexes are combined, performance drops to 2.1%. When we explicitly incorporate environmental covariates (Table S2.11), performance for the full sample increases substantially (adjusted R^2 = 0.582), though this is predominantly due to including sex, since height is a sexually dimorphic phenotype. In sex-stratified samples, adjusted R^2 = 0.091 (males) and = 0.104 (females), indicating that about half of the variance can be attributed to the PGS and half to the 7-8 environmental covariates.

Table 5. Predictive performance of PGS scores. Extended results can be found in Table S2.11.

PGS	**Model** **(R syntax)**	**Model description**	**Sample**	**Adj. R^2**	**R^2**	**R^2 95% CI**
PGS$_{Y1}$: **PGS002800 – GWAS SNP intersection** **(*n* = 898,486)**	Height ~ Score	Height explained by only PGS scores (no other adjustments or covariates)	Full test sample (*n* = 997)	0.026	0.027	0.0075 - 0.0473
			F (*n* = 473)	0.097	0.099	0.0481 - 0.1495
			M (*n* = 524)	0.041	0.042	0.0090 - 0.0763
GWAS$_{env}$ PGS **(*n* = 896,160)**	Height ~ Score	Height explained by only PGS scores (no other adjustments or covariates)	Full test sample (*n* = 997)	0.021	0.022	0.0042 - 0.0404
			F (*n* = 473)	0.056	0.058	0.0173 - 0.0983
			M (*n* = 524)	0.038	0.040	0.0074 - 0.0730

Comparisons with Yengo et al., 2022

We further tested the utility of our PGS equation by assessing its performance relative to a published South Asian PGS [32; PGS002800 in the PGS Catalog], which was developed from a GWAS meta-analysis (n = 77,890) and tested on UKB South Asian participants (n = 9,257). To our knowledge, this PGS equation was developed on the largest ancestry-matched dataset to our sample, and therefore represents the upper limit of predictive performance currently achievable. The PGS002800 score [33] was applied to our test sample using ESCALATOR, a PGS harmonization pipeline [34]. Although the specific participants used in the Yengo et al. [32] PGS test sample are unknown, it is likely that a large proportion are also included in our GWAS and PGS test sample, making our data appropriate for direct comparisons. Performance accuracy (R^2) was compared across several combinations of SNP sets, samples (e.g., sex-specific subsets), and covariate inclusion to account for differing variants included in the PGS as well as environmental variables incorporated in the training GWAS (Tables 5 and S2.11). Additionally, we assessed if PGS002800 performance was due to a larger GWAS sample size or to unadjusted environmental covariates that might inflate effect sizes.

The PGS002800 PGS_Y scores yield comparable results to our PGS_env, with adjusted R^2 values ranging from 0.023 to 0.026 when applied to the full test sample (Tables 5 and S2.11), though this falls short of the reported range of 0.033 to 0.045 [32]. Within sex-stratified subsets, male adjusted R^2 values range from 0.039 to 0.041 while female adjusted R^2 values range from 0.097 to 0.102. When environmental covariates are included in the model (Table S2.11), adjusted R^2 values increase, explaining ~14% of the variance in height in females and ~6.5% in males. To differentiate between factors impacting PGS performance (e.g., if accuracy is increased due to collinearity of genetic effects and un-adjusted environmental influences), each PGS score was then regressed on the GWAS_env environmental covariates (Table S2.12). Across models, adjusted R^2 is predictably low. However, values are higher for all PGS_Y-derived scores than those resulting from the GWAS_env PGS developed here. Comparatively, a larger R^2

(between PGS score and environmental covariates) suggests a higher likelihood that environmental influences are captured by the β weights used to calculate the PGS score (as opposed to directly causal genetic effects). Therefore, PGS_Y scores likely reflect a higher proportion of genetic effects that covary with environmental stratification relative to the $GWAS_{env}$ PGS score.

DISCUSSION

Pre-existing, predominantly European biobank data can be used for genotype-phenotype research, but extra steps must be taken to improve the quality of the genetic data for non-European participants. Currently, the UK Biobank only provides genetic ancestry information for 'White British' participants (data field 22006), leaving the characterization of genetic affinities among non-'White'-identifying participants to researchers. This can result in discrepancies across studies with regards to participant inclusion, and therefore, limit the utility of out-of-sample results and inferences. By investigating the intersection of sociocultural labels and genotype data, we are able to increase the UKB South Asian sample (originally only comprised of Bangladeshi, Indian, and Pakistani participants) by 20%, facilitating the development of larger UKB Asian analytic groups. Increasing sample sizes, including the diversification of biobank demographics, is only a partial solution. We demonstrate that including relevant environmental covariates improves height GWAS portability, and that by taking a quality-centered approach to research design, we are able to generate a PGS model that performs comparably to those trained on datasets incorporating ten times the number of participants.

Genetic affinities of UKB South Asian participants

While the UKB ethnic categories are very limited, there is substantial genetic diversity among participants within the same label, demonstrating the limited correspondence between sociocultural labels and genetic affinities. Currently, geopolitically defined ethnic labels within the UKB (e.g., BIP) have more consistent patterns of genetic affinities than more ambiguous labels such as AOA or WA. Interesting distinctions between sociocultural and genetic affinities arise when considering the intersection of the selected ethnic identities of participants with their region of origin and genetic data. While broad trends are apparent and distinguish the genetic stratification between WA and AOA groups at large, there are no definitive indicators demonstrated by the genetic data that explain participant identities or cultural affiliations at the individual level. Individual participants who report the same region of origin and are assigned to the same SVM classification choose to identify as AOA versus WA (of the available options) for reasons not predictable by genetic analyses. For example, SVM results show that a higher proportion of WA than AOA participants classify as European or UKB Pakistani, however, within SVM classifications the same regions of origin are reflected regardless of participant ethnic identity (e.g., most of the WA participants classified as UKB Bangladeshi are from India, Myanmar, The Guianas, Malaysia, and Mauritius, and most of the AOA participants classified as UKB Bangladeshi are similarly from Mauritius, Myanmar, Nepal, Sri Lanka, and Malaysia [Figures S1.10 and S1.11]). This highlights, and reiterates, that ethnic identity, genetic stratification, and shared ancestry are distinct participant descriptors that cannot be used *a priori* as proxies for one another. Therefore, ethnic groupings may represent an adequate starting point for preliminary sample selection but cannot be used, unaltered, as a homogenous or distinct genetic sample.

