ESHG

www.nature.com/ejhg

ARTICLE OPEN

Check for updates

Ancestral genetic components are consistently associated with the complex trait landscape in European biobanks

Vasili Pankratov[1]✉, Massimo Mezzavilla[2], Serena Aneli[3], Ivan A. Kuznetsov[1], Daniela Fusco[4], James F. Wilson[5,6,7], Mait Metspalu[8], Paolo Provero[4,9], Luca Pagani[2,8] and Davide Marnetto[4]✉

The genetic structure in Europe was mostly shaped by admixture between the Western Hunter-Gatherers, Early European Farmers and Steppe Bronze Age ancestral components. Such structure is regarded as a confounder in GWAS and follow-up studies, and gold-standard methods exist to correct for it. However, it is still poorly understood to which extent these ancestral components contribute to complex trait variation in present-day Europe. In this work we harness the UK Biobank to address this question. By extensive demographic simulations, exploiting data on siblings and incorporating previous results we obtained from the Estonian Biobank, we carefully evaluate the significance and scope of our findings. Heart rate, platelet count, bone mineral density and many other traits show stratification similar to height and pigmentation traits, likely targets of selection and divergence across ancestral groups. We show that the reported ancestry-trait associations are not driven by environmental confounders by confirming our results when using between-sibling differences in ancestry. The consistency of our results across biobanks further supports this and indicates that these genetic predispositions that derive from post-Neolithic admixture events act as a source of variability and as potential confounders in Europe as a whole.

European Journal of Human Genetics; https://doi.org/10.1038/s41431-024-01678-9

INTRODUCTION

In order to uncover the genetic basis of complex traits in Genome Wide Association Studies (GWAS) [1], a large amount of data has been collected in nation-wide population-based Biobanks [1–5]. Despite including predominantly individuals of European ancestry, which poses well-recognized limitations [1, 6–8], these cohorts still contain stratification such as socio-economic disparities, geographic factors and, most importantly, inherent genetic structure, which might result in non-causal associations. A plethora of methods [9–11] have been developed to correct for these unwanted sources of variance that might bias GWAS discovery. Indeed even the finer cases of population structure present in national Biobanks [12–14] have been demonstrated to affect GWAS [14, 15] and, if not carefully addressed, hamper analyses following up on these results, such as polygenic risk scoring [12, 16–19] and polygenic selection testing [20, 21].

Millennia of demographic expansions, migrations, and localized genetic isolation have indeed shaped a far from homogeneous genetic makeup for contemporary Europeans. Besides recent demographic history, predominantly impacting on rare variation structure, a series of admixture events from 8000 BCE to 1000 BCE have been foundational to the European genetic landscape. During that period, Early European Farmers (EEF) and Steppe Bronze Age (SBA) genetic ancestries gradually spread into and across Europe blending with the local Western Hunter-Gatherers (WHG) substratum [22–24], bringing together genetic components that had evolved separately for up to 20,000 years [25]. Divergent phenotypes in these source populations have been previously described for a few traits using polygenic scoring of ancient samples [26–28], and very recently of ancestral segments from modern samples [29], or looking at specific trait-informative Single Nucleotide Polymorphisms (SNPs) [30–32].

These ancestral components can explain a large part of the genetic gradients across Europe [23] and, as a result, are indirectly adjusted for in gold-standard GWAS procedures. In this work, we aim to identify which complex trait variations across the largest European biobank can be explained by the stratification of these ancestral components. Importantly, although genetic predispositions conferred by these components might be indirect and involve pleiotropy and/or complex interactions with the environment, we devote particular care to avoid apparent associations caused by geographic and socio-economic stratifications co-occurring with the ancestral components.

We previously attempted to quantify their differential contribution to the contemporary landscape of complex traits in 35,000 individuals from the Estonian Biobank (EstBB) [33]. Our findings provided the first picture to describe the association of ancestral populations to present day traits; however, it remained unclear

[1]Center for Genomics, Evolution and Medicine, Institute of Genomics, University of Tartu, 51010 Tartu, Estonia. [2]Department of Biology, University of Padua, Padua, Italy. [3]Department of Public Health Sciences and Pediatrics, University of Turin, 10126 Turin, Italy. [4]Department of Neurosciences, University of Turin, 10126 Turin, Italy. [5]Centre for Global Health Research, Usher Institute, University of Edinburgh, Teviot Place, Edinburgh EH8 9AG, Scotland. [6]MRC Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, Scotland. [7]Centre for Genomic and Experimental Medicine, Institute of Genetics and Cancer, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, Scotland. [8]Institute of Genomics, University of Tartu, 51010 Tartu, Estonia. [9]Center for Omics Sciences, IRCCS San Raffaele Scientific Institute, 20132 Milan, Italy. ✉email: vasili.pankratov@ut.ee; davide.marnetto@unito.it

Received: 29 December 2023 Revised: 23 July 2024 Accepted: 25 July 2024
Published online: 10 August 2024

whether our conclusions could be applied to the broader European population. In addition, we could not decisively address the differences in the resulting signal when running the analysis over local candidate regions or over the whole genome. Finally, the study lacked a meticulous simulation testing the limits of the adopted statistic in answering our question.

Here we overcome these limitations by analyzing a total of 53 complex traits in 50,000 UK Biobank [2] (UKBB) donors taken as representative of the Western European metapopulation. We validate the ancestry-trait associations found as European-wide signals, find new ones, and explore the strengths and limitations of our approach with thorough simulations using SLiM [34].

