nature portfolio

Peer Review File

Ancient DNA reveals pervasive directional selection across West Eurasia

Corresponding Author: Dr David Reich

Any redactions in this file are there to maintain patient confidentiality, the confidentiality of unpublished data, or to remove third-party material.

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Referee #1

(Remarks to the Author)
The manuscript 2024-09-19432 entitled “Pervasive findings of directional selection realize the promise of ancient DNA to elucidate human adaptation” by Akbari et al. reports a genome-wide scan for natural selection occurring over the last 14,000 years in West Eurasians. By applying a generalized linear mixed model of allele frequency as a function of time in 8,433 ancient and 6,510 contemporary individuals, they detect 347 independent loci with >99% probability of selection, and suggest that a large fraction of the human genome is in LD with positively selected variants. They discuss several cases of selection on genes related to host defense against pathogens, metabolism, and hair morphology, and revisit the evolutionary history of several disease-causing mutations suspected to evolve under selection, including fluctuating selection. They report robust evidence for coordinated selection on polygenic traits related to skin pigmentation, fat storage, walking pace, smoking, and educational attainment.

This study represents a major step forward in the study of human genetic adaptations. Its novelty and impact lie in the generation of the largest temporal transect of ancient and modern genomes to date, combined with the use of a GLMM to test for association between allele frequency and time, while controlling for population structure. I commend the authors for developing the https://reich-ages.rc.hms.harvard.edu/ browser, which will be an extremely valuable resource for the research community. Nevertheless, the current version of the manuscript lacks a careful evaluation of the methodological approach used, making it difficult to assess whether it is robust to confounding factors. Furthermore, the authors have likely overestimated the number of selected loci due to the criteria used to define independent signals. I have several suggestions that, I believe, the authors should consider to address these potential weaknesses, and to hopefully improve the manuscript.

1. Variant QC filters: The authors have merged three large genomic data sets of different nature, raising the possibility that batch effects can confound some selection signals. To mitigate such biases, they have applied extensive QC filters that I find justified and well-thought. Nevertheless, these filters rely heavily on empirical thresholds, so that the filtered dataset may still contain variants affected by weak batch effects, potentially leading to false positives. Conversely, some selected variants may have been erroneously removed, resulting in false negatives. To explore this potential issue, the authors could evaluate whether their 347 candidate variants show QC statistics close to exclusion thresholds. Specifically, I suggest they highlight the candidate variants on Figures S1.3 and S1.4, and check whether candidate variants are enriched in 1240K or imputed variants. If many candidate variants are imputed, or if most variants with $|X| > 5.45$ at a candidate locus are imputed, it would be informative to check if IQS is lower relative to imputed variants with similar MAF. Additionally, variants with low LD scores and MAF (Figure S2.1) seem to show inflated statistics, possibly due to lower imputation accuracy. Do candidate variants show higher (as expected under selection) or lower LD scores than matched variants? Related to this, why was the non-replicated signal from Mathieson et al. (rs1979866, Figure S5.1) filtered out?

2. Sample QC filters: Can the authors provide details on the number of samples removed based on each QC filter? The authors exclude samples with a reported date that deviates from the date predicted from variants most associated with reported dates across all samples. This filter largely depends on the accuracy of this prediction. Could the authors present a plot showing both the reported and estimated dates, highlighting the excluded samples? Also note that in the Supplementary Information Section 1, the authors state that they excluded samples with a date <15,000 years BP or with the absolute value

of the regression residual <500 years. I guess they meant the inverse.

3. Power of the GLMM method: The authors acknowledge that the strong genomic inflation of their χ^2 statistic (i.e., the squared Z-score for the number of standard errors s is from zero) may be due to batch effects, the quality of genotype imputation, data preprocessing, population structure, genetic drift, background selection, or a large number of true positives. Whereas inflation caused by batch effects and imputation accuracy seems to have been largely addressed through extensive QC filters, the extent to which the K covariance matrix in equation (1) accounts for drift, background selection, and complex selection scenarios remains to be fully evaluated. The authors performed simulations by randomly sampling genotypes constrained by the observed GRM, a given selection coefficient, and initial allele frequency. However, they do not evaluate their FPR under drift, background selection, or positive selection in the source populations only. I strongly suggest that the authors simulate, for example with SLiM, neutral loci, loci under background selection, and loci where a positively selected variant in source populations becomes neutral after admixture, all under a realistic demographic model of west Eurasians (which has been largely established by some authors of this study). Regarding selection in the sources only, which has been addressed by several previous selection studies using ancient genomics data, several of the 347 candidates show trajectories compatible with a temporal change in frequency caused by shifts in ancestry for variants that were already differentiated across ancestries prior to admixture (e.g., candidates 5, 8, 19, 41, 45, 46, 49, etc.).
Note that in Supplementary Information section 2, the authors mention conducting simulations to estimate a control factor for their method, but they provide very few details. Similarly, they refer to 10,000 simulations for a scenario involving population structure, sample size, and temporal distribution of samples matching real data (Extended Data Figure 14). What demographic model did they assume for these simulations?

4. Controlling for FDR: The authors propose an interesting alternative to FWER, by estimating the FDR from the enrichment of variants with significant changes in frequency over time in trait-associated variants. The observed enrichment plateaus at a given significance threshold, suggesting that the enrichment of significant selection signals in GWAS hits, and presumably the proportion of true positives among significant signals, is maximal at this threshold. However, if Pan-UK BioBank GWAS hits include false positives caused by partially controlled population structure, particularly for high-FST variants, the FDR may be underestimated. It is not clear from Karczewski et al., medRχiv 2024 whether residual stratification could affect their results (see their Extended Data Figure 3). Can the authors repeat their enrichment analysis using sibling-based GWAS statistics?
One alternative explanation for the observed plateau is that true positive signals of selection may overlap less with GWAS hits as the selection coefficient increases. Several arguments support this hypothesis: large effects would be expected for large-s variants (assuming a monotonous relation between s and β), whereas GWAS hits typically have low effects on traits. Additionally, there is widespread evidence for negative selection on complex traits studied by GWAS, and actual oligogenic traits that were targeted by strong directional selection have possibly not been studied by GWAS. Is the enrichment decreasing for $|X| > 7$ (although the data is probably noisier)?

5. Independent loci under selection: The authors identify independent loci under selection by considering that all SNPs in LD with the strongest signal of the genomic region reflect the same selective signal, which makes total sense. They considered that two SNPs are in LD if $r^2 > 0.05$ in modern Europeans from the 1000 Genomes Project. However, under a selective sweep, hitchhiked neutral mutations occurring on the selected haplotype may show low r^2 (but high D') with the selected variant. Furthermore, present-day LD levels are probably underestimating past LD levels, as population size in Europe has largely increased over the last ten millennia. Consequently, the authors may overestimate the number of selective events that occurred in West Eurasians over the last 14,000 years. When inspecting the trajectories of the 347 signals, I observed several cases where a relatively rare derived allele, located a few hundred kb apart from a well-known positively selected mutation, increases in frequency tens of generations after the latter. For example, there are nine "independent" signals (signals 31 to 39 in Figures S4.3 and S4.4) within 5 Mb surrounding rs4988235, the selected variant responsible lactase persistence. Among those, rs62160512 and rs143866305 show D' > 0.7 (in 1KG EUR phase 3) with rs4988235. Another example is the TLR1 locus; the authors found ten "independent" signals (signals 62 to 71, Figure S4.6) within 500 Kb of the selected gene. This includes rs3188469 that shows D' = 1 (in GBR) with rs5743618, the likely target of selection at this locus. I strongly suggest that the authors revisit their LD criteria to define independent selection signals, by conducting simulations of selective sweeps and leveraging LD levels in ancient genomes.