The population genetic analyses presented here additionally replicate previously reported trends regarding South Asian genetic affinities, particularly with regards to higher genetic similarity of UKB Pakistani participants to Europeans (Fig.3; Fig.S1.8). This is likely a result of shared ancestry derived from the prehistoric ‘Ancestral North Indian’ (ANI) population [25, 35] rather than more recent gene flow events, though explicitly distinguishing between these hypotheses was not tested here. More recent population movements are reflected when comparing SVM classification results with participants’ region of origin (Figures S1.9 - S1.11), exemplifying links with common diaspora communities. For example, East African countries (Uganda, Kenya, Tanzania) are predominantly comprised of participants classified via SVM as UKB Pakistani or Indian Telugu from the UK, reflecting the 19th century enforced migrations of indentured servants from India, which heavily recruited from Punjab [36]. Regions of origin in the Caribbean (e.g., The Guianas) exhibit a higher proportion of participants classified as Bengali from Bangladesh, UKB Bangladeshi, and Gujarati Indians from Houston, similarly reflecting 19th century enforced migration patterns of indentured servitude that derived individuals mainly from the Northeast states of India [36].

To our knowledge, we are the first group to purposefully counteract the effects of genotype array ascertainment bias in the UKB data, therefore genetically inferring a sample that is likely to be better matched with (i.e. more genetically similar to) South Asians outside of the UKB cohort. If the aim of genotype-phenotype research is to derive meaningful inferences about the genetic etiology of health outcomes towards implementing broadly applicable PGS equations in healthcare contexts, then this bias must be explicitly accounted for in order to mitigate the limitations of PGS portability. Typical MAF filtering thresholds of 0.01 – 0.05 (meant to reduce the effects of genotyping error) are not adequate for quality-control processing of this dataset, when the genotype data are being used to identify genetic patterns of non-Europeans and assess relationships with non-UKB reference data.

Environmentally-adjusted GWAS and PGS development

Genetic associations with height are typically assessed via a linear regression model that is purely additive, ignoring epistasis, as well as gene-environment interactions (GxE).

Currently, model simplification is deemed justifiable due to limitations, such as the low power of GxE tests, measurement error of the environmental variable of interest, and a lack of knowledge of the specific biological pathways underlying the interactive effects [37]. Once identified, GxE can be integrated into GWAS modeling simply by entering the interaction term; however, because of the inherently non-experimental nature of human GxE discovery research it has been difficult to disentangle confounding from causality [38]. Regardless, for complex traits that have been studied, environmental influences are relevant, significant, and often ignored [21, 39].

There are many environmental exposures that have been repeatedly demonstrated to affect human stature [40,41,42,43,44,45,46,47,48,49,50]. Yet most height GWAS only incorporate sex, age, and a genetic relatedness matrix (GRM) and/or simply the first 10 PCs to account for population stratification and adjust for confounders (e.g.: Jeon et al., 2024; Sohail et al., 2023; Wojcik et al., 2019; Yengo et al., 2022; Zoledziewska et al., 2015 [32, 51, 52, 53, 54]). In fact, a primary reason for the accessibility of ample height data in medical records is *because of the relationship between height, environmental factors, and health* including “cumulative net nutrition, biological deprivation, and standard of living between and within populations” [48: 1]. The bulk of population-level differences may be captured by within-sample PCs and GRMs, but it’s unlikely that they capture fine-level environmental effects, and it has been demonstrated that controlling for population stratification in this manner isn’t always sufficient for highly polygenic traits [55].

Environmental covariates

If all environmental variables present in an available data repository are incorporated as covariates in the GWAS, false-positive and false-negative results may arise due to shared genetic causality among some subset of candidate environmental covariates and the phenotype of interest [56]. For example, there may be non-genetic data that can be incorporated into a

GWAS or PGS equation which increases prediction but detracts from our ability to make inferences regarding the specific influences of the explanatory variables.

The environmental-only model presented here demonstrates that each of the environmental variables alone are associated with effects on the order of centimeters, whereas the largest singular allelic effect output by $\mathrm{GWAS}_{\mathrm{null}}$ or $\mathrm{GWAS}_{\mathrm{env}}$ is 1.6mm (for $p < 10^{-6}$). The effects conferred by non-genetic factors may seem small in comparison to absolute adult height, however, when taking into account the highly polygenic architecture of this phenotype, any singular environmental covariate outweighs the effects conferred by any particular genetic locus. Therefore, the selection of environmental covariates included in the GWAS should be considered a non-trivial process.

Lastly, we note that the UKB only provides data on adult participants, reflecting the participants' recent behavior and what can be recalled from childhood. It is assumed that in most cases the participant responses regarding current diet and socioeconomic status do not deviate significantly from childhood patterns, though this cannot be explicitly tested within the data. There is some research to indicate that this is a reasonable assumption, though it is a limitation of the study [57, 58].

GWAS and PGS comparisons

Increased significant variant detection, in conjunction with reduced genomic inflation (smaller values of λ) and β standard errors resulting from our $\mathrm{GWAS}_{\mathrm{env}}$, demonstrate that including select environmental covariates in the model improves GWAS performance. Taken together, this suggests that: 1) adjusting for environmental influences reduces some of the noise that may be mitigating genetic effects; 2) the significance of some variants may be falsely inflated when environmental confounders aren't adjusted for, leading to misidentification of causal genetic loci (e.g., the genetic variant is not causal but covaries with environmental variation); or 3) by incorporating environmental adjustments, SNP β values are more stable (i.e., have lower uncertainty) and likely reflect more accurate estimates of “true” effects. Furthermore, SNPs with large deviations in significance between the GWAS here may represent interactions with one or more of the environmental variables included in $GWAS_{env}$. Additional investigation into these variants may generate candidate SNPs for future GxE research.