SUBJECTS AND METHODS

Framework overview

Our approach is based on *covA*, a measure of relative genetic similarity of an individual from a contemporary population to the distinct ancestries (e.g. WHG; EEF; SBA) that contribute to the genetic makeup of that population. This metric was introduced in our previous work [33] for the analysis of complex traits in EstBB and is ultimately a covariance between allele dosage in a contemporary individual and a given ancestral population, with respect to the contemporary and ancient average frequencies. We regress each complex trait *t* in the present-day dataset on the *covA* for each ancestry *p* so that for each individual *i*:

$$t_i = \beta_0 + \beta_{\mathrm{cov}A(p)} \cdot \mathrm{cov}A(i,p) + \varepsilon_i$$

where the slope $\beta_{\mathrm{cov}A(p)}$ quantifies the association between each ancestry *p* with the trait *t* and ε_i represents the error; covariates can potentially be added (see **Models tested and covariates**).

In order to reduce the influence of genomic confounding factors, we compute *covA* restricting to variants belonging to 20 kb genomic regions around SNPs that have been previously associated with the trait of interest through GWAS, i.e. Trait-Associated Genomic Regions (TAGR, see **Traits and candidate regions**). Note that GWAS-derived summary statistics are used only to identify TAGRs: using these summaries as weights in follow-up analyses, especially when summed across the whole genome, has been shown to produce results difficult to interpret in populations that are even subtly genetically differentiated from the ones where the GWAS was run [12, 21, 35, 36].

Simulations

Below is a brief description of the simulations. See Supplementary Text for further information.

To explore the behavior of *covA* under different heritability, polygenicity, stabilizing selection and differences in trait optima between ancestral populations we performed hybrid simulations using a combination of SLiM [34] and msprime [37]. We simulated 1000 unlinked genomic intervals of 20 kb each (portraying potential TAGRs) with a uniform recombination rate of 1e-8 and a uniform mutation rate of 1.25e-8 [38] under a demographic model relevant for the British population (Fig. S1) with each such interval containing one trait-affecting SNP. In each generation we calculated the genetic value (GV) for each individual based on the genotypes at these SNPs, converted this GV to the trait value by adding environmental noise and then mapped it to fitness using an approach similar to the one used by Yair and Coop, 2022 [39]. All in all, we ran 2370 simulations representing 135 simulation scenarios.

Each simulation run resulted in genotypes and trait values for 10 K contemporary individuals (around 7 K after filtering for relatives) and 100 individuals from each of the three reference populations (WHG, EEF, SBA; no relatives filtering applied). The causal SNPs were ascertained based on the fraction of heritability of the trait they explain to mimic SNP discovery in GWAS, resulting in between 351 and 565 TAGR used to calculate *covA* (Table S1). Finally, the simulated traits and *covAs* were plugged in the regression above without any other covariate.

Sample selection and definition of ancestral groups

UKBB [2] donors were selected among those (a) identified as British of West-European descent (code 1) according to Data-Field 22006 (which includes self-identification and genetic grouping), (b) selected for the UKBB Principal Component Analysis (Data-Field 22020), which excludes extreme heterozygosity, missingness and up to third-degree relatives; and (c) having at most one missing trait among those analyzed. We then extracted a subset of 50,000 samples equally divided between females and males for the analyses.

To validate our results using differences across siblings, we identified 17,319 sibships (35,585 samples) among the UKBB donors identified as British of West-European descent as above, keeping pairs with kinship coefficient between 0.177 and 0.354 and identified as full siblings by KING v2.2.7 based on IBD2 sharing. In case a donor was involved in more than one sibling pair, we required all donors within such a sibship to be identified as full siblings, otherwise the whole sibship was removed.

Ancient samples were extracted from the Allen Ancient DNA Resource [40] v52.2 (AADR) following the approach described in Marnetto et al. [33]. After starting from a manually curated core set for each ancestral group, we expanded these sets to other AADR samples according to distance in a 4-dimensional space defined by dating and first 3 Principal Components (PC). PCs were determined on a set of modern Eurasian and North African individuals west of Iran (included), where the ancient samples were projected; distance cutoff was defined by multi-dimensional ellipses with diameters equal to 3 core set SDs. See Table S2 for a full list of ancestral group classifications, coordinates, dates and PCs.

We used phased, imputed genotype data for the UKBB set and intersected it with the ancient set, obtaining 1,087,822 genotyped variants in the merged dataset.

Traits and candidate regions

Traits were selected to span different domains and favor large effective sample sizes, see Table S3 for a complete list accompanied with the original UKBB Field ID. When multiple data points existed for an individual, the earliest one was considered. For each continuous trait, individuals with values more distant than 4 IQRs from the upper or lower quartile were discarded as outliers, then traits were standardized. Some traits were computed (waist/hip ratio, pulse pressure, caffeine, etc.), log-transformed (body mass index (BMI), creatinine,…) or BMI-adjusted (waist and hip circumference, blood pressure, etc.), see Table S3.

Trait-associated genomic regions (TAGRs) were defined starting from GWAS Catalog [41] data, downloaded on 09/05/2023. This resource collects small-scale variant hits which, depending on the original study, either are genome-wide significant (p-value $< 0.5 \cdot 10^{-8}$) or genome-wide suggestive (p-value $< 10^{-5}$). We selected hits by matching a pattern to the trait definition given by the original study or by selecting a mapped Experimental Factor Ontology term, according to Table S4, then defined 20 kb windows centered on the selected hits and merged them, obtaining a set of TAGRs for each trait.

Partitioned heritability

In order to assess TAGRs contribution to heritability of traits, a Stratified LD Score Regression (sLDSC) was conducted on UKBB GWAS summary statistics estimated by the Neale Lab (http://www.nealelab.is/uk-biobank/) for 50 complex traits. We used 503 samples of European ancestry from the 1000 Genomes Project [42] to compute LD Scores; HapMap Project Phase 3 [43] SNPs, with minor allele frequency above 1% and INFO score above 95% were kept for the analysis. 50 TAGR sets were used as functional categories, in addition to another set of 50 “negative TAGRs” defined excluding 500kbs around each GWAS Catalog hit.