6. Re-evaluating previous studies: I commend the authors for their systematic re-evaluation of previous scans for recent positive selection in ancient and modern West Eurasians. Nevertheless, the authors often make the strong claim that most previous signals not confirmed by their analyses are false positives. While this assertion may largely hold true, given the much larger sample size used in this study, I suggest that the authors provide more nuanced explanations: the variant QC filters applied may be too stringent (see point 1), and their GLMM method may miss true signals of fluctuating selection, such as those associated with OCA2 and TYK2. Prior selection studies have searched for selection that can start at different epochs within the last ten millennia, both on de novo or standing variation, allowing for the detection of fluctuating selection.

7. Evidence for polygenic selection: The authors report "12 traits with significant signals from all three tests after correction for number of traits tested". However, they should clearly state how many of these traits show replication when using Biobank Japan statistics and sibling-based statistics. Furthermore, they should clarify why Extended Data Figure 13 shows that the 95% confidence intervals for γ, γ_sign and r_s include zero for most of the traits, using GWAS statistics for both unrelated people and siblings. Finally, the significance thresholds should be displayed in Extended Data Figure 11.
As previously shown by some authors of this study, signals of polygenic selection can be confounded by uncorrected population stratification. The authors address this issue by testing the association between PRS and time using -1 and 1 as weights instead of effect sizes in the PRS calculation. However, I wonder whether the results shown in Extended Data Figure 10 could be affected by stratification, which might explain the unexpectedly polygenic signal observed for skin

pigmentation. Notably, the authors state in the figure legend that the plot shows “the correlation of a trait with selection summary statistics (r_s)”, whereas my understanding is that r_s represents the correlation between s and β.

Minor comments:
- L. 2-3: The authors state that they developed a method “that leverages an opportunity not utilized in previous scans: testing for a consistent trend in allele frequency change over time”. This is not the case, as Kerner et al. and Irving-Pease et al. estimated s from the full allele trajectory, using ABC and CLUES, respectively.
- L. 15-17: The authors interpret the significant decrease in the PRS for body fat percentage, waist circumference and waist-to-hip ratio as supportive of the “Thrifty Genotype” hypothesis (Neel JV, 1962). This is problematic, because there is evidence that the transition to food production was not a period of “plenty”, but was characterized instead by nutritional deficiency, reflected by the prevalence of low stature and porotic hyperostosis in early European farmers (Cohen & Armelagos, 1984; Macintosh et al., PLoS One 2016; Theodorakopoulou et al., Arch Balk Med 2020; Marciniak et al., PNAS 2022). This observation disagrees with the authors’ results, as the period when negative selection against higher body fat percentage was the strongest corresponds to the period when food production emerged in Europe, ~8,000 years ago (Figure 4). Note also that the hypothesis is not correctly formulated in L. 344-345 (“genetic adaptation to store fat in times of plenty, became deleterious after the transition to food-production”). The “Thrifty Genotype” hypothesis proposes the inverse: genetic adaptation to store fat in times of food scarcity, became deleterious after the transition to food-production.
- L. 54-55: When the authors state that “changes in frequency due to selection are often less than what can be expected from random genetic drift”, they somehow imply that their method is able to detect selection even when its effects are weaker than those of genetic drift, but they provide no such evidence. They should clarify further why the GLMM approach can outperform other approaches, at equal sample size.
- L. 191: The authors decided to focus on 36 candidate variants, out of 347 signals, based on the fact that these candidates were of “particular interest”. These include 7 variants with 64% < π < 98%, which could be false positives. Instead, it would be informative to comment the more confident candidate variants (π > 99%) with, for example, the largest GWAS effect sizes on traits.
- L. 201: rs3891176 is also strongly associated with type 1 diabetes (P = 8.7 × 10-229; Chiou et al., Nature 2019)
- L. 212: Please cite the phenotypes most associated with A and B alleles.
- L. 269: The authors do not test for balancing selection, which would result in a non-significant test. The trajectory of ΔF508 could be compatible with balancing selection.
Perhaps the authors can describe here the independent signal they found in the 5’UTR of CFTR (candidate 199).
- L. 383-384: The significant genetic correlations between traits showing evidence of directional selection are puzzling. The authors may consider this is beyond the scope of this study, but one possibility is to follow the approach in Stern et al., Am J Hum Genet 2021, to disentangle polygenic adaptation on correlated traits, assuming a multivariate version of the Lande approximation and leveraging their GLMM-based s estimates.
- L. 402-404: The authors state that “there have in fact been changes in selection pressures at a number of the loci […] (Extended Data Figure 6), including at HLA-DRB1, TYK2 and HFE”. It is unclear why they have assumed fluctuating selection for some genes and not for others, how many genes were tested for fluctuating selection, and if they reported only the significant results. This should be clarified.
- L. 408: The authors state that “selection coefficients on the alleles must have shifted over time” because the estimated age of favored alleles in a selective sweep is lower than the date of origin of the mutation inferred from RELATE. However, this observation is true in absence of fluctuating selection, as RELATE assumes neutrality and will systematically overestimate the age of a selected mutation (Speidel et al., Nat Genet 2019).
- L. 450-451: The linear mixed model was also already used in the context of genomic scans of natural selection (Yang et al., PNAS 2017).
- L. 460: The authors justify the choice of the GLMM method by the fact that, “in practice, it detected many loci”. These may be false positives.
- L. 512-513: the authors have grouped individuals into clusters of individuals with similar ancestry and coming from similar times, to make analysis tractable. It would be useful to show in a Suppl. Figure how the clustering affects |X| for a few significant signals.
- L. 543: Please clarify if variants used to compute the GRM were filtered on MAF, and in which sample of individuals f_k is computed.
- L. 554: Please replace “proces s” by “process”.
- L. 560-561: Please clarify why y_i was scaled by the standard error of PGS of the modern samples.
- L. 630: What makes the difference between the 56 and remaining 244 newly reported shotgun genomes that sum to 300?
- L. 663: Showing a PCA identifying modern and ancient genomes, as well as sampling regions, would be very useful, in addition to Figure S1.2.
- L. 642/683: The authors refer to either 5,227 or 5,382 newly reported ancient DNA libraries. Please clarify.
- Figure 1c: Please report as a Table.
- Figure 2c: Please replace “Dervied” by “Derived”.
- Extended Data Figure 1: An admixture plot of the genomic transect used would be very useful.
- Extended Data Figure 3b: In the legend, the authors indicate that s = 0.01, but they show curves for variants under neutrality and negative selection as well.
- Extended Data Figure 5: Please define horizontal lines. Clarify in the legend why modern genomes were excluded.
- Extended Data Figure 8: It may be interesting to indicate in the main text that s tends towards 0 for traits 10-12.

Referee #2

(Remarks to the Author)
This manuscript presents a new method for detecting directional selection from time-series data and reported “pervasive findings of directional selection”, using an expanded European ancient DNA dataset. The authors additionally report new signals of polygenic selection.

Unfortunately, there are major concerns about the statistical machinery and interpretations. It remains unclear whether the pervasive findings of directional selection (both locus-specific or polygenic) are genuine or are artifacts of biases from failing to adequately correct for confounders and/or using a loose significance cutoff. As it stands, the paper makes strong claims without adequate support.