The quality of our $GWAS_{env}$ results is reinforced when considering the performance of its derived PGS model relative to that of Yengo et al., 2022, which was trained on a meta-analysis comprising the largest published South Asian sample. Our model works comparably to theirs, though it was trained on a sample size that was 90% smaller, suggesting that increasing sample size is not the full solution to improving PGS performance. Across assessments, the Yengo-derived PGS scores do explain slightly more of the variance in height. However, there is also a larger discrepancy in PGS_{Y} performance between males and females, which is reduced in $GWAS_{env}$ PGS results. Interestingly, PGS_{Y1} performed better than PGS_{Y2} (Table S2.11), although the latter incorporates a larger proportion of possible variants in the PGS002800 scoring file. It’s possible that this is simply due to the quality of the imputed data, as PGS_{Y1} only included the intersection of SNPs meeting our GWAS inclusion info score threshold of 0.8. However, it also could be the case that the additional ~100k SNPs in PGS_{Y2} don’t add much to model performance, which could be due to the multitude of factors generally implicated in PGS portability.

CONCLUSIONS

Even though ancestry-diverse biobank cohorts remain small, there are steps we can take to maximize the utility of these samples within the context of genotype-phenotype research. In addition to commonly applied adjustments, e.g., incorporating LD-informed methods, some fundamental aspects of research design can be implemented to improve the quality of the results. We argue that the improving precision in trait measurement and ancestry assignment, as well as including environmental covariates facilitate the performance of PGS. If the aim of PGS implementation is not just to predict a trait expression or health outcome, but to understand *why* the outcome is predicted, then covariate effects need to be interpretable for individuals across populations. As such, if biobank-level data is accessible, it should be standard practice for GWASs of height to incorporate, at a minimum, covariates that may adjust for within-population variance in height explained by nutritional or socioeconomic variation amongst individuals. Such exploration remains a challenging, but likely fruitful area for new research.

ACKNOWLEDGEMENTS

Above all others, we would like to thank the participants of the UKB for providing their personal health data towards furthering health research. Additionally, we would like to thank Graham Coop and Mark N. Grote for their guidance, critiques, and comments throughout the development of this project (and resulting manuscript).

FUNDING

B.M.H and C.C.R. were supported by the National Institutes of Health grant R35GM133531, and M.L. and C.R.G. were supported by the National Institutes of Health grant R01HL151152. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The funders had no role in the decision to publish or prepare the manuscript.

CODE REPOSITORIES / DATA AVAILABILITY

The results presented here will be returned to the UKB, providing comparable data to data field 22006 for South Asian participants, and will be available to approved researchers. We recommend replicating this pipeline on participants enrolled subsequent to our research agreement to continue building this sample. By characterizing, in depth, the genetic affinities of participants comprising the largest non-‘White British’-identifying sample, we hope that future researchers will be able to more easily incorporate these participants in their studies.

The scripts developed for the analyses discussed here will be housed in an open access repository at https://github.com/ccatram/AnthroGen. The projected PC scores and SVM classification probabilities will be added to the UKB database and made available to researchers approved by the UKB review panel.

MATERIALS AND METHODS

Sample datasets

All data used in this study were either obtained from the UK Biobank via project ID 54084 (downloaded in November, 2020) or, with regards to the 1KG Phase 3 data (build GRCh37), downloaded from the IGSR FTP site in April 2020 [59, 60]. All 1000 Genomes reference groups were included except for American samples. UKB participants were selected based on responses to UKB data field 21000 and were included in our starting sample if any of the following ethnic identities were chosen: Bangladeshi, Indian, Pakistani, Any other Asian, or White and Asian (*n* = 10,288 at the time of our data application [4,832 females, 5,456 males]; Table 1). Information regarding the UKB genotype array (Affymetrix UK Biobank Axiom® array) can be found in the UKB data showcase website (https://biobank.ndph.ox.ac.uk/showcase/), specifically “Resource 1955: SNP Quality Control information” and “Resource 807: Sample processing and preparation of DNA for genotyping”.

The UKB consists of over 7,000 available data fields, 283 of which were requested for this study (Figure S2.2 A). These 283 data fields were selected based on their known or hypothesized effects on skeletal growth and systemic health, or if the data field reflected contextual information surrounding data collection so that any potential biases introduced while obtaining the data could be investigated (Figure S2.2 B).

The data utilized in this study fall under the ethics approvals granted to the UKB. The UKB received approval from the National Information Governance Board for Health and Social Care and the National Health Service North West Centre for Research Ethics Committee (Ref: 11/NW/0382), maintains its own internal ethics board (UK Biobank Ethics Advisory Committee), and is compliant with the General Data Protection Regulation (GDPR) [14, 15]. Participants enrolled in the UKB provided electronic signed consent at the time of enrollment [14, 15].

SNP array quality control

All genotype arrays are aligned to the GRCh37 build. After extracting samples from the UKB and removing duplicates, additional quality control (QC) filtering steps were taken to remove relatives, very rare variants, and filter out SNPs with characteristics more prone to, or indicative of, genotyping errors (Table S1.7). Genotype arrays for chromosomes 1-22 were merged (*n* = 784,256 variants) and filtered using PLINK v2.00a2.3 (--geno 0.10, --hwe 0.000001, --maf 0.000097, --mind 0.90), removing 23,732 variants due to missing genotype data, 13,313 variants due to Hardy-Weinberg equilibrium threshold, and 51,860 variants due to allele frequency threshold. First and second-degree relatives were identified with KING v2.1.8 (--unrelated, --degree 2) and removed, retaining 8,967 samples (1,321 individuals were removed due to relatedness). The array was then filtered once more to remove any remaining rare variants (--maf 0.0001) after relatives were removed, resulting in 692,466 variants remaining.