Models tested and covariates

We regressed each trait *t* on a linear model including the standardized *covA* for the ancestry *p* and a vector of covariates **c**:

$$t_i = \beta_0 + \beta_{\mathrm{cov}A(p)} \cdot \mathrm{cov}A(i,p) + \boldsymbol{\beta}_{\mathbf{c}}\mathbf{c} + \varepsilon_i$$

then estimated the $\beta_{\mathrm{cov}A(p)}$ coefficients. Categorical traits, which were transformed to {0, 1} where 1 stands for the category described in Table S3 and 0 for all the others, were regressed with a logistic regression. We instead adopted an ordinal logistic regression for ordinal traits: Table S3 describes the category order.

The covariates included in the model are (in parenthesis the corresponding UKBB Field ID): age (21022), sex (31), age^2, age×sex, age^2×sex, latitude (22703), longitude (22701), UKBB assessment center (54), urban or rural home area (20118, defined collapsing codes 1,5,11,12 into “urban”; 2,6,13,14,15 into “town”; 3,4,7,8,16,17,18 into “rural” categories), qualifications (6138), age when completed education (845) and Townsend deprivation index at recruitment (22189). We also added two genome-wide *covAs* (*GW-covA*) as covariates in every model to control for genome-wide ancestry which might

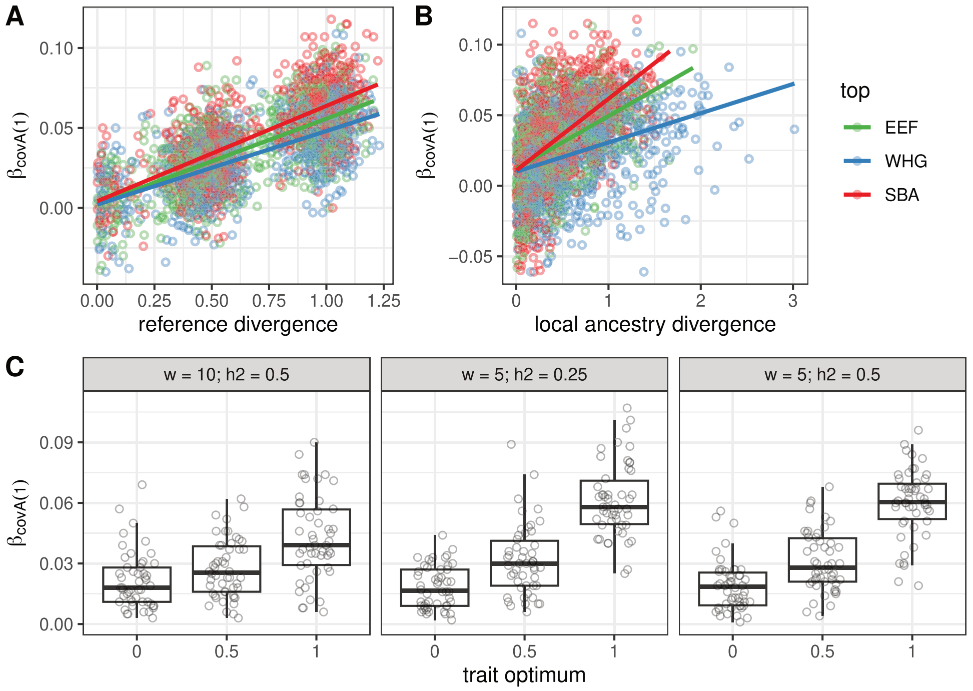

Fig. 1 Simulation results. A Effect on $\beta_{covA(1)}$ of reference populations differentiation, defined as the difference between the highest and the second highest mean genetic values among the three ancestries, normalized by trait standard deviation in the simulated present-day population (see Supplementary Text for details). **B** The same as **A** but plotting on x-axis local ancestry's mean genetic values instead. **C** $\beta_{covA(1)}$ values as a function of trait optimum in the ancestry with the highest genetic value ($p = 1$). The x-axis indicates the trait optimum in this ancestry, while other ancestries have an optimum set to 0. Three sub-panels present results of simulations with different fitness function SD (ω) and heritability (h^2) values. Each boxplot is based on 50 individual simulations. Scenarios with more than 50 simulation runs were randomly downsampled. The boxes show 25th, 50th and 75th quantiles, while the whiskers show values within 1.5 times the interquartile range (IQR).

be passively correlated with environmental confounders. Note that given that *covA* is a relative distance, with three ancestries we only have two degrees of freedom so adding the third would have generated perfect multicollinearity. Throughout the text "*TAGR-covAs*" refers to the statistic computed on the corresponding TAGRs for each trait, while "*GW-covA*" is the genome-wide statistic.

Then, the slope coefficient (β_{covA}), or the Odds Ratio (OR_{covA}) were directly used to assess ancestry-trait association for continuous and categorical traits respectively, double-sided *p*-values were computed by the glm R function. Significance was evaluated at Benjamini-Hochberg False Discovery Rate = 0.05. In all cases when reporting results for each trait we count multi-category traits only once.

When replicating our analysis using siblings we adopted the following model:

$$t_i = \beta_0 + \beta_{covA(p)}{'}\cdot covA(i,p)' + \beta_{covA(p)sib}\cdot covA(i,p)_{sib} + \boldsymbol{\beta_c c} + \varepsilon_i$$

where $covA(i,p)_{sib}$ is the mean *covA* for ancestry *p* of the sibship individual *i* belongs to, and *covA(i,p)′* is individual's *i* residual so that *covA(i,p)′* + $covA(i,p)_{sib}$ is the total *covA* value of individual *i* for ancestry *p*. It is then the $\beta_{covA(p)}{'}$ coefficient that we evaluate. The covariate vector includes age, sex, age^2, age×sex, age^2×sex and, when *covAs* are based on TAGRs, two *GW-covA(i,p)* as above.

