Ancestral genomic contributions to complex traits in contemporary Europeans

Davide Marnetto[1,2*], Vasili Pankratov[1], Mayukh Mondal[1], Francesco Montinaro[1,3], Katri Pärna[1,4], Leonardo Vallini[5], Ludovica Molinaro[1], Lehti Saag[1,6], Liisa Loog[7], Sara Montagnese[8], Rodolfo Costa[5,9,10], Estonian Biobank Research Team[1], Mait Metspalu[1], Anders Eriksson[1], Luca Pagani[1,5*]

November 11, 2021

[1] Institute of Genomics, University of Tartu, Tartu, 51010, Estonia.
[2] Department of Neurosciences 'Rita Levi Montalcini', University of Turin, Torino, 10126, Italy.
[3] Department of Biology, University of Bari, Bari, 70125, Italy
[4] Department of Epidemiology, University of Groningen, Groningen, 9700 RB, The Netherlands.
[5] Department of Biology, University of Padova, Padova, 35131, Italy.
[6] Research Department of Genetics, Evolution and Environment, University College London, London, WC1E 6BT, UK.
[7] Department of Genetics, University of Cambridge, Cambridge, CB2 3EH, UK.
[8] Department of Medicine, University of Padova, Padova, 35128, Italy.
[9] Institute of Neurosciences, National Research Council (CNR), Padova, 35121, Italy.
[10] Faculty of Health and Medical Sciences, University of Surrey, Guildford, GU2 7YH, UK.
∗ Correspondence: davide.marnetto@unito.it (DM); lp.lucapagani@gmail.com (LP)

Summary

The contemporary European genetic makeup was formed in the last 8000 years when local Western Hunter-Gatherers mixed with incoming Anatolian Neolithic farmers and Pontic Steppe pastoralists[1–3]. This encounter combined genetic variants with distinct evolutionary histories and, together with new environmental challenges faced by the post-Neolithic European farmers, unlocked novel human adaptations[4].

Previous research efforts have inferred phenotypes in these source populations, using either a few

single loci[5–7] or polygenic scores based on genome-wide association studies[8–10], and investigated the strength and timing of natural selection on traits such as lactase persistence or standing height[6,11,12]. However, how ancient populations contributed to present-day phenotypic variation is poorly understood.

Here we investigate how the unique tiling of genetic variants inherited from different ancestral components drives the complex traits landscape of contemporary Europeans, and quantify selection patterns associated with these components. Using matching individual-level genotype and phenotype data for 27 traits in the Estonian biobank[13] and genotype data directly from the ancient source populations, we quantify the contributions from each ancestry to present-day phenotypic variation in each complex trait.

We find substantial differences in ancestry for eye and hair colour, body mass index, waist/hip circumferences and their ratio, height, cholesterol levels, caffeine intake, heart rate and age at menarche. Furthermore, we find evidence for recent positive selection linked to four of these traits and, in addition, sleep patterns and blood pressure.

Our results show that these ancient components were differentiated enough to contribute ancestry-specific signatures to the complex trait variability displayed by contemporary Europeans.

Keywords

Human Genetics; Population Genetics; Complex Traits; Ancestry; Biobank; Europe; Estonia

Results and Discussion

We identified 27 complex traits of interest, based on information availability in the Estonian Biobank[13] (EstBB) and GWAS catalog[14]. EstBB contains matching genotype and phenotype information for individuals from a relatively homogeneous population, that contains all three ancestry components found in Europe, with the proportion of remnant Hunter Gatherer ancestry among the highest in Europe and an additional minor ($< 5\%$) Siberian component associated with Iron Age movements[15,16]. In order to associate specific ancient European ancestry components with predicted phenotypes, we introduce *covA*, a measure of the relative similarity between any contemporary individual and the ancestries that contributed to its genetic makeup. For each sample in the EstBB we computed its *covA* with each of the ancestral source populations, focusing on genomic regions potentially connected to each trait. We then used *covA* as a predictor to model traits, also in comparison with the same statistic computed for the whole genome. Finally we test if those regions associated with the genetic contribution from a specific ancestry experienced a post-admixture selective pressure on top of the observed local unbalance in contributing ancestries.

covA measures similarity with ancestral groups

Here we introduce *covA*, the covariance between allele frequency (p) in a contemporary individual i (i.e. its allele dosage) and a given ancestral population j with respect to the contemporary and ancient average frequencies (p_C and p_A respectively):

$$covA(i,j) = (p_i - p_C)(p_j - p_A) \tag{1}$$

The *covA* statistic is expected to be high when the allele frequencies of the individual i and the ancestry j are similar in comparison with the differences within the contemporary population and across the ancestries that contributed to its genetic makeup. Furthermore, *covA* can be computed averaging over the contribution of multiple Single Nucleotide Polymorphisms (SNPs), either across the whole genome or for specific regions of interest.

In order to test the potential of *covA* to distinguish between genetic contributions from different ancestries, we simulated polygenic traits in a modern population composed of three ancestral groups and verified that when predicting simulated traits, *covA* estimated coefficient correlates well with their ancestral specificity (Pearson's correlation coefficient ρ=0.919-0.937, Figure S1a). See Methods, Supplementary Notes and Figure S7 for further discussion of *covA* properties and simulation details, including the definition of ancestral specificity.