Principal components analysis

Principal component analysis (PCA), support vector machines (SVM), and ADMIXTURE analyses were performed to characterize the genetic affinities of UKB participants who self-identify as Bangladeshi, Indian, Pakistani, “White and Asian” (WA), and “Any other Asian”

(AOA). To identify if Bangladeshi-, Indian-, and Pakistani-identifying participants represent a genetically South Asian metapopulation and to elucidate the genetic affinities of AOA and WA participants, a principal component analysis (PCA) was performed using FRAPOSA and the 1KG data as a reference for global genetic diversity (Table S1.6) [23, 60].

To conduct these analyses, the QC-ed UKB array was merged with the 1KG WGS data, resulting in an intersection of 668,051 shared SNPs. The 1KG data were used to calculate 20 PCs and the UKB data were projected onto this PC space using FRAPOSA [23]. PC plots were visually assessed to identify large-scale patterns within and between the UKB groups.

However, unexpected data issues were discovered after visualizing the PC plots derived from the initial merge with 1KG data. PCA results indicated that CEU and GBR explained most of the genetic variance within the 1KG reference data, prompting further investigation of the data quality (see SI Section 1.2.3; Figures S1.3 – S1.5). The 1KG data were LD thinned (Plink v2.00a2.3; --indep-pairwise 50 20 0.5), and 113,571 variants were removed from the merged dataset (retaining 514,708 variants). To test if unidentified strand-flipping between the merged datasets was contributing to the issues (i.e., if strand orientation had not been maintained throughout the plink QC pipeline), snpflip v0.0.6 was run (Biocore NTNU, 2017), but only seven variants were identified as possibly flipped. Next, F_{ST} was calculated (using PLINK v1.90p --fst flag) for samples putatively representing the same population: UKB Bangladeshi and 1KG Bengali in Bangladesh (BEB), and UKB "White"/British with 1KG GBR. All SNPs across all checks with $F_{ST} > 0.15$ were subsequently removed, though this only filtered out 144 variants. The PCA was repeated but visualizations of the results demonstrated that this did not resolve the issue.

Next, outliers (samples) were sequentially removed from the 1KG PCA and the PCA was repeated. Because the first several rounds of "outlier identification" kept implicating CEU samples as the culprit, the entire CEU sample was removed from analyses. Subsequently, GBR samples began driving axes of variation across abnormally low PCs. The most obvious irregularity in PCA outputs remained the inflated weighing of European genetic variance relative to global patterns of genetic diversity, therefore, ascertainment bias in the UKB array was suspected.

To test if European genetic variance was being falsely inflated by design of the array, all but the most common genetic variants were removed and the PCA was repeated. Variants were filtered out at 0.1, 0.05, and 0.02 MAF for PCs 1-10 and the results were visualized to assess if excluding rare variants recovered expected patterns of global genetic diversity. Of these three thresholds visualized, only the most stringent (removing SNPs with MAF < 0.1) replicated expected patterns in global genetic diversity in which European variance isn't artificially inflated. Therefore, this intersection of 199,495 shared SNPs in both the UKB and 1KG arrays were utilized for the PCA (representing variants with MAF > 0.1 and < 0.99). To additionally assess if rare variants were having undue influence on PC weights, MAF was plotted against the absolute PC weight, for PCs 1-10 (Fig. S1.2). Visual assessment indicated this was true for six of the first 10 PCs.

Supervised genetic ancestry assignment

An SVM model was then developed to quantitatively integrate information from PCs 1-15 to identify the most genetically similar reference group(s) to WA and AOA participants, using R package “e1071” [61]. The SVM model was trained on a reduced sample of UKB Bangladeshi (*n* = 208) and UKB Pakistani (*n* = 1,484) participants, all 1KG Asian populations, a combined 1KG African group (AFR) (all African populations pooled; *n* = 474), and a combined 1KG European group (EURO) (all European populations pooled; *n* = 394). See SI Section 2.3.2 for further details. This model was then used to classify the most probable population affiliation for each AOA and WA participant. SVM classifications were compared to the participants' region of origin (UKB data field 20115) as an applied assessment of general accuracy (e.g., if AOA or WA participants classified as STU report being from Sri Lanka). Because the training population options were limited, each AOA and WA participant was also assigned to a metapopulation based on the sum of classification probabilities across all training populations comprising a subcontinental region. AOA and WA participants with a South Asian metapopulation summed probability score ≥ 0.7 were subsequently added to the UKB South Asian group, formerly only composed of Bangladeshi-, Indian-, and Pakistani-identifying (BIP) participants.

Genetic affinities of the newly formed UKB South Asian group were assessed further by comparing SVM classifications and reference populations via ADMIXTURE analyses. The same 1KG data as utilized in the PCA were subjected to unsupervised ADMIXTURE analyses (run once for k = 3 - 9) and were used to guide the supervised analyses of the UKB data. The results were visualized in *pong* to determine if the SVM classifications reasonably aligned with the reference populations [62, 63].

Environmental covariates

To maximize the retention of participants incorporated in the GWAS, environmental covariates were imputed using classification and regression trees (CART) within the R statistical package MICE (Multiple Imputation by Chained Equations), resulting in five completed datasets [64, 65] (Figure S2.3). Missingness of each covariate prior to imputation is reported in Table 2 (refer to SI Section 2.2 for further details regarding the environmental covariate imputation process).

After receiving the initial 283 environmental variables requested, the data were segregated into broad categories (Figure S2.2 B). Within each domain, the etiology of each variable, and its relationships with all other variables and height were considered (Figure S2.2 B). Correlations among variables within domains were tested, and data conferring the same or highly correlated information were removed (e.g., “Long-standing illness” and “Overall health rating”). Data hypothesized or known to have a strong genetic etiology that may be shared with the GWAS phenotype were noted (e.g., sitting height and “comparative body size at age 10”) following the guidance for covariate selection outlined in Aschard et al. [56]. Variables of this nature are not recommended for use as environmental covariates in the GWAS but were retained to maximize imputation performance (described below). The data were then assessed for missingness and all variables falling above 20% missingness were discarded prior to any further analyses.