RESULTS

Simulations

We started by exploring covA strengths and limitations in the simulation framework described above. To assess the performance of *covA*, we first ranked the three reference populations (equivalent to the sampled ancient genomes used to calculate *covA*) for their mean genetic value (GV), that is the expected phenotypic value without considering the environmental deviation. Therefore in each run $p = 1,2,3$ is the population with the highest, the median, and the lowest mean GV respectively. In scenarios of directional selection $p = 1$ corresponds to the ancestry experiencing a positive shift in trait optimum. We then tracked the slope obtained when regressing the simulated trait on *covA(1)* i.e. $\beta_{covA(1)}$.

We started by verifying whether β_{covA} depends on the difference in the trait's mean genetic value between reference populations. On average, higher $\beta_{covA(1)}$ values are observed as the difference in GV between ancestries $p = 1$ and $p = 2$ increases (Fig. 1A). In other words, the more genetically differentiated are the reference populations at TAGRs, the stronger is the association between *covA* and the trait value in present-day individuals. On the other hand, we see no independent effect of heritability, polygenicity or strength of stabilizing selection on the relationship between β_{covA} and genetic value differentiation (Fig. S2).

Notably, we can reliably identify the reference ancestry $p = 1$ as the one with the maximal β_{covA} as long as it is sufficiently differentiated from the ancestry $p = 2$: the true-positive rate is 78–87% if the difference between GV is within 0.25–0.5 of the trait's SD in the present-day population (Fig. S3A). Note, however, that β_{covA} values are negatively correlated between ancestries (Fig. S4), meaning that high absolute β_{covA} for one ancestry will be often complemented by high absolute β_{covA} for at least one other ancestry but with an opposite sign.

We next asked if β_{covA} can be informative about differences in trait optima between ancestries. We hypothesize that drift under stabilizing selection with the same optimum in all ancestries would result in lower $\beta_{covA(1)}$ as compared to scenarios with optimum shift in one of the ancestries. Although we do observe such a trend (Fig. 1C, S5), rather high values of $\beta_{covA(1)}$ (97.5 percentile of 330 simulations being equal to 0.048, see Table S5) can also be obtained in simulations with constant trait optimum.

Finally, we checked whether β_{covA} can be informative about the differential genetic contribution of ancestral populations to present-day trait variation. Intuitively, this can be measured as the expected GV of a present-day individual with all trait-affecting loci inherited from a specific ancestry. This value might deviate from the reference population's average GV due to post-admixture drift and selection. As can be seen in Fig. 1B and Fig. S6, β_{covA} remains informative about the difference in local ancestry GV between ancestries $p = 1$ and $p = 2$ but the true-positive rate of identifying the local ancestry with the highest GV is lower (57–78% if GV difference is within 0.25–0.5 of the trait's SD in the present-day population; Fig. S3B). The lowest sensitivity is observed for the WHG local ancestry likely because of its low contribution to the present-day population (12%) and hence stronger effect of post-admixture drift. However, as exemplified by scenarios **E** and **F** in Fig. S6, when the reference populations had the same trait optimum, our approach is blind to differences in local ancestry genetic value arising from post-admixture drift (**E**) or selection (**F**). Full simulation results are reported in Table S1.

Association with UK Biobank complex trait landscape

We applied the same approach to 50,000 unrelated individuals of European descent from the UKBB [2], so as to match the order of magnitude of our previous EstBB analysis [33]. The reference groups representing WHG, EEF and SBA (here defined as genetic ancestries, rather than as actual cultures or populations) included 95 WHG-, 118 EEF-, and 83 SBA-like ancient genomes, see list in Table S2, Fig. S7.

We identified TAGRs for each trait exploiting GWAS Catalog [41] hits for congruous traits (see Methods for details). Indeed, TAGRs were enriched in SNP heritability for the corresponding trait in almost all cases (stratified LD score regression [44], nominal P-value < 0.05, see Fig. S8A). We then computed *TAGR-covA*s for each ancestry on these regions and regressed phenotypic values for each trait on them. Notably, to capture any possible physiology-related, geographic and socio-economic confounder that might produce spurious false positives, we also include in this regression a large set of covariates: sex, age, age2, sex×age, sex×age2, latitude, longitude, UKBB assessment center, Townsend deprivation index (TDI), qualifications, and age when completed education. Importantly, we also included genome-wide *covA* for two ancestries (capturing all *GW-covA* variability) to control for genome-wide population structure.

An independent model was built for each of the ancestries as our simulations showed that a genetic value shift in a single ancestry could explain significant β_{covA} in all ancestries, and that the ancestry with the maximal absolute coefficient ($max(|\beta_{covA}|)$) is likely to be the most divergent one. As seen in Fig. 2A, 32 out of 53 traits tested show at least one significant β_{covA} (or OR_{covA}) at 5% Benjamini-Hochberg False Discovery Rate (FDR). Affected traits include several biological domains and will be addressed in the last results section. Notably, no trait exhibits a $max(|\beta_{covA}|) > 0.048$, which based on our simulation results would have been a strong indication of pre-admixture selection.

Assessing the robustness of TAGR-covA effects

Although we include multiple covariates in our model, and we show that there is no strong collinearity between them and TAGR-covA (see Supplementary Text), we wanted to further ensure that we are not detecting associations between traits and *covA* due to unaccounted environmental confounding. We therefore aimed at validating our results by using trait and *covA* differences between siblings within UKBB (see Methods for a description of the model). Although siblings might differ in their environment, those differences do not correlate with genetics due to population structure. Hence, this approach is robust to confounding due to overlapping genetic and environmental stratification. 16 out of 32 traits significant in our main analysis were recovered with nominal significance and concordant effect direction in at least one ancestry (Fig. 2A). We observe a general effect concordance (no significant effects in the opposite direction, see also Fig. S9): most of the missing signals can be explained by a lower statistical power.

Next, for 21 traits already explored in the Estonian Biobank (EstBB) [33] we compared the regression statistics obtained across biobanks, finding a remarkable consistency, see Fig. 2B. Even though we refrain from assigning a p-value due to the plausible correlation across several trait pairs, we identify a Pearson's correlation coefficient between biobanks of $\rho = 0.729$, $\rho = 0.542$, and $\rho = 0.726$ respectively for WHG, EEF, and SBA ancestries.