For each EstBB individual, we computed genome-wide *covA* between the individual and each of the ancestries among Western Hunter-Gatherers (WHG), Neolithic Farmers from Anatolia (Anatolia_N), Yamnaya Pastoralists from the Pontic Steppes (Yamnaya), and Siberian (Siberia). We defined these ancestry groups based on genetic and chronological proximity to a set of identified focal individuals, see Methods and Table S1 for a list of the ancient genomes assigned to each group. As expected, *covA*s calculated on the different ancestries are strongly interdependent, because they include as term the average ancestral frequency (p_A) and because of varying grades of similarity among the ancestries for historical demographic reasons (see *covA* joint distributions in Figure S2). In particular *covA* tends to be negatively correlated between different ancestral components with the exception of Yamnaya and WHG, reflecting complex demographic relationships between the two, involving WHG-like Eastern Hunter Gatherer ancestry presence in Yamnaya[2,3,17]. Furthermore, although the Estonian population is considered relatively genetically uniform, some geographic differences exist with the south-eastern inland counties having higher haplotype sharing with Latvians, Lithuanians and Russians compared with the rest of the country, as recently shown in Pankratov *et al.* [18]. This result is also mirrored in our analyses with median *covA* for WHG being higher in south-eastern inland counties, see Figure S3a. Conversely, as shown by median *covA* for the Siberian component in Figure S3d, the Siberian ancestry seems to be more abundant in north-east Estonia, consistently with Finnish ancestry shown by Pankratov and colleagues[18]. Yamnaya and Anatolia_N *covA*s are instead more evenly distributed (Figure S3b,c).

Phenotype-associated genomic regions show specific ancestry similarity patterns

We examined 27 complex traits (31 if considering separate classes of pigmentation) for which we had sufficient records in the Estonian Biobank (see Table S2). We corrected and adjusted them for confounding covariates, including sex, age, genotyping platform and others as specified in Table S2. As our analysis relies on SNPs overlapping between ancient and contemporary genetic data, a portion of the genetic influence over these traits, especially when conveyed by rare alleles, might elude our experimental setting[19]. Nevertheless, our data set captures a genetic basis for most of them, as confirmed by the trait heritability measured in our sample (Figure 1).

We defined three sets of candidate regions for any given trait by considering windows of 5kb, 50kb or 500kb centered around GWAS catalog[14] hits for corresponding trait categories (see Methods and Table S3). As shown in Figure S4, these genomic regions harbor a higher heritability intensity (h^2/Mb) than the whole genome, supporting their suitability as candidate regions for the traits of interest.

Next, we used *covA*s computed on the candidate regions as a predictor to model traits, and asked whether they showed significantly different regression coefficients when compared to 50 size-matching random genomic sets: this was found true in 11 out of 27 traits (double-sided Z-test, Benjamini-Hochberg FDR = 0.05), see Z-scores in Figure 2. This analysis has the advantage of automatically controlling for virtually all potential confounders that apply to the genome in its entirety, e.g. social, economic and cultural statuses, thus allowing us to not include any such covariates in the model. In addition, this analysis pinpoints genetic signals that are likely to be functionally connected to the trait. Among others, blood cholesterol levels are shown to be positively correlated with similarity to Yamnaya in cholesterol-associated regions with respect to the rest of the genome, while the opposite is true for WHG.

Since *covA* exhibits a high correlation across ancestries, we avoided implementing a model with largely multicollinear predictors including *covA* for all ancestries and instead adopted separate models for each ancestry, complementing them with a regression on *covA* Independent Components (ICs) (Figure 2b). We used the loadings from a Principal Component (PC) Analysis on whole genome *covA*s (Figure 2c) to transform region-specific *covA*s into ICs. This, though not returning actual PCs in each candidate region, drastically reduces the collinearity (highest Variance Inflation Factor=1.62 in hair color 50kb candidate regions), while allowing simpler interpretation and, crucially, cross-region comparisons required for Z-scores computation. While *covA*s (Figure 2a) highlight the overall excess or lack of certain ancestries in relation with a given phenotype but are largely intertwined, ICs (Figure 2b) can be interpreted as independent axes defined by 2 or 3 *covA*s. We therefore adopted ICs to discriminate significant ancestry-trait associations, as they are independent variables in a comprehensive predictive model. Significant results, interpreted in light of the ICs, are summarized in Table S4 and discussed below. Among others, this analysis confirms the association between cholesterol levels and the Yamnaya-WHG axis previously mentioned.

Comparison with genome-wide ancestry similarity

We followed up the association between phenotypes and local excess or lack of a given ancestry and explored whether a similar pattern held at whole genome level by computing genome-wide *covA*s. Here, being unable to correct for environmental confounders with a Z-score approach and avoiding genotype-based PCs as covariates in order not to hinder potential genome-wide signals, we run the risk of obtaining spurious ancestry-trait associations. This is due to uneven ancestry similarity across Estonia concurrent with geographically associated socio-economic differences that can potentially confound genotype-phenotype associations. Although the confounding effect of population structure is minimised by the inclusion of a relatively uniform population, small differences related to historical reasons[18] are still visible in *covA* (see Figure S3). Therefore, we include a city/countryside residency covariate in the models, defined as 1 for people living in the wealthiest and most populous county (Harju county) and 0 otherwise, and a covariate for educational attainment, which is a good proxy for family socioeconomic status[20,21]. This control allows us to suggest a significant influence of genome-wide ancestry on 16 traits out of 27, as shown in Figure S5, even when geographical and social stratification is present (coefficient p value significant at Benjamini-Hochberg FDR=0.05). Again, *covA*-based PCs were used to interpret significant results.

Interestingly, we do not always observe concordance between the region-specific and genome-wide results, as shown in Figure 3, pointing to the fact that region specific trends are not always sufficient to drive genome-wide signals to significance, or might even arise in a contrasting genomic background. This is especially true for less polygenic traits (e.g. pigmentation), but also for more polygenic ones, as indicated by height association with WHG. On the other hand, we also find genome-wide ancestry-trait connections which are not exacerbated in candidate regions, thus losing Z-score significance. This can occur for a single ancestry (e.g. Anatolia_N or Siberia and height) or cause the loss of trait associations altogether, as for alcohol consumption, depression, sleep duration, social jetlag, diopters, pulse pressure, creatinine levels. Finally, we observed that genome-wide *covA*s for WHG and Yamnaya tend to be linked to most phenotypes in a similar fashion, in contrast with results found in candidate regions where the two ancestries behave in a more independent manner (Figure S5).