Variables that passed the missingness threshold were then subjected to a multiple imputation (MI) procedure, in order to retain the largest sample of individuals with both genotype and standing height data. MI has been demonstrated to perform as well as or better than the methods typically employed for dealing with uncertainty in environmental data within genetic studies [66,67,68]. Additionally, utilizing MI provides an estimate of the uncertainty produced by using imputed values in downstream analyses, which more transparently reflects the final results [69,70,71].

Several of these variables (uniformly, within each of the five completed datasets) were then recoded for simplicity and to better represent the variable of interest, as it pertains to skeletal health. For example, a dietary restriction variable (UKB data field 6144) originally allowed for six categorical responses describing which food products the participant “never eats”, with “dairy products” as an available option. Of all possible dietary data fields with dairy consumption information, this data field had the least missingness (~2%) within our study population. Therefore, this variable was recoded as a binary response, such that all responses that selected “dairy products” were recoded as “excludes dairy” and all others were recoded as “does not exclude dairy”.

Subsequent to imputation, nine variables were selected for inclusion as environmental covariates for the GWAS (Table 2). To test if these specific environmental variables are significantly associated with height, and to identify the variance in height explained by additive environmental components alone, an ordinary least squares (OLS) regression model was assessed. The full UKB South Asian sample with un-imputed data (*n* = 7,331) were used for this assessment, using the base R statistical framework (version 4.2.2).

Data recoding scripts are available at https://github.com/ccatram/AnthroGen.

Imputation of genetic data

Imputed genetic data were utilized for the GWASs. Although the UKB provides imputed data, derived from a combined British and global reference panel (UK 10K and 1000 Genomes), imputation accuracy has not been tested for non-European participants (See Bycroft et al., 2018 [14], UKB Resource 531[https://biobank.ndph.ox.ac.uk/ukb/refer.cgi?id=531], UKB Category 100319 [https://biobank.ndph.ox.ac.uk/showcase/label.cgi?id=100319], and Marchini et al., 2016 [72] for details on the UKB imputation procedures for data versions 1-3). Performance was only tested for imputing the UKB array variants on a sample of 10 Europeans, using data inferred from the British-specific reference panel, alone. Additionally, the info score provided in the imputed data files reflects the average info score across the entire UKB sample, which is predominantly comprised of British, Irish, or “white” identifying participants. Therefore, the provided info score realistically only reflects the certainty in predicting genotypes in Europeans and is of low utility for all other participants. A different imputation was deemed necessary for our sample population (UKB whole genome sequences were not yet released), and the Genome Asia 100K Project (GAsP) was selected for use as the reference panel as it reflected the most appropriate collection of samples publicly available at the time of imputation (2023) [73]. At that time, there were 489 South Asian participants within the full reference panel used by the UKB and 724 within GAsP (phase 1 data).

To prepare the UKB genotype data for imputation, the full array (*n* = 692,466) for the UKB South Asian sample (*n* = 8,178; defined in Ch.2) was QC-ed using Plink v2.00a2.3 (--geno 0.10, --hwe 0.00001, --maf 0.00006, --mind 0.90) and the remaining variants (*n* = 690,869) were converted to vcf format (--recode vcf). Bcftools v1.19 [74] was used to orient the strands to the hg19 reference, split the data by chromosome, and sort by genomic position. These data were then used as the input for imputation via the Michigan Imputation Server (pipeline version 1.7.3, using minimac 4-1.0.2 for imputation, eagle-2.4 for phasing, and R^2 filter of 0) [75]. SNPs with info scores below 0.8 were removed from the returned imputed data, leaving 6,946,575 remaining variants to be used for the GWAS.

Height GWAS models

To assess how environmental covariate inclusion may affect GWAS results, two GWAS models were run on a subset of the sample (*n* = 6,954; 1,000 randomly chosen participants were removed for PGS development) and their results compared. $\text{GWAS}_{\text{null}}$ includes sex, year of birth (YOB), a genetic relatedness matrix (GRM), and the first 15 genetic PCs as covariates in the model, mirroring the typical covariates included in other published height GWAS. GWAS_{env} includes all variables used in $\text{GWAS}_{\text{null}}$ plus seven additional covariates encompassing socioeconomic, dietary, and demographic influences (Table 2).

The GWASs were run using GCTA v1.93 MLMA-LOCO (--mlma-loco), a powerful mixed linear model based association analysis [76]. The genetic relatedness matrix (GRM) used in the GWAS models was derived from the UKB genotype array data, QC-ed using Plink v2.00a4 (--geno 0.05, --maf 0.1, --max-maf 0.99; *n* = 251,556 remaining variants) and calculated with GCTA. The GRM was then used as input for the GCTA PCA (--pca 20) that generated the 15 PCs included in the GWAS models. Although 20 PCs were calculated, visual inspection suggested that only PCs 1-15 provided relevant information regarding population substructure.

GWAS_{env} was run once for each environmental covariate imputed dataset (i.e., five GWAS were run) and the results were pooled following Rubin's recombining rules, providing pooled effect sizes, standard errors, and p-values for each variant [70] (Figure S2.3).

The effects of environmental covariate inclusion in the GWAS models were evaluated in two primary manners: comparing the number of significant associations and assessing changes in effects for significant associations. To assess the effects of environmental variables on our ability to detect significant genetic associations with height, the number of variants meeting several p-value thresholds were calculated for both GWAS outputs and these absolute values were compared. To characterize differences in SNP effects between GWAS results, three approaches were taken. First, variants with $p < 10^{-6}$ output by GWAS_{env} were identified and the GWAS_{env} beta values (β_{env}) were subtracted from the $\text{GWAS}_{\text{null}}$ beta values (β_{null}) to ascertain the relative change in effect sizes from the “default” height GWAS model. Secondly, for this same sample of SNPs, the absolute differences and percent differences between β_{null} and β_{env} were calculated. Thirdly, to qualitatively assess discrepancies in β distributions, density plots were visualized, both genome-wide and within chromosomes.