Major comments:

This manuscript used a GLMM which is motivated by the LMM commonly used in GWAS to correct for potential population structure. However, it is not immediately clear from theory why this should be well-calibrated and simulation verification is critically needed. The simulations as performed (L610) are inadequate as they are based entirely on the assumed model rather than probing the complexities of real population genetic scenarios.

It's essential to evaluate the following: (1) whether the GLMM works for a no-gene-flow case, with a well-calibrated p-value for a few Ne values (forward-in-time simulations); (2) whether the GLMM works for a realistic scenario of complex European history, population movement and gene-flow; (3) what is the impact of linked selection from background selection? What impact do these scenarios have on the distribution of test statistics?

I’m perplexed that the central output of the analysis is to model and report frequency changes as a function of time. It is well known – thanks in large part to the Reich lab -- that there have been major ancestry turnovers in European history. These have the potential to drive significant allele frequency changes. For example, if the WHG, Farmers or Steppe populations had markedly different frequencies at any given locus, the model could interpret this as a rapid selective shift. Browsing Figure 3 suggests several cases where this may be the correct interpretation. In some cases, outlier variants between populations may indeed result from past selection, but the time and place of that selection would be far removed from what is reported by the model. I view it as essential to recast the model in a more biologically motivated manner.

I am troubled by how heavily the paper relies on the overlap with GWAS loci to set the significance threshold. Genome regions containing GWAS hits differ from random loci by many measures (eg gene constraint and conservation, gene density, measures of chromatin landscape, predicted background selection, etc). Without really understanding the properties of the GLMM very well, how sure are we that these features aren’t correlated with the probability of large values of X? I’d worry particularly about background selection as it’s entirely unclear if sigma_j can adequately capture this effect.

Moreover, the authors don’t really spell out exactly why they expect the overlap with GWAS variants to be a reliable metric for their purpose. This might make more sense to me if we thought these traits were the actual targets of selection, but instead the UKB traits provide an eclectic sample of a small part of the possible trait space. Presumably the authors have in mind some implicit hypothesis that there is enough pleiotropy between the measured traits and the targets of selection. But there’s also a second problem here: presumably most of these GWAS hits have quite small effects on specific traits--perhaps explaining 1/1000th of the variance. Why should these have selection coefficients of ~1%? And third, why would we see selection acting on specific complex trait variants and not on other variants that affect the same trait? For example, why does one particular balding variant show up in the data, while there is no polygenic signal for balding?

I couldn’t tell if HLA is included in this analysis. HLA could have a huge biasing effect in this type of analysis and it would be best to exclude it from all genome-wide analyses and only report HLA results in single-locus analyses.

The paper claimed the signals arise from directional selection, but directional selection is not the only mechanism to cause frequency shifts over time. Background selection and stabilizing selection can also change the distributions of allele trajectories of common variants. The effects of background selection cannot be correctly modeled as a simple varying Ne along the genome, as represented by the locus-specific noise term in the GLMM.

The paper also reports potentially interesting results on polygenic adaptation. However, it is well known that these tests are highly sensitive to stratification. The quality of GWAS has certainly improved since the 2019 eLife papers on this topic, but this is still a critical concern. One potential issue with replication in a second population sample is that this strategy can be biased if the SNPs are initially ascertained in a stratified sample (https://elifesciences.org/articles/61548).

Family GWAS, on the other hand, is less likely to be biased by stratification (depending on how SNP ascertainment is performed). However, the authors were unable to replicate the signals of polygenic selection using family GWAS. I accept that the sib GWAS may be underpowered, but it may also be a warning sign that the signal is spurious.

One might be particularly worried that traits including education, intelligence, and household income show up so strongly in this analysis. The emerging consensus is both that these traits have relatively low true SNP heritability, and that much of the polygenic score may be driven by stratification. Added to general concerns about portability of PGS across environments (how portable are these PGS to the neolithic?) it’s hard not to be at least a little skeptical about these results. Among many possible references:
https://pubmed.ncbi.nlm.nih.gov/38225408/
https://pmc.ncbi.nlm.nih.gov/articles/PMC9627830/

https://www.medrxiv.org/content/10.1101/2024.10.01.24314703v1

The limited data release policy (L642) is unacceptable. Future researchers studying selection would certainly need access to more detailed meta-data including, but not limited to, geo-locations and basic archaeological context such as putative population group. Beyond enabling effective future use of the data, there's also a serious concern that if some of these data are never reported elsewhere, they could remain in limbo indefinitely, with no meaningful metadata and open-ended publication restrictions. And it's worth noting that the proposed data release policy is literally the opposite of the Fort Lauderdale principles since the point of the Fort Lauderdale agreement was to create a framework for full data release to allow resource users "to use the data in any creative way. There should be no restrictions on the use of the data… In some cases, this might best be done by discussion or coordination with the resource producers".
https://www.genome.gov/Pages/Research/WellcomeReport0303.pdf

Minor comments:
It could be helpful to draw the p-value QQ-plot, and check if the inflation is global or just the tail. The latter would support the pervasive findings of directional selection, while the former would support poorly-calibrated statistical test and the selection coefficients cannot really be trusted;

The word "Derived" is mis-spelled in the x-axis label in Figure 2c;

The HAF-score results (Supplementary Information Section 3) are derived under a very simple case of constant population size, and here it could be varying population size and massive gene-flow. It is hard to justify the correctness of these results;

The authors have inferred allele frequencies of ancestral populations with qpAdm, can the authors check how much frequency change can be attributed to mixture composition change versus outliers which are likely true selection targets? (Supplementary Information Section 6)

The title rather overpromises: even if we take the 347+ hits as correct, it's not yet the case that ancient DNA now "realize(s) the promise" to "elucidate human adaptation", any more than finding hundreds of GWAS hits has elucidated the molecular pathways of complex traits such as schizophrenia.

Referee #3

(Remarks to the Author)
Summary comments

Using the largest ancient DNA dataset ever assembled for this purpose, the paper claims to have found evidence for an order of magnitude more signals of adaptive responses to natural selection in humans than have been discovered in previous studies. Notably, it does this primarily by claiming to be able to overwrite the existing statistical significance standards of this field via a new enrichment procedure. In my opinion, there is not enough evidence to justify this.

To briefly summarize, the authors used a generalized linear mixed model (GLMM) with a logistic link function to model allele frequency change as a function of time in a sample of over 8,000 ancient individuals. To account for variation in ancestry, they modeled the among-individual covariance structure using a genetic relatedness matrix (GRM) estimated from genome-wide data. The test statistic is heavily inflated relative to the null expectation. The authors argue that this inflation is due to an abundance of real signal and devise a procedure for determining a significance cutoff as a function of the enrichment of GWAS hits among selection signals. I did not find this approach very convincing. The authors also analyzed polygenic scores and reported some signals. I have concerns about whether the efforts to avoid population structure confounding have been successful.

Major comments on single locus outlier scan

The authors' GLMM approach is very similar to environmental association methods that control for a GRM, but it uses time as the predictor rather than an environmental variable (e.g., LFMM, Bayenv, etc.). In contrast, previous papers analyzing aDNA time trends have used only the top principal components of this GRM. The authors argue that the PCA approach suffers from collinearity between the PCs and the time axis, which reduces its power, whereas the GLMM does not have this problem.