Selection signatures at candidate regions with ancestry-trait association

So far we only explored associations between a given trait and a local or genome-wide excess of a given ancestry. The observed local admixture unbalance points to a role of that ancient contribution in explaining a given phenotype. However, these results alone do not show whether after the admixture event the incoming genetic material also underwent a selective sweep within the recipient population, altering population-wide allele frequencies as investigated in Mathieson *et al.* [6]. In other words, the local admixture imbalances we detected so far are not necessarily transferred to the whole population.

We independently asked whether the phenotype-associated regions above also exhibit signs of recent natural selection. We applied CLUES[22] to the list of GWAS hits used as index for our

candidate regions to obtain per-SNP evidence of recent (up to 500 generations ago) natural selection, and to see which phenotypes show enrichment in SNPs with strong selection signals compared to a random set of GWAS hits. Out of the genomic regions responsible for ancestry-trait association shown in Figure 2, pigmentation-related SNPs (eye and hair color) showed extremely high CLUES logLR values (Figures 4a, S6) in accordance with previous results[6,9,23], as well as SNPs related to BMI and cholesterol, pointing to ongoing or recent selection at these loci. Diastolic blood pressure (DBP) and sleep-related SNPs also showed the same extreme signature, but the candidate regions encompassing them did not reach significance in ancestry-trait association.

The recent and putatively ongoing nature of the inferred selective pressure on the six traits shown in Figure 4a is further exemplified by the steep increase in derived allele frequencies over time inferred for the top 3 SNPs of each trait and shown in Figure 4b. These include some loci previously shown to be selected in West Eurasians (rs4988235 at MCM6/LCT[24], pigmentation-related SNPs at HERC2/OCA2, TYRP1, TYR, TPCN2[9,23,25], rs653178 at ATXN2[26]) and some other, yet to be explored. In particular rs17630235, associated with BMI and DBP, is an expression Quantitative Trait Locus (QTL) in several epithelial tissues[27] for ALDH2, an aldehyde dehydrogenase known for its role in the alcohol metabolism[28]. Although this selective signal might be due to rs17630235 proximity with ATXN2, it is tempting to speculate about the changed role of ALDH2 in a post-neolithic society, which made available several fermentable substrates. Other selected SNPs include rs74555583 and rs11539148, both associated with sleep patterns (chronotype). Most notably, the latter is a missense variant in the catalytic domain of QARS1, for which also functions as splicing QTL[27]. QARS1 itself encodes an enzyme involved in the glutaminyl-tRNA synthesis and, when mutated, leads to microcephaly, cerebral-cerebellar atrophy and seizures[29].

Discussion

Here we combined existing knowledge on genotype-phenotype associations and the availability of ancient genomes to assess the impact of ancient migrations on the phenotypic landscape of contemporary Europeans. We leveraged on traits measured in living individuals, complementing previous works where phenotypes were inferred for ancient genomes instead. As a whole, the most affected traits include pigmentation and anthropometric traits together with blood cholesterol levels, caffeine consumption, heart rate and age at menarche.

Importantly, while our genome-wide results highlight an overall excess of an ancestry in the carriers of a given phenotype, this is not necessarily mirrored at the genetic loci for which the genotype-phenotype association is ascertained in the literature. A genome-wide excess can completely explain a regional signal, leading to non-significant Z-scores, and even indicate a different direction for the same ancestry. While the first scenario can be due to the extreme polygenicity of a trait, possibly coupled with an inaccurate tagging of the actual functional regions by the GWAS catalog hits, the second might indicate an incomplete correction of non-genetic factors in the genome-wide analysis. Indeed, it is possible that place of residence and educational attainment alone cannot fully account for confounding environmental effects such as

socioeconomic status. Conversely, candidate region Z-scores are disentangled from background confounders, and virtually free from collinearity issues when they also agree with the relevant ICs. In this light, we here chose to report and discuss results showing region-specific significance for *covA*s and matching ICs (as reported in Table S4), hence refraining from making inferences on traits such as eye pigmentation in Yamnaya, among others.

WHG ancestry in present day individuals is linked to lower cholesterol levels, higher BMI and putatively contributed brown hair and light eye color to the contemporary Estonian population. This last association has been previously described based on the HERC/OCA2 haplotypes found in ancient WHG samples[5,23]. In addition, loci associated with these features also appear to have undergone selection in Estonians. Other region-specific associations for this ancestry include decreased hip circumference, and increased caffeine consumption and heart rate.

An enriched Yamnaya ancestry is linked to a strong build, with tall stature (in agreement with previous studies[6,8]) and increased hip and waist circumferences, both at genome-wide and region-specific levels, but also to black hairs and high cholesterol concentrations when focusing on candidate regions. The associations of Yamnaya and WHG ancestries to respectively higher and lower cholesterol levels, together with the observed signatures of selection at loci connected to cholesterol and BMI, add a new component to our understanding of post-neolithic dietary adaptation[7,30,31] with potential implications to disease risk and outcomes in present-day populations.

Anatolia_N enrichment in trait-related genomic regions is connected with a reduced BMI-corrected waist to hip ratio, reduced BMI, light (but not green) eyes and fair hair, increased age at menarche and reduced heart rate. Notably, *covA(i,Anatolia_N)* has a substantial weight only in IC2, the single IC that reaches significance when predicting heart rate, suggesting a prominent role for this ancestry in determining this trait.

Finally, the Siberian ancestry is connected with dark hair pigmentation, higher heart rate, lower caffeine consumption and most prominently green eye color and lower age at menarche. Importantly, while the results connected to the Siberian ancestry are not of broad applicability to all European populations, *covA(i,Siberia)* and relative ICs received effect-sizes with mixed significance in all the previous traits except for age at menarche and pigmentation, suggesting that other ancestries might have a larger impact. In other words, we do not find other phenotypes that can be best explained by similarity with Siberia, implying that the presence of this ancestry in the Estonian genome does not significantly affect the inference based on the other, pan-European ancient components.