Lastly, FUMA-GWAS [77] was used to summarize functional annotations of the GWAS_{env} output and assess overlap with SNP associations currently published in the GWAS catalog [78]. Input parameters and results of the analyses can be found in SI Section 2.3.1.

Polygenic score (PGS) assessments and comparisons

A height PGS model derived from GWAS_{env} summary statistics was developed using LDpred2-auto within the ‘bigsnpr’ R package [30, 31]. Only SNPs also found within HapMap3 were incorporated in the PGS (n = 896,160), and all default values recommended for the ‘snp_ldpred2_auto()’ function were used for the analysis (following Privé’s 2023 vignette, which can be found here: https://privefl.github.io/bigsnpr/articles/LDpred2.html [30]). LDpred2-auto was chosen for our PGS development as it infers SNP heritability (h^2) and polygenicity from the LDpred model without necessitating a validation set for parameter-tuning, allowing us to retain a larger sample size than would be possible otherwise [30]. Performance was tested by obtaining adjusted R^2 between PGS scores and height for UKB South Asian participants removed from the sample prior to GWAS analyses (n = 997; 1,000 participants were randomly selected, but three were later removed due to data missingness).

The utility of our PGS equation was then tested by assessing its performance relative to the South Asian PGS equation created by Yengo et al., 2022 [32], which was developed from a GWAS meta-analysis (*n* = 77,890) and tested on UKB South Asian participants (*n* = 9,257). The associated PGS scoring file (PGS002800) was downloaded from the PGS Catalog [33] and applied to our PGS test sample using ESCALATOR v.2.0.0 (Lin and Fisher, 2022 [34]; [https://github.com/menglin44/ESCALATOR]), a PGS harmonization pipeline.

PGS002800 performance was assessed for two subsets of SNPs: PGS_{Y1} : the subset of SNPs found in our GWAS (and passing ESCALATOR quality control filtering) (*n* = 898,486), and PGS_{Y2} : the subset of variants that were available within our imputed data (*n* = 1,032,672). The Yengo et al. South Asian PGS makes use of 1,156,741 variants in its entirety, but we were unable to include the complete set in our analyses due to poor imputation quality for 12% of SNPs [32].

Additionally, to test if PGS002800 performance may be inflated due to unadjusted environmental influences (as opposed to the alternative hypothesis that higher relative performance is due to a better-powered training sample), R^2 between PGS scores and the model covariates were calculated for each combination of the data subsets referenced above (Table S2.12). While we expect correlations between PGS scores and environmental variables to be very low, we predict that if the effects of collinearity impacted the development of PGS002800 β weights then the resulting PGS scores will exhibit higher R^2 values compared to the PGS scores derived from our environmentally adjusted GWAS data. Further investigations into the influence of environmental adjustments on PGS development can be found in SI Section 2.4.2.

REFERENCES

1. Sirugo G, Williams SM, Tishkoff SA. The missing diversity in human genetic studies. Cell. 2019;177(1):26–31.

2. Ben-Eghan C, Sun R, Hleap JS, Diaz-Papkovich A, Munter HM, Grant AV, et al. Don’t ignore genetic data from minority populations. Nature. 2020;585(7824):184–6.

3. Privé F, Aschard H, Carmi S, Folkersen L, Hoggart C, O’Reilly PF, et al. Portability of 245 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups from the same cohort. Am J Hum Genet. 2022;109(1):12–23.

4. Weissbrod O, Kanai M, Shi H, Gazal S, Peyrot WJ, Khera AV, et al. Leveraging fine-mapping and multipopulation training data to improve cross-population polygenic risk scores. Nat Genet. 2022;54(4):450–58.

5. Ding Y, Hou K, Xu Z, Pimplaskar A, Petter E, Boulier K, et al. Polygenic scoring accuracy varies across the genetic ancestry continuum. Nature. 2023;618:1–8.

6. Duncan L, Shen H, Gelaye B, Meijsen J, Ressler K, Feldman M, et al. Analysis of polygenic risk score usage and performance in diverse human populations. Nat Commun. 2019;10(1):3328.

7. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet. 2019;51:584–91. https://doi.org/10.1038/s41588-019-0379-x

8. Novembre J, Stein C, Asgari S, Gonzaga-Jauregui C, Landstrom A, Lemke A, et al. Addressing the challenges of polygenic scores in human genetic research. Am J Hum Genet. 2022;109(12):2095–2100. https://doi.org/10.1016/j.ajhg.2022.10.012

9. Wang Y, Guo J, Ni G, Yang J, Visscher PM, Yengo L. Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. Nat Commun. 2020;11(1):3865.

10. Cavazos TB, Witte JS. Inclusion of variants discovered from diverse populations improves polygenic risk score transferability. Hum Genet Genom Adv. 2021;2(1):100017.

11. Ruan Y, Lin YF, Feng YCA, Chen CY, Lam M, Guo Z, et al. Improving polygenic prediction in ancestrally diverse populations. Nat Genet. 2022;54(5):573–80.

12. Tian P, Tsai HC, Wang Y, Yang Y, Yin G, Zhang YD. Multiethnic Polygenic Risk Prediction in Diverse Populations through Transfer Learning. Front Gen. 2022;13. https://doi.org/10.3389/fgene.2022.906965.

13. Coram MA, Candille SI, Duan Q, Chan KHK, Li Y, Kooperberg C, et al. Leveraging Multi-ethnic Evidence for Mapping Complex Traits in Minority Populations: An Empirical Bayes Approach. Cell. 2015;96(5):740–52.

14. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018;562(7726):203.

15. Littlejohns TJ, Holliday J, Gibson LM, Garratt S, Niels Oesingmann O, Alfaro-Almagroet F, et al. The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. Nat Commun. 2020;11:2624. https://doi.org/10.1038/s41467-020-15948-9

16. Constantinescu AE, Mitchell RE, Zheng J, Bull CJ, Timpson NJ, Amulic B, et al. A framework for research into continental ancestry groups of the UK Biobank. Human Genom. 2022;16(1):1–14.

17. Cardena MM, Ribeiro-dos-Santos A, Santos S, Mansur AJ, Pereira AC, Fridman C. Assessment of the relationship between self-declared ethnicity, mitochondrial haplogroups and genomic ancestry in Brazilian individuals. PloS one. 2013;8(4):e62005.

18. Benn Torres J. Anthropological perspectives on genomic data, genetic ancestry, and race. Am J Phys Anthro. 2020 May;171:74–86.

19. Mathieson I, Scally A. What is ancestry?. PLoS Genet. 2020;16(3):e1008624.

20. Wojcik GL. Genetic distance informs polygenic score predictive accuracy. Trends Genet. 2023;39(11):813–15.

21. Hou K, Ding Y, Xu Z, Wu Y, Bhattacharya A, Mester R, et al. Causal effects on complex traits are similar for common variants across segments of different continental ancestries within admixed individuals. Nature Genet. 2023;55(4):549–58.

22. Yang Z, Cieza B, Reyes-Dumeyer D, Montesinos R, Soto-Añari M, Custodio N, et al. A benchmark study on current GWAS models in admixed populations. Briefings Bioinform. 2024;25(1):bbad437.

23. Zhang D, Dey R, Lee S. Fast and robust ancestry prediction using principal component analysis. Bioinform. 2020;36(11):3439–46. https://doi.org/10.1093/bioinformatics/btaa152

24. Sengupta D, Choudhury A, Basu A, Ramsay M. Population stratification and underrepresentation of Indian subcontinent genetic diversity in the 1000 genomes project dataset. Genome Biol Evol. 2016;8(11):3460–70.

25. Reich D, Thangaraj K, Patterson N, Price AL, Singh L. Reconstructing Indian population history. Nature. 2009;461(7263):489–94.

26. Coop G. Genetic similarity versus genetic ancestry groups as sample descriptors in human genetics. arXiv. 2022;arXiv:2207.11595.

27. Lewontin RC. The Apportionment of Human Diversity. In: Dobzhansky T, Hecht MK, Steere WC, editors. Evolutionary Biology. Springer, New York, NY; 1972. p. 381–398.

28. Abrams SA. Normal acquisition and loss of bone mass. Hormone Res. 2003;60 Suppl 3:71–6. https://doi.org/10.1159/000074505.

29. Waldron T. Palaeopathology. Cambridge: Cambridge University Press; 2008. (Cambridge Manuals in Archaeology).

30. Privé F, Albiñana C, Arbel J, Pasaniuc B, Vilhjálmsson BJ. Inferring disease architecture and predictive ability with LDpred2-auto. Cell. 2023;110(2):2042–55.

31. Privé F, Luu K, Blum MG, McGrath JJ, Vilhjálmsson BJ. Efficient toolkit implementing best practices for principal component analysis of population genetic data. Bioinformatics. 2020;36(16):4449–57.

32. Yengo L, Vedantam S, Marouli E, Sidorenko J, Bartell E, Sakaue S, et al. A saturated map of common genetic variants associated with human height. Nature. 2022;610:704–12.

33. Lambert SA, Wingfield B, Gibson JT, Gil L, Ramachandran S, Yvon F, et al. Enhancing the Polygenic Score Catalog with tools for score calculation and ancestry normalization. Nat Genet. 2024. https://doi.org/10.1038/s41588-024-01937-x.

34. Lin M, Fisher M. ESCALATOR: polygEnic SCore hArmonization and calcuLATiOn scRipts [Software]. Github; [cited 2024 Oct 27]. Available from:https://github.com/menglin44/ESCALATOR

35. Metspalu M, Mondal M, Chaubey G. The genetic makings of South Asia. Curr Opin Genet Dev. 2018;53:128–33.

36. Rangaswamy P. (2005). South Asian Diaspora. In: Ember M, Ember CR, Skoggard I, editors. Encyclopedia of Diasporas. Springer, Boston, MA; 2005. p. 285–96. https://doi.org/10.1007/978-0-387-29904-4_28

37. McAllister K, Mechanic LE, Amos C, Aschard H, Blair IA, Chatterjee N, et al. Current challenges and new opportunities for gene-environment interaction studies of complex diseases. Am J of Epidemiol. 2017;186(7):753–61.

38. Keller MC. Gene x environment interaction studies have not properly controlled for potential confounders: the problem and the (simple) solution. Biol Psychiatry. 2014;75(1):18–24.

39. Cooley PC, Clark RF, Folsom RE. Assessing gene-environment interactions in genome-wide association studies: statistical approaches. RTI Press; 2014.

40. Holick MF, Dawson-Hughes B. Nutrition and bone health. Totowa, NJ: Humana Press; 2004

41. Alacevich C, Tarozzi A. Child height and intergenerational transmission of health: Evidence from ethnic Indians in England. Econ Hum Biol. 2017;25:65–84.

42. Anderson, JJB. The important role of physical activity in skeletal development: How exercise may counter low calcium intake. Am J Clin Nutr. 2000;71:1384–6.

43. Bogin B, Smith P, Orden AB, Varela Silva MI, Loucky J. Rapid change in height and body proportions of Maya American children. Am J Hum Biol. 2002;14(6):753–61.

44. Costa-Font J, Gil J. Generational effects and gender height dimorphism in contemporary Spain. Econ Hum Biol. 2008;6(1):1–18.