However, I think this represents a misunderstanding, or at least a misstatement, of the difference between the two methods. The GLMM method can be viewed as a model that includes all PCs, but with a prior on the magnitude of the contribution from each PC that is proportional to the eigenvalue associated with it, and a single common scale of variance shared across all PCs (see, e.g.,
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0075707, or https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3002847). In this analysis, the GLMM primarily uses the scale of variation observed on "lower" PCs not correlated with time to determine how much variation there "should" be along the time axis. Therefore, the question is not whether including the kinship matrix in the GLMM results in the inclusion of covariates collinear with the time axis (it surely does), but whether the GLMM correctly estimates how much variation there should be along these axes in the absence of selection.

The authors suggest they performed simulations of selection to compare the two approaches (e.g., Extended Data Figure 14). However, based on the descriptions in the Methods, it appears they simulated data from the GLMM itself (with a time trend to simulate selection) and then compared the GLMM that generated the data against the PCA-based method. If this is correct, it is not surprising that the GLMM performs better, as it is the data-generating model. If this is wrong, the simulations need to be explained more clearly.

It is clear, given the extreme inflation of the test statistic, that the GLMM does not fit the real data well. The question is whether this is 1) because it is a poor null model, or 2) because there is rampant adaptation, as the authors suggest.

1) It is difficult to assess how poor the GLMM is as a model because the authors do not appear to have tested it in any real evolutionary simulations.

One possible issue with the GRM is the influence of background selection. The authors are cognizant of this potential confounder and try to address it, but I am not sure I am convinced. First, it is not clear that independently estimating sigma for each SNP is sufficient. The effects of background selection take multiple generations to accumulate and thus more strongly impact comparisons among more divergent groups than among less divergent groups. Because the GLMM effectively uses the variance on the lower axes to estimate how much variance there should be on the time axis (see above), it could underestimate the variance on the time axis in the presence of background selection.

I recognize that realistic simulations of the impact of background selection in this context would be challenging, but given that much of the GWAS enrichment signal appears to be confounded with variation in the strength of background selection (Extended Data Figure 3a; see also comments below), it seems an important and still unresolved issue.

Also, for the record, I did not find the HAF score analysis helpful. HAF scores are a measure of frequency change and will necessarily be large given that these loci were identified by having a significant correlation with time. Therefore, there is no new information. The simulations in the supplement about the distribution of HAF scores do not deal with this ascertainment bias and are not convincing.

2) The authors' preferred conclusion is that the test statistic is inflated because there is an abundance of true signals, and thus applying significance thresholds used in previous studies would be too conservative. Notably, if the authors apply a standard genomic control correction, followed by a standard genome-wide significance threshold of $p < 5e\text{-}8$, they find 48 signals, consistent with previous studies. Therefore, the order of magnitude increase in the number of signals reported follows entirely from the authors' conclusion that this threshold is too conservative.

This conclusion is justified by estimating a false discovery rate (FDR) as a function of the enrichment of selection signals among GWAS hits from UKBB traits. However, the authors need to clarify their FDR estimation procedure. The supplemental material contains some underexplained math, and it is unclear how this math relates to the curve-fitting procedure used to determine the FDR. Is the fact that the FDR plateaus at approximately 1 actually data-driven, or does it simply follow from the assumption that once the enrichment curve levels off, the FDR must be close to 1, as suggested by the verbal model in the main text? Furthermore, is the idea that the FDR is approximately 1 once the enrichment levels off consistent with the observation that the enrichment is fairly modest (approximately 3.9)? More generally, does the total level of enrichment play any role? For example, when the authors repeat the enrichment analysis while controlling for local background selection effects (Extended Data Figure 3), the enrichment drops to only approximately 2.1. I find it difficult to reconcile such a modest enrichment with the strength of the claim it supposedly supports.

Another question is whether all future selection scans should adopt this GWAS enrichment method to determine significance thresholds. Do the results of this paper suggest we should expect an order of magnitude increase in significant hits in future selection scans? In the discussion, the authors compare their results to earlier work (from over ten years ago) that searched for evidence of adaptation on deeper timescales. They try to explain the discrepancy in the number of signals detected. However, it remains unclear whether those earlier scans would have found so few signals if they had used the same procedure to determine a significance threshold.

I also want to note a general conceptual disconnect in this enrichment procedure. The effect sizes these variants have on their GWAS traits are presumably quite small and therefore likely not the actual cause of the putative selection signal. So it is fairly hazy as to exactly why this is supposed to provide such strong evidence. Thus, while the enrichment is interesting, I don't think there is enough evidence to overwrite the existing statistical significance standards in this field.

Major comments on polygenic score analysis

I have two main concerns about the polygenic score analyses. As always, the primary concern with this sort of analysis is that residual environmental confounding in the GWAS may have led to a bias in the distribution of polygenic scores, which can manifest as a signal of selection. This is particularly likely to be a problem when lenient p value thresholds are used to construct the polygenic score, as is the case here.

The authors are aware of these concerns and take steps to address them. However, it is unclear whether these steps have been successful, largely due to some ambiguity in the text regarding what was done.

The currently accepted best method for overcoming stratification bias in this kind of analysis is to use effect sizes from a

study with a population structure that does not overlap with the test panel. This can be accomplished using either family-based effects or a population GWAS that is evolutionarily diverged. The authors attempt both, using effects estimated from siblings and from East Asians.

However, this method is only guaranteed to prevent bias if the unconfounded effect sizes are used both for ascertaining SNPs to be included in the polygenic score and for actually calculating the polygenic score itself. Using unconfounded effects only for the calculation but not for the ascertainment can still result in biased polygenic scores (see Figure 6 of Zaidi and Mathieson), because this ascertains a biased subset of SNPs. The text is unclear about the authors' exact approach, so it is difficult to judge whether the East Asian/sibling analyses are unconfounded.

Regarding the sibling effect analysis, I also want to note that the authors attribute the lack of meaningful insight to limited sample sizes in family-based GWAS (Extended Data Figure 13). However, they do not provide any power analysis to support this conclusion, and in fact several signals are no longer significant. While limited sample size may indeed be a factor, the authors should clearly state that multiple signals do not replicate.

Minor comments:

- The first sentence of the abstract is simply untrue. For example, here are three papers that test for consistent allele frequency changes over time. There may be more these are just three that came immediately to mind.

Ju and Mathieson
https://www.pnas.org/doi/abs/10.1073/pnas.2009227118

Mathieson and Terhorst
https://genome.cshlp.org/content/32/11-12/2057.short

Fine and Steinruecken
https://www.biorxiv.org/content/10.1101/2024.05.10.593575v1

- the title of SI section three is “HAF score analysis proves directional selection”
Above I said that I didn’t find this section persuasive, but I also think the authors should remove the world “proves” and talk about “evidence” or something, especially in a scenario such as this where we are relying on fraught statistical inferences about things we can never fully experimentally confirm.

Version 2:

Reviewer comments:

Referee #1

(Remarks to the Author)
In the revised manuscript 2024-09-19432B, Akbari et al. have partly addressed several of my initial comments, by (i) exploring the power of their GLMM model and the expected proportion of GLMM-significant variants among trait-associated variants with forward-in-time simulations of neutral and selected mutations, assuming different models of selection under a semi-realistic demographic model, (ii) estimating the number of independent selected alleles by using different clumping strategies based on both r^2 and D’; (iii), showing that the null distribution of the GLMM model is not normal and the test statistics are inflated for low-frequency variants; and (iv) reporting correlations between polygenic selection test statistics using GWAS in Europeans and East Asians. Although conclusions in the revised manuscript are strengthened, several of my main concerns have not been satisfactorily addressed.