A general caveat about significance levels observed in this study is that as we refrain from reducing interdependent traits by arbitrary choices, even testing multiple alternatives of the same trait, we expose ourselves to inflated false negatives. We deemed it best to acknowledge and control this risk by avoiding overly stringent multiple testing corrections as Bonferroni, and adopting the Benjamini-Hochberg procedure to control FDR instead. In addition, as highly significant traits tend to have higher heritability, it is likely that our analysis might not have enough statistical power for poorly heritable traits. Nevertheless, as we are able to highlight ancestry-trait associations for caffeine consumption ($h^2 = 0.087 \pm 0.009$), brown hair color

($h^2 = 0.079 \pm 0.009$) and even heart rate ($h^2 = 0.044 \pm 0.009$), this condition should be limited only to the very few traits exhibiting lower heritabilities.

Importantly, our inferences are applicable to contemporary individuals of European ancestry, where the phenotypes were collected. Conversely, using them to extrapolate features of ancient populations, although tempting, should be done with caution due to the interaction of their genetic legacy with a radically different lifestyle and environment. Furthermore, when seeking a biological interpretation of our results, it should be kept in mind that certain narrowly defined, contemporary phenotypes such as caffeine consumption may point to broader biological pathways.

Taken together, our results show that the ancient components that form the contemporary European landscape were differentiated enough at a functional level to contribute ancestry-specific signatures on the phenotypic variability displayed by contemporary individuals, regardless of which target population one may examine. In particular, when looking at Estonians, for 11 out of 27 traits surveyed here we could confirm a significant relationship between presence of a given ancestry in genetic regions associated with a given phenotype and how this is expressed by contemporary individuals. While showing that both autochthonous (WHG) and incoming groups contributed genetic material that shapes the phenotype landscape observed today, we also demonstrated that a subset of these loci further underwent positive selection in the last 500 generations. Although not determining whether the selected alleles (and phenotypes) were predominantly contributed by the autochthonous or incoming groups, by connecting genotypic ancestry and complex traits measured in a large dataset, our results reveal both neutral and adaptive consequences of the post-neolithic admixture events on the European phenotype landscape.

Acknowledgements

This work is supported by the European Union through the European Regional Development Fund, project No. 2014-2020.4.01.16-0024, MOBTT53 (DM, KP, LM, LP); MOBEC008 (VP, MMo, MMe, AE); 2014-2020.4.01.16-0030 (FM, MMe); 2014-2020.4.01.15-0012 (MMe); through the Horizon 2020 research and innovation programme grant no. 810645 (VP, MMo, MMe, AE) and through the Horizon 2020 MSCA Initial Training Network, grant no. 765937 (RC). LS, MMe are supported by the Estonian Research Council through PUT PRG243. SM is supported by the STARS@UNIPD 2019 Consolidator Grant for the project CircadianCare. We would like to thank Paolo Provero for helpful discussions. Most of the analysis was run on the High-Performance Computing Center at the University of Tartu. The Estonian Biobank Research Team includes Mari Nelis, Lili Milani, Tõnu Esko, Andres Metspalu, Reedik Mägi.

Author Contributions

DM, LP conceived and designed the study; AE contributed in the statistical design; DM, VP performed data analyses; MMo, FM, KP, LV, LM, LP contributed to data analyses; SM, RC provided analyses and expertise about sleep traits; FM, LS, LL, MMe contributed with ancient genetics expertise; DM, LP drafted the manuscript; all authors reviewed and approved the submitted paper.

Declaration of interests

The authors declare no competing interests.

Figure legends

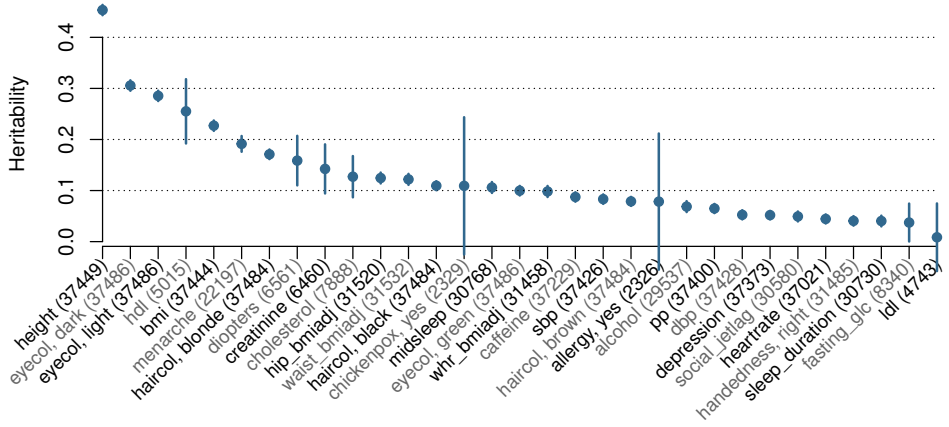

Figure 1: Traits and their heritability. All traits analyzed and their estimated heritability after covariate adjustment. Bars indicate standard errors of the estimate. Numbers in parentheses indicate the number of unrelated samples for which phenotypic information was available for each trait.