We repeated the analysis by excluding *TAGR-covA*s and instead evaluating the coefficients of *GW-covA*. These were previously used as genomic control covariates, thus allowing us to only keep physiology-related, geographic and socio-economic factors as covariates this time. In this setting we can still see 11 out of 53 significant signals, see Fig. S10, seemingly suggesting that the global similarity with a specific ancestry could be associated with a certain trait. However, the signals are often discordant and have lower absolute effect size on average: for above-mentioned reasons we refrain from conducting a formal test but the trend is appreciable in Fig. 2D. Moreover, the consistency with the sibling analysis and across biobanks is completely absent if we compare *GW-covAs* (Fig. S9 and Fig. 2C, respectively) suggesting that these signals are likely resulting from yet unaccounted confounders. Indeed, although likely harboring variants contributing to the polygenic traits analyzed, the remaining non-TAGR genome was often depleted in trait heritability (see Fig. S8B) but equally subject to confounders impacting at genomic level.

Ancestry-specific diverging traits

We interpret significant β_{covA} as ancestry divergence in genetic value and designate the genetic component showing $\max(|\beta_{covA}|)$ as the most divergent one. Note, however, that in opposition to our simulations, where only one ancestry was divergent by design, the ancestry with $\max(|\beta_{covA}|)$ might not be the only one that experienced a shift in GV. In the following description, we will privilege associations validated in our sibling analysis, but all significant associations in Fig. 2A could in theory be validated with a larger sibling sample size.

Resting heart rate, platelet count and heel bone mineral density, which increase together with WHG ancestry similarity, show the highest absolute β_{covA}. As shown by our simulations, adaptation might not be necessary to observe such divergence. Nevertheless, given health-relevant changes in platelet count and heart rate variability during winter [45, 46], coupled with the complex heart rate reactions to cold exposure [47, 48], it is tempting to speculate about an adaptation to colder climates that WHG suffered when colonizing Europe soon after the last glacial maximum (after 17 kya [49]). Interestingly, platelet count was found to be selected for lower levels since the early Neolithic [50, 51] thus supporting the idea of higher values being WHG-specific and possibly maladaptive in environments tied to non-hunter-gatherer lifestyles.

We replicate known signals of genetic divergence across ancestries for anthropometric and pigmentation traits [33], most notably an association of SBA ancestry with tall stature [26], large waist and waist/hip ratio, but also with darker hair and skin pigmentation, mirrored by opposite effects associated with EEF component. This agrees with ancestry-specific risk scores for these traits estimated by Irving-Pease et al. [29], considering that our SBA component should conflate their Caucasus Hunter-Gatherers, Eastern Hunter-Gatherers and Steppe ancestries. Overall, these results corroborate the hypothesis that pigmentation decrease in Europe is due to post-admixture selection rather than to the impact of incoming SBA migrations during the Bronze Age [27, 50].

For some traits such as caffeine intake, as well as for previously mentioned traits (e.g. waist-derived traits, pigmentation), two ancestries are both significant for opposite trait values. This is expected due to the negative correlation between β_{covA} (Fig. S4) which is given by design but can be further enhanced by divergent evolution of the two ancestries. In some scenarios, only one directional association is replicated by the sibling analysis, resulting from the lower statistical power of this analysis: this is the case for eosinophil percentage, hypertension prevalence and ease