1. TPR, FPR and FDR of the GLMM. The reported forward-in-time simulations were carefully designed and the way purifying, stabilizing and directional selection were modelled is very reasonable. These simulations show unambiguously that no enrichment of trait-associated variants in significant GLMM variants is expected under stabilizing selection. However, because the authors investigated true positive rates using a different significance threshold than their empirical threshold ($|X| > 5.45$, corresponding to P-value < $1.95 \times 10\text{-}16$; Fig. S3.2), the simulations do not inform the actual accuracy of their method, under either background selection or pre-admixture selection. In Fig. S2.10, they assumed FPR = $1.0 \times 10\text{-}5$; for the case of pre-admixture selection, they considered a Bonferroni corrected P-value = $1.25 \times 10\text{-}4$. Therefore, I invite the authors to report the TPR and FPR of the GLMM and GLM models under models 1 to 3 considering $|X| > 5.45$ as the significance threshold (the enrichment plateaus at this value in both empirical and simulated data; Extended Data Fig. 4). Plotting the X distribution for positively selected mutations, as well as mutations in coding, noncoding and neutral regions and mutations under pre-admixture selection would also be informative. Notably, it is currently unclear whether the GLMM has lower FPR than GLM in the presence of population structure. This is important because GLMM and GLM show similar TPR, when the number of PCs is carefully chosen, FPR = $1.0 \times 10\text{-}5$ and MAF > 0.1 (Extended Data Fig. 16, Fig. S2.10b). Furthermore, it is also unclear which simulated model was used in Figure S2.11 and S2.12. Finally, as initially requested by Reviewer 3, I invite the authors to provide more explanations about equations in pages 40-41, and detail the assumptions made.

2. Empirical consistency checks. The authors have evaluated whether their simulations reproduce some aspects of the true data, but these checks remain partial. The authors compared simulated and true PCA, and acknowledged that Irving-Pease

et al.'s demographic model largely underestimates variation in ancestry proportions among Europeans. This raises some concerns regarding power comparisons between GLMM and GLM using these simulations. In particular, Fig. S2.11 shows that $1 - R^2$ is lower for GLM in simulations, relative to observed data, suggesting underestimation of GLM power in the simulations (due to higher correlation between sampling time and PCs, relative of empirical data). Furthermore, I suggest the authors add observed heterozygosity in Fig. S2.4 (relative to intergenic regions, possibly), and the observed β/EAF relationship for a trait under stabilizing selection in their empirical analyses in Fig. S2.6 (Koch et al., bioRχiv 2024).

3. Pre-admixture selection. Instead of using forward-in-time simulations, the authors decided to simulate genotypes by sampling from a binomial distribution, assuming admixture proportions from Irving-Pease et al.'s model. However, simulating temporally varying s is easy to implement in SLiM, and would enable the authors to determine if the HAF statistic is inflated under pre-admixture selection, which is not currently investigated.

4. s estimation. I suggest that the authors use the simulations already performed to evaluate if their selection coefficient estimates are unbiased. In several parts of the manuscript, the authors comment GLMM s estimates, but they have not evaluated estimation accuracy.

5. LD clumping. The authors reported in a new Supplementary table the number of independent signals estimated considering different clumping strategies. Although some strategies result in a substantial reduction in the number of independent signals, and estimating such a number is a central objective of their study, they decided not to mention these results in the main text. The authors claim that their "approach to identifying distinct loci is conservative, because genuinely selected alleles in LD with nearby stronger ones will be missed" (l. 188-190), but they provide no evidence that their approach is indeed conservative. For example, by just taking a pair of nearby variants with very similar trajectories and s values in Fig. 3 (rs3891176 for locus 1 and rs7775397 for locus 19), I find with PLINK1.9 --ld that have $r^2 = 0.215411$ and $D' = 0.737$ in Europeans from 1KG (so one of the two variants should be filtered). Furthermore, their FDR approach assumes that variants are independent, which may not be true due to hitch-hiking effects, questioning the way they define genome-wide significance. Therefore, I strongly recommend that the authors investigate which clumping strategy provides the best estimate of the number of independent selected variants, based on the simulations already performed.

Minor comments.
X and HAF. The authors use HAF as an independent line of evidence, in addition to GLMM X. Can the authors compute the correlation between X and HAF in the different simulated models, without and with directional selection?

Simulations. Could the authors please clarify further how many 10-Mbp loci were simulated per model? Can the authors clarify how many 10-Mbp regions were used to compute PCs and the GRM?

Extended Data Fig. 3b. Please report HAF distribution under the semi-realistic demographic model.

Figure legends. The authors refer either to a proportion of SNPs associated in GWAS or an enrichment, when describing the y axes of plots reporting these enrichments. Please clarify and homogenise descriptions across panels and figures.

Figure 1 legend. Please remove 'perfectly' (l. 950).

Figure 2b and Extended Data Fig. 1. Please replace 'coeificient' by 'coefficient'.

Extended Data Fig. 15. The authors may caution that these genetic correlations can be inflated due to assortative mating.

Extended Data Fig. 16a. The curve for GLM (1) is not visible.

(Remarks on code availability)
I was able to download the code and run a toy example. Documentation remains relatively limited.

Referee #2

(Remarks to the Author)
This paper reports a series of very interesting results, and an extraordinary data set that will be of great interest to the field. The analysis is highly complex and there are details where one could still ask for more simulations or rationale, but at this point I am convinced that the major claims are likely correct. In summary, I am now supportive of publication in principle, and I congratulate the authors on what will be an important and widely-discussed paper.

I do have several specific comments that I would like to see addressed before acceptance:

1. The presentation of the GWAS FDR is absolutely critical to the paper, yet poorly described. I'd like to see the authors rewrite this.

-- The supplementary section on FDR control (around p. 40) is poorly explained and difficult to follow. I encourage the authors to improve the clarity of the mathematical derivations and provide a more transparent explanation of the logic behind the procedure. Is there some unstated assumption here? Maybe that the FDR is required to decrease all the way down to zero with increasing values of the test statistic? Please clarify that in the main text.

-- Beyond the math for this section, I'd like to see the authors articulate more clearly what exactly they think they are testing, and what is the rationale for what they are doing. They had some of this in the Response to Reviewers, but it's so central to the paper that I think it should be in the main text.

I agreed with some of what they wrote in the Response, but I would actually frame it a bit differently. They are including (nearly) all SNPs that are genome-wide significant in any of a huge number of traits. This encompasses around 12.5% of all SNPs. Importantly, they are NOT attempting in any way to colocalize the selection signals with GWAS hits. So we should really think of this as a type of genome annotation that spans ~12% of the genome, and which is highly enriched for likely functional variants, as well as SNPs that are in even pretty weak LD with functional variants. Note that the SNPs in this set must surely have higher LD than all SNPs genome-wide. So this is a very blunt annotation, and the authors could perhaps have done about as well by using other types of functional annotations: eg perhaps some combination of highly functional genes + conserved noncoding regions. That's ok in my view, but I think it would be very helpful to the reader to frame it in this way, rather than using language that may be misunderstood by readers to suggest that the selection signals are actually colocalizing with specific GWAS hits.

All that said, the 5-fold enrichment that the authors report here is very impressive, and this is the main observation that helps me to overcome any other concerns that I may have with the analysis pipeline. If, for example, they had found a 1.2-fold enrichment, this could also have been a highly significant overlap, but I would have had much more concern about possible confounders.