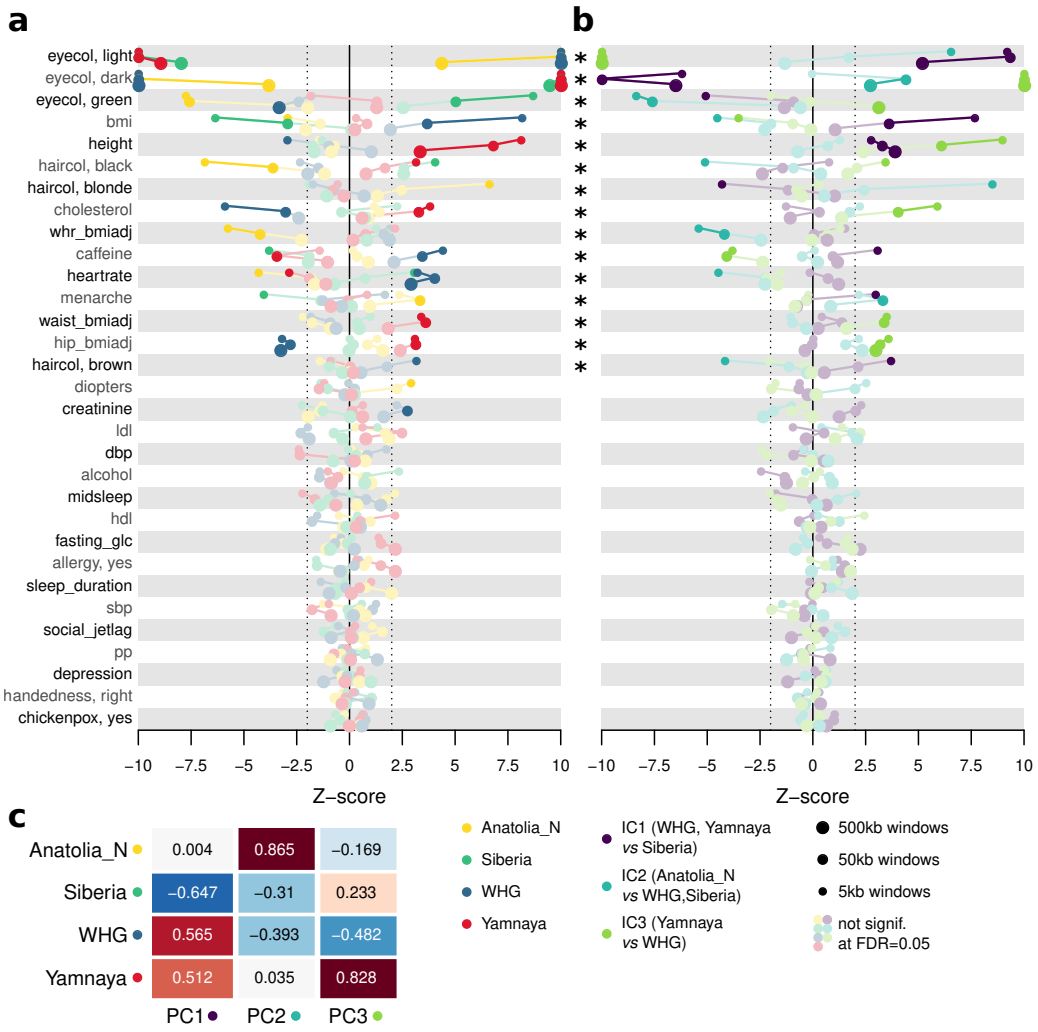

Figure 2: ancestry-trait association on candidate regions. **a** Z-scores of *covA* coefficients, the color refers to the ancestry tested. **b** Z-scores of coefficients associated with *covA* independent components (IC) computed with whole genome-based *covA* PC loadings. Each color is associated with one of the three ICs. For each trait we show the Z-score of the standardized coefficient associated with candidate regions against a distribution of 50 random genomic regions of matching size. Candidate regions are determined around GWAS hits for appropriate traits as windows with three different widths: 5 (small dot), 50 (medium dot) and 500 (large dot) kilobases. Pastel dots are deemed not significant at Benjamini-Hochberg FDR = 0.05, p value from double-sided Z-test; asterisks mark traits to be considered significant according to **b**; dotted lines correspond to absolute Z-scores = 2. **c** Loading matrix for genome wide *covA*s and their PCs, used to transform *covA*s into their ICs. The three genome wide PCs accounted for 0.498, 0.327 and 0.175 *covA*s variance, respectively. PCs and relative ICs can be interpreted as axes defined by 2 or 3 *covA*s.

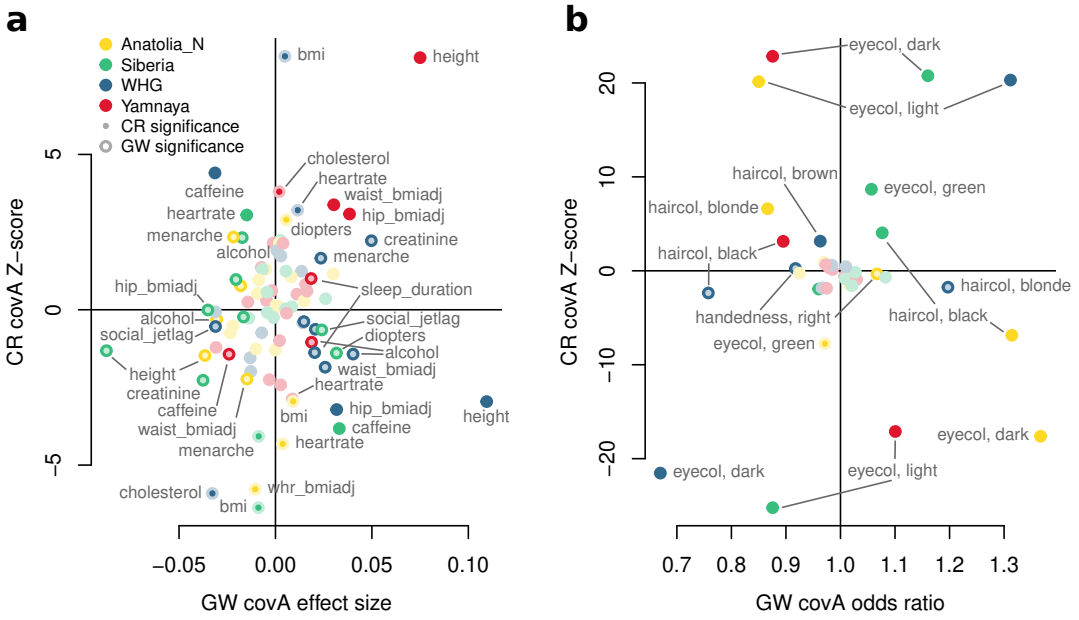

Figure 3: Comparison between Genome-wide and region-specific ancestry-trait associations. The Y-axis represent Z-scores of *covA* coefficients, for *covA* computed on candidate regions (CR) of 5 kilobases as in Figure 2. X-axes represent genome-wide (GW) *covA* estimated coefficients: we report *beta* effect sizes for continuous traits in **a** and Odds Ratios for categorical traits in **b**. Independent models are run for different *covA*s. Colors label the ancestry tested, while inner and outer color intensity represents significance of CR *covA* Z-score and GW *covA* coefficients, respectively. Pastel colors indicate not significant results at Benjamini-Hochberg FDR = 0.05 (double-sided Z-test p value for CR *covA* Z-score or double-sided coefficient p value for GW *covA* coefficients). Labels indicate selected outlying ancestry-trait associations.