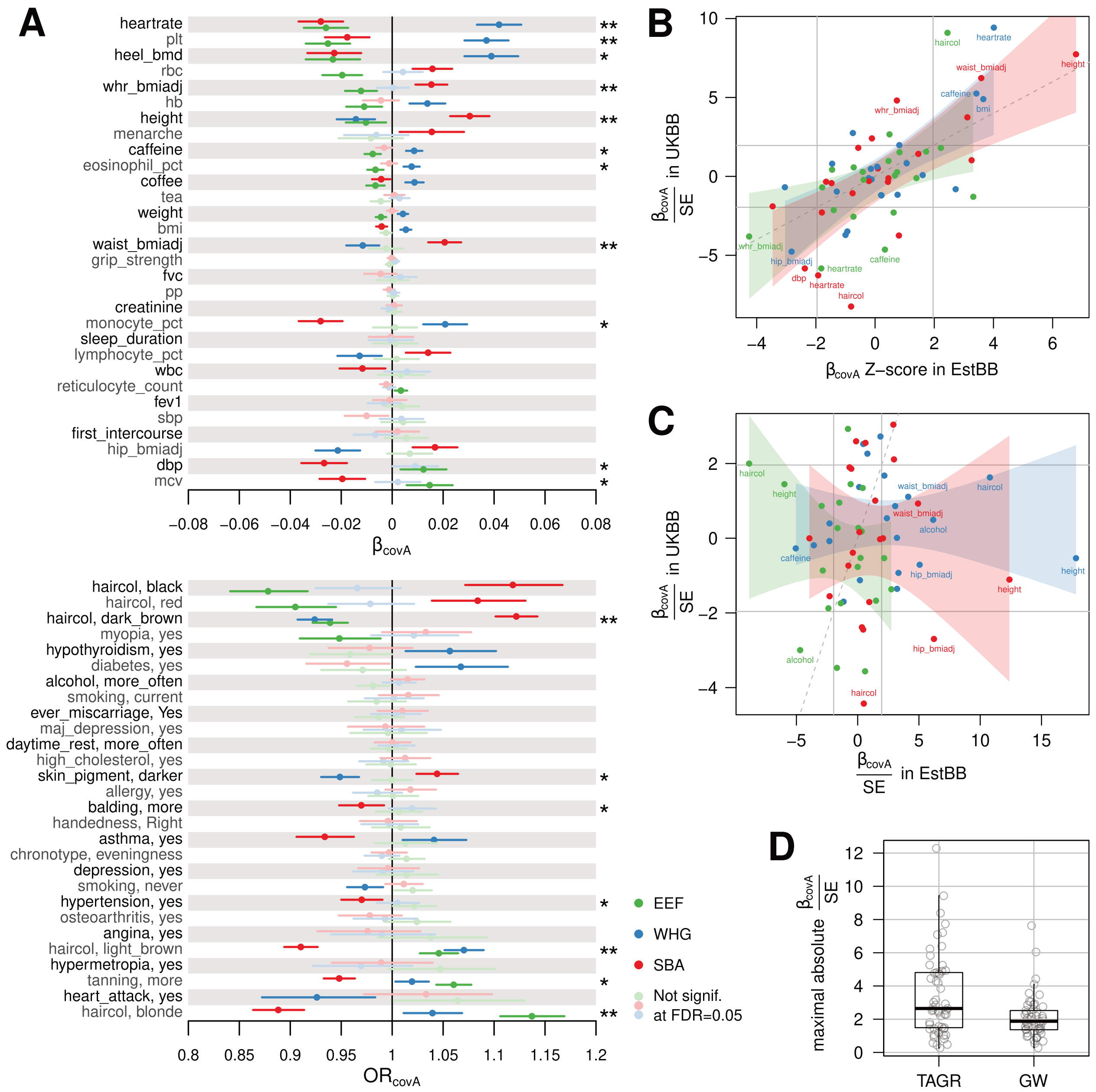

Fig. 2 Trait-ancestry associations in UKBB and across Biobanks. A *CovA* effect size (β_{covA} or OR_{covA}) estimates together with their 95% confidence interval, for each ancestry and for 53 traits analyzed. For each trait, *TAGR-covAs* are used in independent models as trait regressors together with several covariates (including two *GW-covAs*). *TAGR-covAs* not adding significant information according to double-sided coefficient *p*-value at 5% FDR are shown in pastel colors. Trait-ancestry associations significant and concordant in the sibling analysis are marked with one asterisk ($p < 0.05$) or two ($FDR < 0.05$). All β_{covA} represented in this figure, including those estimated in the sibling analysis, are reported in Table S6-7. **B** *TAGR-covA* test statistics compared between UK Biobank and Estonian Biobank. Z-scores in X axis are extracted from the original paper [33], and computed differently than β/SE. Shaded areas represent linear regression slope confidence intervals for each ancestry independently, the solid gray lines mark nominal significance boundaries for β_{covA} while the $y = x$ line is represented as dashed. Traits showing an absolute value > 4 in any of the two axes are labeled **C** Same as **B** but here covA is computed genome-wide. **D** Maximal absolute *TAGR-covA* and *GW-covA* test statistics. Each dot is a trait.

of tanning with EEF remaining significant; and monocyte percentage, mean corpuscular volume, diastolic blood pressure and balding with SBA as the validated component.

DISCUSSION

We applied *covA* to the largest European biobank to identify which complex trait population gradients can be explained by genetic similarity with one of the three main European ancestries: WHG, EEF, and SBA. In our work, we found that relative similarity with a specific ancestral component at trait-associated genomic regions significantly explains a portion of the variance in 32 out of 53 complex traits.

Our findings are globally consistent with an independent analysis conducted on the Estonian Biobank for 21 overlapping traits, thus confirming that these results should be taken as indicative of continental patterns rather than just regional ones. Most of these associations were also confirmed in sibling analysis thus providing evidence against these findings being due to unaccounted environmental confounders conflated with population structure.

Noteworthy, whole-genome results were inconsistent across biobanks, unrelated vs. sibling analysis, and when compared with TAGRs results within the same dataset (Fig. 2C, D, S9); in general whole-genome ancestry similarity was a poor trait predictor. In other words: although we took measures to control for confounders by including multiple geographic and socio-economic variables into our models, only by enriching for biologically-relevant regions, and by controlling for *GW-covA* stratification, we could expose robust and consistent signals. This finding mirrors the known difficulty in discerning functional gene-trait associations from spurious correlations mediated by genetic structure [16–18, 20, 21], which are indeed independent between different populations and different biobanks with different recruitment strategies. In addition, while combining GWAS-derived effect sizes at the genome scale is sensitive to subtle correction-surviving biases [19], our approach does not rely on it. Our results are partially overlapping with an independent investigation with a different approach [29], where a different choice of source ancestral groups and use of GWAS-derived weights for the trait-associated genetic variants may explain emerging differences.

While we are confident about the robustness and the continental scope of the results presented in Fig. 2A, the biological and evolutionary dynamics that can explain these associations remain rather complex to establish. We thus complemented the empirical analysis with extensive simulations to provide insights useful in interpreting our results.

Specifically, the simulations show that significant β_{covA} for at least one of the ancestries can be most directly interpreted as differences in genetic values between sampled reference populations. Although in our simulations such differences can be reached under scenarios with all reference populations having the same optimum, scenarios with optimum shift in one of the ancestries in general result in higher absolute β_{covA} (Fig. 1C) for that ancestry. As the highest β_{covA} values we observe on the real data are close to the 97.5th percentile of the β_{covA} distribution under equal optima in simulations, it is likely that the top of our list contains at least some traits that underwent directional selection in different ancestral populations. Indeed, our results highlight that resting heart rate, blood platelet count and bone mineral density present divergence patterns compatible with adaptive selection. These patterns are comparable to those observed in extensively studied traits like pigmentation and height, warranting a similar level of investigation.

Our simulation results also highlight the potential limitations of our study. These mostly relate to interpreting β_{covA} as an indication of the differential genetic contribution of ancestries to trait variation in the present-day population. On the one hand, such contribution can be overlooked if genetic differentiation between ancestries post-dates the available aDNA samples used as references (Fig. S6E). Furthermore, as most common alleles are expected to pre-date the split between the three ancestries, present-day individuals with higher GV for a given trait might show higher *covA* with an ancestry enriched for trait-increasing alleles, even without any direct genetic contribution from that ancestry. Thus, choosing and exhaustively including all relevant ancestral groups is crucial for a sound biological interpretation of *covA* signals. This limits the applicability of our approach to populations for which the demographic history is less well understood and proper reference aDNA samples are missing.