45. Deaton A. Height, Health, and Inequality: The Distribution of Adult Heights in India. Am Econ Rev. 2008;98(2):468–74.

46. Kaplan BA. Environment and Human Plasticity. Am Anthropol. 1954;56(5):780–800.

47. Whiting CD. Stature and Build of Hawaii-Born Youth of Japanese Ancestry. Am J Phys Anthropol. 1961;19:159–67.

48. Perkins JM, Subramanian SV, Davey Smith G, Özaltin E. Adult height, nutrition, and population health. Nutr Rev. 2016;74(3):149–65.

49. Whiting SJ, Vatanparast H, Baxter-Jones A, Faulkner RA, Mirwald R, Bailey DA. Factors that Affect Bone Mineral Accrual in the Adolescent Growth Spurt. J Nutr. 2004;134(3):696S–700S.

50. Yeboah J. Diet, height, and health. Am J Clin Nutr. 2017;106(2):443–444.

51. Zoledziewska M, Sidore C, Chiang CWK, Sanna S, Mulas A, Steri M, et al. Height-reducing variants and selection for short stature in Sardinia. Nat Genet. 2015;47(11):1352–6.

52. Jeon S, Choi H, Jeon Y, Choi W, Choi H, An K, et al. Korea4K: whole genome sequences of 4,157 Koreans with 107 phenotypes derived from extensive health check-ups. GigaScience. 2024;13:giae014, https://doi.org/10.1093/gigascience/giae014.

53. Sohail M, Palma-Martínez MJ, Chong AY, Quinto-Cortés CD, Barberena-Jonas C, Medina-Muñoz SG, et al. Mexican Biobank advances population and medical genomics of diverse ancestries. Nature. 2023;622:775–83.

54. Wojcik GL, Graff M, Nishimura KK, Tao R, Haessler J, Gignoux CR. Genetic analyses of diverse populations improves discovery for complex traits. Nature. 2019;570(7762):514–518.

55. Berg JJ, Harpak A, Sinnott-Armstrong N, Joergensen AM, Mostafavi H, Field Y, Boyle EA, Zhang X, Racimo F, Pritchard JK, Coop G. Reduced signal for polygenic adaptation of height in UK Biobank. Elife. 2019;8:e39725.

56. Aschard H, Guillemot V, Vilhjalmsson B, Patel CJ, Skurnik D, Ye CJ, et al. Covariate selection for association screening in multiphenotype genetic studies. Nat Genet. 2017;49(12):1789–95.

57. Emmett PJ, Jones LR, Golding J. Pregnancy diet and associated outcomes in the Avon Longitudinal Study of Parents and Children. Nutr Rev. 2015;73 Suppl 3:154–74.

58. Luque V, Escribano J, Closa-Monasterolo R, Zaragoza-Jordana M, Ferré N, Grote V, et al. Unhealthy Dietary Patterns Established in Infancy Track to Mid-Childhood: The EU Childhood Obesity Project. 2018;148(5):752–59.

59. Fairley S, Lowy-Gallego E, Perry E, Flicek P. The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. Nucleic Acids Res. 2020;48(D1):D941–D947.

60. 1000 Genomes Project Consortium et al. A global reference for human genetic variation. Nature. 2015;526(7571):68.

61. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F, Chang CC. Misc Functions of the Department of Statistics (e1071). The R J. 2019.

62. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19:1655–64.

63. Behr AA, Liu KZ, Liu-Fang G, Nakka P, Ramachandran S. pong: fast analysis and visualization of latent clusters in population genetic data. Bioinform. 2016;32(18):2817–23.

64. van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. Stat Methods Med Res. 2007;16(3):219–42. https://doi.org/10.1177/0962280206074463.

65. Doove LL, van Buuren S, Dusseldorp E. Recursive partitioning for missing data imputation in the presence of interaction effects. Comput Stat Data Anal. 2014;72:92–104.

66. Palmer C, Pe'er I. Bias Characterization in Probabilistic Genotype Data and Improved Signal Detection with Multiple Imputation. PLOS Genet. 2016:12(6):31006091. https://doi.org/10.1371/journal.pgen.1006091.

67. Ramstein GP, Lipka AE, Lu F, Costich DE, Cherney JH, Buckler ES, et al. Genome-wide association study based on multiple imputation with low-depth sequencing data: application to biofuel traits in reed canarygrass. G3 (Bethesda, Md.). 2015;5(5):891–909.

68. Croiseau P, Génin E, Cordell HJ. Dealing with missing data in family-based association studies: a multiple imputation approach. Hum Hered. 2007;63(3–4):229–38. https://doi.org/10.1159/00010048.

69. Nakagawa S, Freckleton RP. Model averaging, missing data and multiple imputation: A case study for behavioural ecology. Behav Ecol Sociobiol. 2011;65(1):103–16. https://doi.org/10.1007/s00265-010-1044-7.

70. Little RJ, Rubin DB. Statistical analysis with missing data. 3rd ed. Hoboken (NJ): John Wiley & Sons; 2019.

71. Schafer JL, Olsen MK. Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective. Multivariate Behav Res. 1998;33:545–71. https://doi.org/10.1207/s15327906mbr3304_5

72. Marchini J, Abecasis G, Durbin R. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet. 2016;48(10):1-279

73. GenomeAsia100K Consortium. The GenomeAsia 100K Project enables genetic discoveries across Asia. Nature. 2019;576:106–11. https://doi.org/10.1038/s41586-019-1793-z

74. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021;10(2):giab008. https://doi.org/10.1093/gigascience/giab008

75. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016;48:1284–87.

76. Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. Advantages and pitfalls in the application of mixed-model association methods. Nat Genet. 2014;46(2):100–6.

77. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. Nat Commun. 2017;8:1826.

78. Sollis E, Mosaku A, Abid A, Buniello A, Cerezo M, Gil L, et al. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. Nucleic Acids Res. 2023;51(D1):D977-D985. https://doi.org/10.1093/nar/gkac1010.