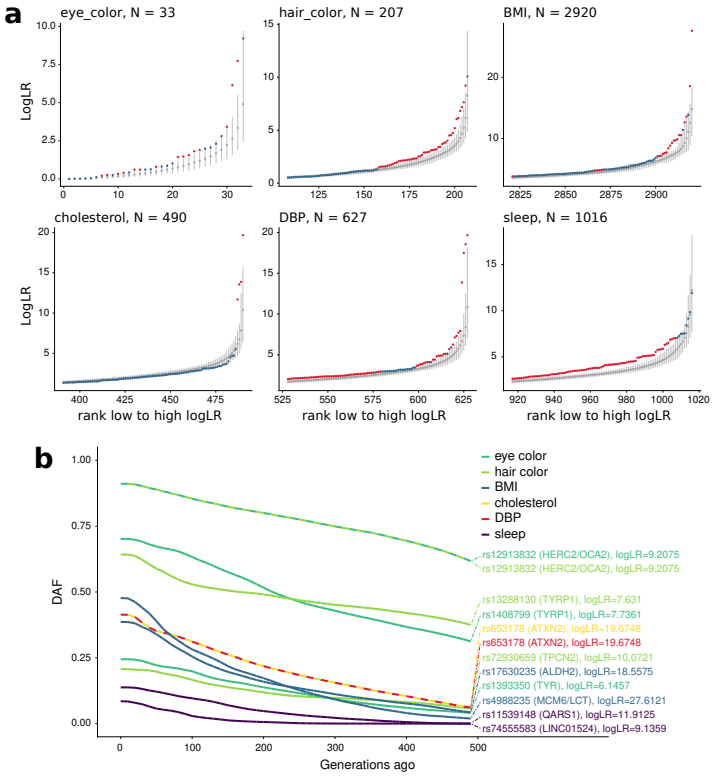

Figure 4: Selection signatures. **a** CLUES log likelihood ratios (logLR) values distribution for GWAS hits for six selected phenotypes. For each phenotype at most 100 top SNPs with highest logLR values and the corresponding ranks from the random GWAS hits distribution are shown. Grey dots show mean values for each rank in the background distribution while the whiskers show the 5-95 percentile range. The logLR values for tested SNPs are shown in red or blue depending on whether the value lies above the 95th percentile of the values from the background distribution with a given rank. Number of tested SNPs for each phenotype are shown in panel titles. Sleep indicates SNPs connected to all sleep traits as indicated in table S3. **b** Maximum likelihood estimates of derived allele frequency trajectories for top 3 SNPs with highest logLR values for each phenotype. When more than one SNPs come from the same locus, only the top-scoring SNP is shown.

STAR Methods

0.1 Resource Availability

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Davide Marnetto (davide.marnetto@unito.it).

Data and code availability

This paper analyzes existing, publicly available data. These accession numbers for the datasets are listed in the key resources table. Data from Estonian Biobank are under managed access and subject to approval of the Estonian Committee on Bioethics and Human Research; accessed with Approval Number 285/T-13 obtained on 17/09/2018 by the University of Tartu Ethics Committee.

All original code has been deposited at `bitbucket_repository_url_upon_publication` and is publicly available as of the date of publication.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

0.2 Method details

Sample selection and ancient European grouping

We used 50,353 sequenced or genotyped individuals from the Estonian Biobank[13] as contemporary Estonian sampleset. After removing second-degree relatives (pi-hat > 0.25) we obtained a subset of 37,952 individuals and used it as a scaffold to perform a PC Analysis (PCA) with Eigensoft-6.1.4. Other individuals were projected on the same PCA space. Outliers identified in this process (with parameters `numoutlieriter: 5` `numoutlierevec: 10` `outliersigmathreshold: 6`) were discarded. Samples that on the first round of genome-wide *covA*s were more distant than 8 Interquartile Ranges (IQR) from the upper or lower quartile against any of the ancestries were also discarded, resulting in 49811 individuals included in our sample set. For each trait of interest we first removed individuals with missing data for traits and covariates and subsequently discarded second-degree relatives.

To define ancestral European groups we started from the Allen Ancient DNA Resource (AADR)

V44.3 merged with present-day individuals typed on the Human Origins array (see Data Availability section). From this set we defined a manually curated core set for each ancestral group, then performed a PCA on a space defined by modern Eurasian and North African individuals west of Iran (included), where the ancient samples were projected. We expanded these core sets to other individuals from AADR dataset using multi-dimensional ellypses with diameters equal to 3 core set SDs. We used 4 dimensions: the annotated dating and the first 3 PCs generated above. With this process we selected 90 WHG, 92 Anatolia_N, 74 Yamnaya S1. In addition, from the ones available from the same dataset, we took 7 samples as representative of the broader Siberian ancestry, assuming any Siberian individual would be equidistant to the other ancestral European groups: S_Even-3.DG, S_Even-1.DG, S_Even-2.DG, Bur1.SG, Bur2.SG, Kor1.SG and Kor2.SG. 957,869 SNPs remained in our dataset after merging the contemporary and ancient sets.