Finally, two aspects need to be emphasized. First, covA effect sizes were almost never above 0.1, even in the most favorable simulation scenarios: this means that the majority of trait variation should be traced to genetic variation common across ancestral groups, or to environmental effects. Second, the differences in genetic predispositions between ancestries to the tested phenotypes emerge from contemporary phenotypic landscapes and might involve complex interactions with the environment thus weakening claims about actual trait values in the ancestral populations.

In conclusion, the reported ancestry-trait associations are strongly indicative of ancestry-specific genetic predispositions, possibly due to pre-admixture selection, and under certain conditions indicate actual genetic contribution to the contemporary trait landscape. Especially for the most significant associations described above, it is important to stress how the European population has to be assumed as inherently stratified due to its demographic history. This bias will need to be addressed in analyses following up on GWAS effect sizes, in order to avoid potential spurious results. This stratification acts therefore both as a source for variability in such traits and as a potential significant confounder for GWAS study across European cohorts.

DATA AVAILABILITY

UK Biobank data can be accessed through their procedure including project ethical assessment and access fees. Ancient DNA data from Allen Ancient DNA Resource is available at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/FFIDCW. All produced results will be available as supplementary materials or through an online GIT repository.

CODE AVAILABILITY

Simulations: https://github.com/vasilipankratov/SLiM_quantitative_trait. All remaining code will be available as an online GIT repository upon acceptance.

REFERENCES

1. Abdellaoui A, Yengo L, Verweij KJH, Visscher PM. 15 years of GWAS discovery: realizing the promise. Am J Hum Genet. 2023;110:179–94. S0002929722005456.
2. 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;562:203–9.
3. Leitsalu L, Haller T, Esko T, Tammesoo ML, Alavere H, Snieder H, et al. Cohort profile: Estonian biobank of the Estonian Genome Center, University of Tartu. Int J Epidemiol. 2015;44:1137–47.
4. Sijtsma A, Rienks J, Van Der Harst P, Navis G, Rosmalen JGM, Dotinga A. Cohort profile update: lifelines, a three-generation cohort study and biobank. Int J Epidemiol. 2022;51:e295–302.
5. Kurki MI, Karjalainen J, Palta P, Sipilä TP, Kristiansson K, Donner KM, et al. FinnGen provides genetic insights from a well-phenotyped isolated population. Nature 2023;613:508–18.
6. 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.
7. Popejoy AB, Fullerton SM. Genomics is failing on diversity. Nature 2016;538:161–4.
8. Kachuri L, Chatterjee N, Hirbo J, Schaid DJ, Martin I, Kullo IJ, et al. Principles and methods for transferring polygenic risk scores across global populations. Nat Rev Genet. 2024;25:8–25.
9. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38:904–9.
10. Loh PR, Tucker G, Bulik-Sullivan BK, Vilhjálmsson BJ, Finucane HK, Salem RM, et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat Genet. 2015;47:284–90.
11. Zhou W, Nielsen JB, Fritsche LG, Dey R, Gabrielsen ME, Wolford BN, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat Genet. 2018;50:1335–41.
12. Kerminen S, Martin AR, Koskela J, Ruotsalainen SE, Havulinna AS, Surakka I, et al. Geographic variation and bias in the polygenic scores of complex diseases and traits in Finland. Am J Hum Genet. 2019;104:1169–81.
13. Pankratov V, Montinaro F, Kushniarevich A, Hudjashov G, Jay F, Saag L, et al. Differences in local population history at the finest level: the case of the Estonian population. Eur J Hum Genet. 2020;28:1580–91.
14. Cook JP, Mahajan A, Morris AP. Fine-scale population structure in the UK Biobank: implications for genome-wide association studies. Hum Mol Genet. 2020;29:2803–11.
15. Helgason A, Yngvadóttir B, Hrafnkelsson B, Gulcher J, Stefánsson K. An Icelandic example of the impact of population structure on association studies. Nat Genet. 2005;37:90–5.