2. There should be straightforward ways to estimate how ancestry proportions change over time in the aDNA panel, and I encourage the authors to perform such checks. In particular, they could test whether variants strongly associated with ancestries that increase or decrease over time show systematically lower p-values. If so, this would indicate that the ancestry correction is still incomplete. In that case, the authors should explicitly note this limitation in the main text.

3. I suggest reconsidering the section “LD score and pseudo-intercept”. In linear mixed model-based GWAS, LD score regression is generally not appropriate because the LMM changes the expected mean of the chi-squared statistics, making the intercept hard to interpret. The proposed “pseudo-intercept”, defined using SNPs above an LD-score threshold, does not appear to solve this underlying issue. The statement in the supplement (p. 41) that agreement with the pseudo-intercept provides independent support for the approach is therefore not convincing, and I recommend removing or revising this section.

Version 3:

Reviewer comments:

Referee #1

(Remarks to the Author)
In the revised manuscript (2024-09-19432C), Akbari et al. have now largely addressed my previous comments. I had emphasised that it would be highly informative to determine which LD clumping strategy provides the most accurate estimate of the number of independent loci under directional selection. The authors argue that this question is too difficult to address. While I respectfully disagree, I do accept that the manuscript in its current form already presents numerous novel results and methodological innovations and represents as such one of the most important advances to date in the study of human evolutionary history.

Because the exact number of independent loci is inherently difficult to estimate, owing to (i) uncertainty in the significance threshold, (ii) residual LD among candidate loci, and (iii) the presence of genuinely independent, closely linked variants, I recommend that the authors adopt more cautious language when reporting the number of independent loci.

More specifically, I propose the following minimal changes in the main text:
- L. 10: “By applying this to 15836 West Eurasians who lived the past 18000 years and 6438 contemporary people, we find several hundred genome-wide significant signals with >99% probability of selection, an order of magnitude more than reported in previous studies.”
- L. 192: “Moreover, although residual LD may exist among candidate loci (Supplementary Information section 5), genuinely selected alleles in LD with nearby stronger ones can also be missed.”
- L. 504: “Although the exact number of independent loci cannot be estimated with high confidence, we project that there are at least 3800 independent signals of directional selection”

Please also find below a few other minor changes:
- L. 123: “The plateau occurs at the same place when we control for “background selection”, the loss of neutral variation linked to deleterious mutations removed by purifying selection, although the enrichment in GWAS variants is reduced (~2.4-times, Extended Data Figure 3a).”
- L. 178-179: Please consider citing Sikora et al., Nature 2025
- L. 379: “after correction for [the] number of GWAS (traits?) examined”
- L. 542: “Three theoretical reasons indicate that [it] cannot be explaining our observations”

(Remarks on code availability)
The code is now sufficiently documented.

Referee #2

(Remarks to the Author)

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.
In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.
The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.
To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/

Referee expertise:

Referee #1: human evolution and population genetics

Referee #2: population and statistical genetics

Referee #3: population and statistical genetics

Referees' comments:

Referee #1 (Remarks to the Author):

The manuscript 2024-09-19432 entitled “Pervasive findings of directional selection realize the promise of ancient DNA to elucidate human adaptation” by Akbari et al. reports a genome-wide scan for natural selection occurring over the last 14,000 years in West Eurasians. By applying a generalized linear mixed model of allele frequency as a function of time in 8,433 ancient and 6,510 contemporary individuals, they detect 347 independent loci with >99% probability of selection, and suggest that a large fraction of the human genome is in LD with positively selected variants. They discuss several cases of selection on genes related to host defense against pathogens, metabolism, and hair morphology, and revisit the evolutionary history of several disease-causing mutations suspected to evolve under selection, including fluctuating selection. They report robust evidence for coordinated selection on polygenic traits related to skin pigmentation, fat storage, walking pace, smoking, and educational attainment.

This study represents a major step forward in the study of human genetic adaptations. Its novelty and impact lie in the generation of the largest temporal transect of ancient and modern genomes to date, combined with the use of a GLMM to test for association between allele frequency and time, while controlling for population structure. I commend the authors for developing the https://reich-ages.rc.hms.harvard.edu/ browser, which will be an extremely valuable resource for the research community. Nevertheless, the current version of the manuscript lacks a careful evaluation of the methodological approach used, making it difficult to assess whether it is robust to confounding factors.

We appreciate the positive assessment, and agree about the importance of rigorous evaluation of our new methodology. We have spent a year exhaustively exploring potential confounding factors, and are now even more confident of our findings (the findings themselves are qualitatively unchanged). Most importantly, our revision now presents forward-in-time simulations using SLiM, which we use to evaluate the robustness of our approach with respect to potential confounding factors. None of a range of simulated processes that are not directional selection—including stabilizing selection and purifying/background selection—can produce the correlation of selection signals to GWAS signals and HAF score we observe in data (Supplementary Information section 2, Figures 1b,d, and Extended Data Figures 4,5). In contrast, when we simulate directional selection, we see the empirically observed pattern, and show we can use the enrichment to identify a valid threshold for genome-wide statistical significance.

Furthermore, the authors have likely overestimated the number of selected loci due to the criteria used to define independent signals. I have several suggestions that, I believe, the authors should consider to address these potential weaknesses, and to hopefully improve the manuscript.

For identifying candidate SNPs, we used the clumping approach from PLINK with a 10 Mbp window size and $r^2 < 0.05$, both of which are extremely aggressive parameters. Nevertheless, this is a simplistic and heuristic approach, and identifying independent signals of selection is challenging and interesting but beyond the scope of this paper. To address this concern, we added the following analysis in our revised manuscript in Supplementary Information section 5.

"We applied an additional pruning step to the 479 candidate SNPs identified after the initial clumping based on r^2. In this step, we calculated D′ among the candidate SNPs within a specified window size, prioritized variants by their GLMM p-values, and recursively designated tag SNPs. Any variant with D′ above a given threshold relative to a tag SNP was pruned. This process was repeated until all remaining variants were retained as tag SNPs or pruned. Thresholds for D′ were set at 1, 0.9, 0.5, 0.2, and 0.05, and window sizes were set at 100 kbp, 200 kbp, 500 kbp, 1 Mbp, 2 Mbp, 5 Mbp, and 10 Mbp. Table S5.1 reports the number of SNPs retained and pruned at each D′ threshold, separately for the HLA region and outside the HLA region and for each window size."