Phenotypes treatment and heritability

Continuous traits were treated as specified in Table S2 and regressed against the covariates according to the same table. Individuals with traits or covariates more distant than 4 IQRs from the upper or lower quartile were considered as outliers and discarded. After adjusting traits as described, their heritability was computed using LDAK 5.0[32]. First we computed a kinship matrix with the LDAK-Thin Model: we thinned down SNPs on the non-related sample set defined above with parameters `--window-prune .98 --window-kb 100`, then used `--calc-kins-direct` with the resulting weights and `--power .25`. Finally we estimated heritability using REML solver.

covA definition

covA is the covariance in allele frequency (p) within a contemporary individual i (i.e. its allele dosage) with the ancestral group of interest j, computed respectively against the allele frequency p_C of the contemporary population C and the average frequency p_A in all the A ancient groups:

$$covA(i,j) = (p_i - p_C)(p_j - p_A) \tag{2}$$

When comparing *covA* with outgroup $f_3(i,j;Yoruba)$[33], where j is one of the four ancestral groups, the statistics are different but strongly correlated (see Figure S7): this is expected when the f_3 outgroup population is an outlier to all populations, contemporary and ancestral, considered in *covA*, as in $f_3(i,j;Yoruba)$. Indeed $covA(i,j)$ has a strong relationship with f-statistics[34], i.e. $covA(i,j) = f_4(i,C;j,A) = f_3(i,j,A) - f_3(C,j,A)$ where C is the contemporary population (Estonians in our case) and A is an ideal population with $p = p_A$. Nevertheless, as opposed to f-statistics, which include allele frequencies in groups that portray actual populations, $covA(i,j)$ includes p_A, an average allele frequency which only serves as balanced comparison for the ancestries under analysis. In relation with our aim, this constitutes an advantage of *covA*, which does not take into account drift or selection occurred in the branch

that connects the outgroup population with the internal node shared by the other populations under analysis.

Predicting traits with *covA* and *covA*-based PCs

We fitted each standardized trait t_i with a model including one standardized *covA* for each ancestry j and estimated its coefficient: $t_i = \beta_j covA_j + \epsilon_i$. We adopted a logistic regression for categorical traits, which were transformed to $\{0, 1\}$ where 1 stands for the specified category and 0 for all the others. In addition, each trait was regressed against three PC-transformed *covA*s: $t_i = \beta_1 PC_1 + \beta_2 PC_2 + \beta_3 PC_3 + \epsilon_i$. Notably, we transformed all *covA*s using the loadings obtained from a PC analysis run on whole genome *covA*s, thus obtaining components that were largely independent, yet not strictly principal. These Independent Components (ICs) were standardized and included together as predictors. To evaluate association we used coefficient Z-scores computed against the same parameter extracted from 50 random genomic sets with matching size.

In the genome-wide analysis, we adopted similar steps, performing individual regressions for all the *covA*s and coupling this with a model including all *covA*s PCs, but socioeconomic variables were added as covariates in all models as described in the result section. Note that in this analysis ICs are not needed anymore, but actual PCs are used. Then, the standardized coefficient (β or effect size), or the Odds Ratio (OR) were directly used to assess ancestry-trait association for continuous and categorical traits respectively. This analysis was restricted to samples for which socioeconomic covariates were defined, i.e. 38,996 samples (including relatives): the actual sample size for this analysis is therefore less than reported in Figure 1 and Table S2.

Candidate genomic regions

We downloaded GWAS hits from GWAS catalog[14] (date of download: 20/11/2020) and then extracted for each trait a set of hits connected to it filtering on the reported trait ("TRAIT/DISEASE" field) or selecting the appropriate trait in the Experimental Factor Ontology (EFO) field, as specified in Table S3. Then we took windows of 5, 50 and 500 Kbs centered on the selected hits and merged them where overlapping, obtaining three sets of candidate regions for each trait. To perform the Z-score analysis, for each of them we obtained 50 matching window sets randomly placed across the genome.

Testing for signals of positive selection

In order to test individual SNPs for signatures of positive selection we utilized the Relate/CLUES pipeline[22,35]. This was applied on a curated subset of 1800 unrelated samples; further details on its application are described in Relate/CLUES Supplementary Methods. CLUES was run

once for each of the 14,712 unique GWAS hits for traits analyzed here with a derived allele frequency (DAF) above 1% and passing the 1000 Genomes strict mask. To obtain an expected distribution we randomly sampled 10,000 GWAS hits from the GWAS catalog meeting the same conditions and ran CLUES for positions not present among the 14,712 SNPs. Next, for each phenotype we compared its distribution of the logLR values to that of random GWAS hits. We took 1000 random subsets (with replacement) from the 10,000 logLR values each of the same length as the number of GWAS hits for a given phenotype and ranked the logLR values from lowest to highest within each subset. In this way we obtained 1000 values for each logLR rank from 1 to N where N is the number of SNPs analyzed for a given phenotype. For each rank we calculated the mean and the 5^{th} and 95^{th} percentiles. Finally, we rank SNPs within each trait and compare each logLR value to the mean and $5^{th} - 95^{th}$ percentiles range for the corresponding rank of the background distribution. As we are interested in deviations in the higher ranks we focus on the top 100 ranks for each phenotype. Such an approach is conservative as we are testing not against presumably neutral SNPs but against random GWAS hits that are shown to be enriched in signals on natural selection compared to random SNPs in the genome[35].

0.3 Quantification and statistical analysis

Statistical details to obtain any p value or significance assessment mentioned in the text are given immediately in the text and in the figure captions. remaining statistical methods and softwares are specified in "Method details" and listed in the "Key resources table".