16. Zaidi AA, Mathieson I. Demographic history mediates the effect of stratification on polygenic scores. eLife 2020;9:1–30.
17. Pärna K, Nolte IM, Snieder H, Fischer K, Estonian Biobank Research Team, Marnetto D, et al. A principal component informed approach to address polygenic risk score transferability across European cohorts. Front Genet. 2022;13:899523.
18. Mostafavi H, Harpak A, Agarwal I, Conley D, Pritchard JK, Przeworski M. Variable prediction accuracy of polygenic scores within an ancestry group. eLife 2020;9:e48376.
19. Barton N, Hermisson J, Nordborg M. Why structure matters. eLife 2019;8:e45380.
20. Berg JJ, Harpak A, Sinnott-Armstrong N, Joergensen AM, Mostafavi H, Field Y, et al. Reduced signal for polygenic adaptation of height in UK Biobank. eLife 2019;8:1–47.
21. Sohail M, Maier RM, Ganna A, Bloemendal A, Martin AR, Turchin MC, et al. Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. eLife 2019;8:1–17.
22. Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature 2014;513:409–13.
23. Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, et al. Massive migration from the steppe was a source for Indo-European languages in Europe. Nature 2015;522(Jun):207–11.
24. Allentoft ME, Sikora M, Sjögren KG, Rasmussen S, Rasmussen M, Stenderup J, et al. Population genomics of bronze age Eurasia. Nature 2015;522:167–72.
25. Marchi N, Winkelbach L, Schulz I, Brami M, Hofmanová Z, Blöcher J, et al. The genomic origins of the world's first farmers. Cell 2022;185:1842–59.e18.
26. Cox SL, Ruff CB, Maier RM, Mathieson I. Genetic contributions to variation in human stature in prehistoric Europe. Proc Natl Acad Sci USA. 2019;116:21484–92.
27. Ju D, Mathieson I. The evolution of skin pigmentation-associated variation in West Eurasia. Proc Natl Acad Sci. 2021;118:e2009227118.
28. Berens AJ, Cooper TL, Lachance J. The genomic health of ancient hominins. Hum Biol. 2017;89:7.
29. Irving-Pease EK, Refoyo-Martínez A, Ingason A, Pearson A, Fischer A, Barrie W, et al. The selection landscape and genetic legacy of ancient Eurasians. Evol Biol; 2022. Available from: https://doi.org/10.1101/2022.09.22.509027.
30. Olalde I, Allentoft ME, Sánchez-Quinto F, Santpere G, Chiang CWK, DeGiorgio M, et al. Derived immune and ancestral pigmentation alleles in a 7000-year-old Mesolithic European. Nature. 2014;507:225–8.
31. Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, et al. Genome-wide patterns of selection in 230 ancient Eurasians. Nature. 2015;528:499–503.
32. Saag L, Vasilyev SV, Varul L, Kosorukova NV, Gerasimov DV, Oshibkina SV, et al. Genetic ancestry changes in Stone to Bronze age transition in the East European plain. Sci Adv. 2021;7:eabd6535.
33. Marnetto D, Pankratov V, Mondal M, Montinaro F, Pärna K, Vallini L, et al. Ancestral genomic contributions to complex traits in contemporary Europeans. Curr Biol. 2022;32:1412–9.e3.
34. Haller BC, Messer PW. SLiM 4: multispecies eco-evolutionary modeling. Am Nat. 2023;201:E127–39.
35. 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;618:774–81.
36. Sakaue S, Hirata J, Kanai M, Suzuki K, Akiyama M, Lai Too C, et al. Dimensionality reduction reveals fine-scale structure in the Japanese population with consequences for polygenic risk prediction. Nat Commun. 2020;11:1569.
37. Baumdicker F, Bisschop G, Goldstein D, Gower G, Ragsdale AP, Tsambos G, et al. Efficient ancestry and mutation simulation with msprime 1.0. Browning S, editor. Genetics. 2022;220:iyab229.
38. Scally A, Durbin R. Revising the human mutation rate: implications for understanding human evolution. Nat Rev Genet. 2012;13:745–53.
39. Yair S, Coop G. Population differentiation of polygenic score predictions under stabilizing selection. Philos Trans R Soc B Biol Sci. 2022;377:20200416.
40. Mallick S, Micco A, Mah M, Ringbauer H, Lazaridis I, Olalde I, et al. The Allen Ancient DNA Resource (AADR): a curated compendium of ancient human genomes. Genomics; 2023. Available from: https://doi.org/10.1101/2023.04.06.535797.
41. 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:D977–85.
42. The 1000 Genomes Project Consortium, Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, et al. A global reference for human genetic variation. Nature 2015;526:68–74.
43. The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. Nature. 2010;467:52–8.
44. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh PR, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet. 2015;47:1228–35.
45. Kristal-Boneh E, Froom P, Harari G, Malik M, Ribak J.Summer-winter differences in 24 h variability of heart rate.Eur J Cardiovasc Risk. 2000;7:141–6.
46. De Lorenzo F. Cold adaptation and the seasonal distribution of acute myocardial infarction. QJM 1999;92:747–51.
47. Mourot L, Bouhaddi M, Regnard J. Effects of the cold pressor test on cardiac autonomic control in normal subjects. Physiol Res. 2009;58:83–91.
48. LeBlanc J, Dulac S, Cote J, Girard B. Autonomic nervous system and adaptation to cold in man. J Appl Physiol. 1975;39:181–6.
49. Bortolini E, Pagani L, Oxilia G, Posth C, Fontana F, Badino F, et al. Early Alpine occupation backdates westward human migration in Late Glacial Europe. Curr Biol. 2021;31:2484–93.e7.
50. Le MK, Smith OS, Akbari A, Harpak A, Reich D, Narasimhan VM. 1000 ancient genomes uncover 10,000 years of natural selection in Europe. Genomics; 2022. Available from: https://doi.org/10.1101/2022.08.24.505188.
51. Kerner G, Neehus AL, Philippot Q, Bohlen J, Rinchai D, Kerrouche N, et al. Genetic adaptation to pathogens and increased risk of inflammatory disorders in post-Neolithic Europe. Cell Genom. 2023;3:100248.

ACKNOWLEDGEMENTS

We would like to thank Francesco Montinaro and Alessandro Raveane for insightful discussions. Data analyses for this study were carried out in the High-Performance Computing Center of the University of Tartu. D.F. is a PhD student enrolled in the National PhD in Artificial Intelligence, XXXVIII cycle, course on Health and life sciences, organized by Università Campus Bio-Medico di Roma.

AUTHOR CONTRIBUTIONS

V.P., L.P. and D.M. conceived and designed the study; V.P. and D.M. developed the computational pipelines and performed the bulk of statistical analyses; M.Mez. contributed to demographic simulations; M.Mez., S.A., I.K. and D.F. contributed to data analysis; J.F.W., M.Met., P.P., L.P. provided data and contributed to results interpretation; V.P. and D.M. drafted the manuscript; all authors reviewed and approved the submitted paper.

FUNDING

D.M. was supported by the University of Turin through Grant For Internationalization 2022. V.P., I.K. and M.Met. were supported by the European Union through Horizon 2020 Research and Innovation Program under Grant No. 810645 and the European Union through the European Regional Development Fund Project No. MOBEC008. M.Met. was supported by the Estonian Research Council grant PUT (PRG1899).

COMPETING INTERESTS

The authors declare no competing interests.

ETHICAL APPROVAL

This research has been conducted using the UK Biobank Resource under Application Number 19655. Researchers do not require separate ethical clearance and can operate under the UK Biobank approval.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41431-024-01678-9.

Correspondence and requests for materials should be addressed to Vasili Pankratov or Davide Marnetto.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.