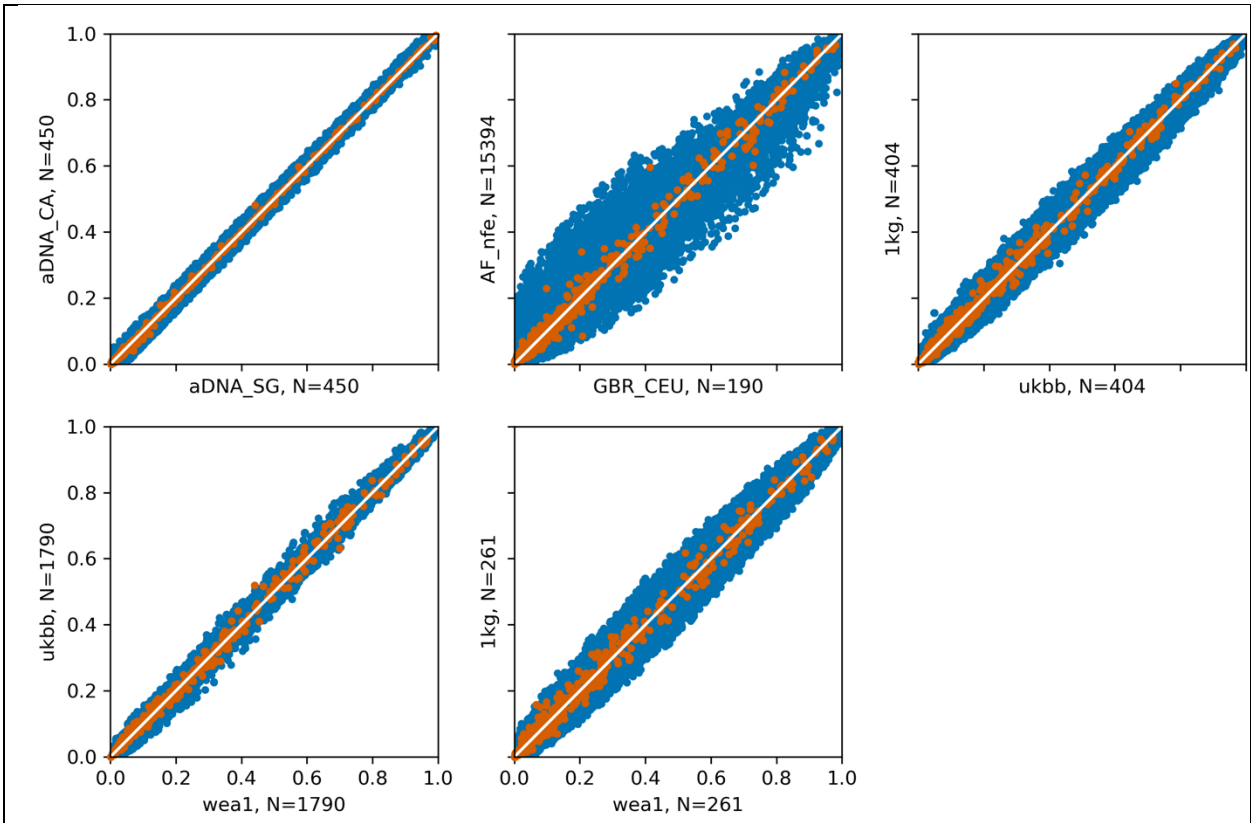
This analysis is summarized in Table S5.1. For a window size of 5 Mbp and $D' \leq 0.05$, we lose 50% of SNPs outside the HLA region (leaving 207 independent signals) and 99% in the HLA region (leaving just 1 signal). For a window size of 5 Mbp and $D' \leq 0.5$, we lose 23% of SNPs outside the HLA region (leaving 315 independent signals) and 94% within the HLA region (leaving just 4 signals). Both these filters seem too extreme given that the likelihood of selection on many HLA alleles in a time period when, even with this extraordinarily stringent filtering using D' as well as r^2, we find hundreds of signals of selection enriched in immune function.

In conclusion, different approaches to identifying independent signals of selection yield different SNP sets, but the main findings of the paper (pervasiveness of direction selection leading to hundreds of independent loci) remain robust regardless of the approach, and we retain the version of the analysis based entirely on the r^2 filter in the main text.

1. Variant QC filters: The authors have merged three large genomic data sets of different nature, raising the possibility that batch effects can confound some selection signals. To mitigate such biases, they have applied extensive QC filters that I find justified and well-thought.

Nevertheless, these filters rely heavily on empirical thresholds, so that the filtered dataset may still contain variants affected by weak batch effects, potentially leading to false positives. Conversely, some selected variants may have been erroneously removed, resulting in false negatives. To explore this potential issue, the authors could evaluate whether their 347 candidate variants show QC statistics close to exclusion thresholds. Specifically, I suggest they highlight the candidate variants on Figures S1.3 and S1.4,

Below is the requested figure showing that the majority of these variants are not close to the exclusion threshold.



Overplotting of the 479 candidate tag SNPs on 1 million random variants from Figure S1.1.

and check whether candidate variants are enriched in 1240K or imputed variants. If many candidate variants are imputed, or if most variants with |X| > 5.45 at a candidate locus are imputed, it would be informative to check if IQS is lower relative to imputed variants with similar MAF.

We performed an enrichment analysis for variants with |X| > 5.45 using MAF-matched controls from the remaining variants, where MAF matching was performed in 100 bins. The enrichment for the 1240k SNP set was 1.0764 ± 0.0167, and for IMPUTE2's INFO score it was 1.0088 ± 0.0003. These results show that variants with |X| > 5.45 are, if anything, higher in genotyping quality than SNPs with |X| < 5.45, providing no evidence for an enrichment in false-positives.

Additionally, variants with low LD scores and MAF (Figure S2.1) seem to show inflated statistics, possibly due to lower imputation accuracy. Do candidate variants show higher (as expected under selection) or lower LD scores than matched variants?

These are important topics that we hadn't addressed properly in the initial submission. In the new Supplementary Information section 2, subsection "Non-normality of the null distribution in GLMM test statistics," we show through simulations that SNP ascertainment in the imputation reference panel, together with the non-linear logit transformation in generalized linear (mixed) models, indeed generates an allele frequency-dependent bias and inflation, which we are not able

to eliminate even using variant QC and other quality control measures in which we tried to be stringent. This is the main reason we do not use traditional methods to define the genome-wide significance threshold. Instead, we adopt an alternative approach, leveraging the rich GWAS dataset to estimate and define the significance level based on FDR. We now discuss these issues explicitly, and provide evidence that this works to address bias.

Related to this, why was the non-replicated signal from Mathieson et al. (rs1979866, Figure S5.1) filtered out?

This SNP showed a discrepancy between the shotgun and 1240k ancient DNA sequences. The following line has been added to Supplementary Information Section 6 in "Re-evaluation of results from Mathieson et al. 2015":

"It showed a 5.4% mismatch (P = 4.8e-8) between the genotypes of 450 individuals with both Shotgun and 1240k sequences. Any variant with an error rate greater than 5% or a P value for error less than 1e-5 is filtered out"

2. Sample QC filters: Can the authors provide details on the number of samples removed based on each QC filter?

In Supplementary Information section 1, subsection "Sample Quality Control," we added a description of the filtering process and the number of samples removed at each QC step, with a summary presented in Table S1.3.

The authors exclude samples with a reported date that deviates from the date predicted from variants most associated with reported dates across all samples. This filter largely depends on the accuracy of this prediction. Could the authors present a plot showing both the reported and estimated dates, highlighting the excluded samples?

In Supplementary Information section 1, subsection "Sample Quality Control," we added Figure S1.7, which highlights samples removed as date outliers.

Also note that in the Supplementary Information Section 1, the authors state that they excluded samples with a date <15,000 years BP or with the absolute value of the regression residual <500 years. I guess they meant the inverse.

Thanks for catching this. This is fixed in the revised Supplementary Information section 1, subsection "Sample Quality Control."