0.4 Key resources table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Estonian genetic and phenotypic data	Estonian Biobank	https://genomics.ut.ee/en/access-biobank
AADR and Human Origins dataset	Allen Ancient DNA Resource	https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data
GWAS hits	GWAS catalog on 20/11/2020	https://www.ebi.ac.uk/gwas/
eQTL and sQTL data	GTEx portal on 18/10/2021	https://www.gtexportal.org/home/
1000 Genomes strict mask	The 1000 Genomes Project Consortium [36]	
Software and algorithms		
PLINK 1.9	Chang *et al.* [37]	https://www.cog-genomics.org/plink2
Eigensoft-6.1.4	Patterson *et al.* [38]	https://alkesgroup.broadinstitute.org/EIGENSOFT/
LDAK 5.0	Speed *et al.* [32]	https://dougspeed.com/ldak/
Relate	Speidel *et al.* [35]	https://myersgroup.github.io/relate/
CLUES	Stern *et al.* [22]	https://github.com/35ajstern/clues
Analysis Pipeline	This paper	bitbucket_repository_url_upon_publication

References

1. Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513,** 409–13 (2014).
2. Haak, W. *et al.* Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522,** 207–211 (2015).
3. Allentoft, M. E. *et al.* Population genomics of Bronze Age Eurasia. *Nature* **522,** 167–172 (2015).
4. Racimo, F. *et al.* The spatiotemporal spread of human migrations during the European Holocene. *Proceedings of the National Academy of Sciences* **117,** 8989–9000 (16 2020).
5. Olalde, I. *et al.* Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European. *Nature* **507,** 225–8 (2014).

6. Mathieson, I. *et al.* Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528,** 499–503 (2015).

7. Saag, L. *et al.* Genetic ancestry changes in Stone to Bronze Age transition in the East European plain. *Science Advances* **7,** eabd6535 (2021).

8. Cox, S. L., Ruff, C. B., Maier, R. M. & Mathieson, I. Genetic contributions to variation in human stature in prehistoric Europe. *Proceedings of the National Academy of Sciences of the United States of America* **116,** 21484–21492 (2019).

9. Ju, D. & Mathieson, I. The evolution of skin pigmentation-associated variation in West Eurasia. *Proceedings of the National Academy of Sciences of the United States of America* **118,** e2009227118 (2021).

10. Berens, A. J., Cooper, T. L. & Lachance, J. The Genomic Health of Ancient Hominins. *Human Biology* **89,** 7 (1 2017).

11. Berg, J. J. & Coop, G. A Population Genetic Signal of Polygenic Adaptation. *PLoS Genetics* **10** (8 2014).

12. Racimo, F., Berg, J. J. & Pickrell, J. K. Detecting Polygenic Adaptation in Admixture Graphs. *Genetics* **208,** 1565–1584 (2018).

13. Leitsalu, L. *et al. International Journal of Epidemiology* **44,** 1137–1147 (2015).

14. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research* **47,** D1005–D1012 (2019).

15. Tambets, K. *et al.* Genes reveal traces of common recent demographic history for most of the Uralic-speaking populations. *Genome Biology* **19,** 139 (2018).

16. Saag, L. *et al.* The Arrival of Siberian Ancestry Connecting the Eastern Baltic to Uralic Speakers further East. *Current Biology* **29,** 1701–1711.e16 (2019).

17. Damgaard, P. d. B. *et al.* 137 ancient human genomes from across the Eurasian steppes. *Nature* **557,** 369–374 (2018).

18. Pankratov, V. *et al.* Differences in local population history at the finest level: the case of the Estonian population. *European Journal of Human Genetics* **28,** 1580–1591 (2020).

19. Yang, J. *et al.* Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nature Genetics* **47,** 1114–1120 (10 2015).

20. Liu, H. Genetic architecture of socioeconomic outcomes: Educational attainment, occupational status, and wealth. *Social Science Research* **82,** 137–147 (2019).

21. Morris, T. T., Davies, N. M., Hemani, G. & Smith, G. D. Population phenomena inflate genetic associations of complex social traits. *Science Advances* **6,** eaay0328 (2020).

22. Stern, A. J., Wilton, P. R. & Nielsen, R. An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data. *PLOS Genetics* **15,** e1008384 (2019).

23. Key, F. M., Fu, Q., Romagne, F., Lachmann, M. & Andres, A. M. Human adaptation and population differentiation in the light of ancient genomes. *Nature Communications* **7,** 1–11 (2016).

24. Bersaglieri, T. *et al.* Genetic signatures of strong recent positive selection at the lactase gene. *American journal of human genetics* **74,** 1111–20 (2004).

25. Pickrell, J. K. *et al.* Signals of recent positive selection in a worldwide sample of human populations. *Genome Research* **19,** 826–837 (2009).

26. Ding, K. & Kullo, I. J. Geographic differences in allele frequencies of susceptibility SNPs for cardiovascular disease. *BMC medical genetics* **12,** 55 (2011).

27. The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369,** 1318–1330 (6509 2020).

28. Wang, W., Wang, C., Xu, H. & Gao, Y. Aldehyde Dehydrogenase, Liver Disease and Cancer. *International Journal of Biological Sciences* **16.** ISSN: 1449-2288 (6 2020).

29. Mutations in QARS, Encoding Glutaminyl-tRNA Synthetase, Cause Progressive Microcephaly, Cerebral-Cerebellar Atrophy, and Intractable Seizures. *The American Journal of Human Genetics* **94** (4 2014).

30. Buckley, M. T. *et al.* Selection in Europeans on Fatty Acid Desaturases Associated with Dietary Changes. *Molecular Biology and Evolution* **34,** 1307–1318 (2017).

31. Mathieson, S. & Mathieson, I. FADS1 and the Timing of Human Adaptation to Agriculture. *Molecular Biology and Evolution* **35,** 2957–2970 (2018).

32. Speed, D., Holmes, J. & Balding, D. J. Evaluating and improving heritability models using summary statistics. *Nature Genetics* **52,** 458–462 (2020).

33. Reich, D., Thangaraj, K., Patterson, N., Price, A. L. & Singh, L. Reconstructing Indian population history. *Nature* **461,** 489–494 (2009).

34. Peter, B. M. Admixture, population structure, and f-statistics. *Genetics* **202,** 1485–1501 (2016).

35. Speidel, L., Forest, M., Shi, S. & Myers, S. R. A method for genome-wide genealogy estimation for thousands of samples. *Nature Genetics* **51,** 1321–1329 (2019).

36. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

37. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4** (1 2015).

38. Patterson, N., Price, A. L. & Reich, D. Population Structure and Eigenanalysis. *PLoS Genetics* **2** (12 2006).