3. Power of the GLMM method: The authors acknowledge that the strong genomic inflation of their χ^2 statistic (i.e., the squared Z-score for the number of standard errors s is from zero) may be due to batch effects, the quality of genotype imputation, data preprocessing, population structure, genetic drift, background selection, or a large number of true positives. Whereas inflation caused by batch effects and imputation accuracy seems to have been largely

addressed through extensive QC filters, the extent to which the K covariance matrix in equation (1) accounts for drift, background selection, and complex selection scenarios remains to be fully evaluated. The authors performed simulations by randomly sampling genotypes constrained by the observed GRM, a given selection coefficient, and initial allele frequency. However, they do not evaluate their FPR under drift, background selection, or positive selection in the source populations only. I strongly suggest that the authors simulate, for example with SLiM, neutral loci, loci under background selection, and loci where a positively selected variant in source populations becomes neutral after admixture, all under a realistic demographic model of west Eurasians (which has been largely established by some authors of this study). Regarding selection in the sources only, which has been addressed by several previous selection studies using ancient genomics data, several of the 347 candidates show trajectories compatible with a temporal change in frequency caused by shifts in ancestry for variants that were already differentiated across ancestries prior to admixture (e.g., candidates 5, 8, 19, 41, 45, 46, 49, etc.).

These is an important set of suggestions, which we have implemented in our revision. We conducted comprehensive forward-in-time simulations of a previously published demography based on European population history (Supplementary Information section 2). With the aim of being biologically realistic, the simulated genomes include three types of regions: coding, functional noncoding, and neutral regions. We tested various models with distinct selection mechanisms and verified that each of our simulations provides a qualitative match to real data in the sense of reducing variation with approximately the correct magnitudes in coding, functional noncoding, and neutral regions (we recapitulate the expected background selection effects). Each model includes an experimental condition with directional selection and a corresponding negative control, differing only by the absence of directional selection. We show that our approach is fully robust to various confounding factors, including population structure, linked purifying selection, sampling bias, and selection in the ancestral population prior to the beginning of our time transect (Supplementary Information section 2, Figures 1b,d, and Extended Data Figures 4,5). These simulations greatly increase confidence in our findings.

Note that in Supplementary Information section 2, the authors mention conducting simulations to estimate a control factor for their method, but they provide very few details. Similarly, they refer to 10,000 simulations for a scenario involving population structure, sample size, and temporal distribution of samples matching real data (Extended Data Figure 14). What demographic model did they assume for these simulations?

We have completely redone this part of the paper, using comprehensive forward-in-time simulations with SLiM which are far more realistic than the simulations we had before. We use these simulations to evaluate the robustness of our approach with respect to potential confounding factors (Supplementary Information section 2, Figures 1c,d, and Extended Data Figures 4,5). This newly added section is a major improvement and increases confidence in the findings.

4. Controlling for FDR: The authors propose an interesting alternative to FWER, by estimating the FDR from the enrichment of variants with significant changes in frequency over time in trait-associated variants. The observed enrichment plateaus at a given significance threshold, suggesting that the enrichment of significant selection signals in GWAS hits, and presumably the proportion of true positives among significant signals, is maximal at this threshold. However, if Pan-UK BioBank GWAS hits include false positives caused by partially controlled population structure, particularly for high-FST variants, the FDR may be underestimated. It is not clear from Karczewski et al., medRχiv 2024 whether residual stratification could affect their results (see their Extended Data Figure 3).

Published literature indicates that the Pan-UKBB GWAS statistics we are using are highly unlikely to include a substantial fraction of false-positive genome-wide associations. The GWAS results apply state-of-the-art mixed models (GLMM or LMM) through the SAIGE package to control for population structure and cryptic relatedness. This framework includes the top 10 principal components as fixed effects and a kinship matrix as a random effect, and Karczewski et al. 2024 demonstrate robust correction for stratification. We also used effect size estimates from the "white British" subset of the Pan-UKBB which is a relatively homogeneous population (Berg et al. 2019), further reducing the risk of false positives from stratification.

Nevertheless, we agree that uncorrected stratification in GWAS if present could have the potential to affect our findings, and to further guard against such concerns, we restricted to a subset of Pan-UKBB GWAS results that behave as one would expect for a GWAS where signals of association were well-calibrated and not affected by residual stratification. Critically, we restricted to 452 out of 7200 phenotypes that passed multiple criteria for a high-quality GWAS, including a well-behaved LDSC regression that has the expected linear form indicating that standard approaches correct fully for stratification.

Can the authors repeat their enrichment analysis using sibling-based GWAS statistics?

In our revision we do use family-based GWAS for some analyses, increasing confidence in results. Most notably, we now use family-based GWAS from Tan et al. medRχiv 2024 to replicate the years of schooling phenotype. This shows that not only in theory but also in practice, uncontrolled population stratification is not explaining the years of schooling result.

But for some analyses, both high sensitivity and specificity are essential, and the UKBB dataset is far superior to family-based GWAS for this. While sibling-based GWAS may offer high specificity, it suffers from severe power limitations due to much smaller sample sizes, reducing sensitivity. We thus continue to report enrichment analyses only for population-based GWAS.

One alternative explanation for the observed plateau is that true positive signals of selection may overlap less with GWAS hits as the selection coefficient increases. Several arguments support this hypothesis: large effects would be expected for large-s variants (assuming a monotonous relation between s and β), whereas GWAS hits typically have low effects on traits.

Additionally, there is widespread evidence for negative selection on complex traits studied by GWAS, and actual oligogenic traits that were targeted by strong directional selection have possibly not been studied by GWAS. Is the enrichment decreasing for $|X| > 7$ (although the data is probably noisier)?

Indeed the data become noisier at the extreme end because only a few loci remain, making the trend less interpretable.

We agree that variants with extraordinarily strong selection coefficients—which are expected to give the strongest signals by our test of selection—could plausibly show different patterns of correlation to complex traits than variants with only strong or modest selection coefficients, for the reasons the referee mentions. Thus, there is no guarantee that if we plot selection coefficient against enrichment for GWAS signals, there will be a plateau. There would plausibly be a decline above a sufficiently high selection coefficient threshold, or alternatively an increase.

But the plateau is not driven by extreme outlier loci like *LCT*, *HLA*, *FADS1,* and *SLC45A2* with extraordinarily strong selection coefficients, but by the much larger number of genome-wide significant signals with strong or modest selection coefficients, down to around s=0.5%. These weaker selection effects dominate our signal as shown by our scatterplot of minor allele frequency against estimated selection coefficient in Figure 2b for 479 genome-wide significant variants. Figure 2b shows a negative correlation between minor allele frequency and strength of effect, which is qualitatively similar to that observed for effects on traits in GWAS studies, and is not characteristic of the extreme signals of selection detected in the first ancient DNA time-transects like *LCT*, *HLA*, and *FADS1, SLC45A2*. This suggests that the computer simulations we have added in our revision, which specify a well-behaved functional coupling between selection coefficients and effects sizes, are a good proxy for the variants driving our plateau. Encouragingly, we observe exactly the plateauing effect in these simulations, and show that we can use it read off a well-calibrated FDR of natural selection for each variant.

5. Independent loci under selection: The authors identify independent loci under selection by considering that all SNPs in LD with the strongest signal of the genomic region reflect the same selective signal, which makes total sense. They considered that two SNPs are in LD if $r^2 > 0.05$ in modern Europeans from the 1000 Genomes Project. However, under a selective sweep, hitchhiked neutral mutations occurring on the selected haplotype may show low r^2 (but high D') with the selected variant. Furthermore, present-day LD levels are probably underestimating past LD levels, as population size in Europe has largely increased over the last ten millennia. Consequently, the authors may overestimate the number of selective events that occurred in West Eurasians over the last 14,000 years. When inspecting the trajectories of the 347 signals, I observed several cases where a relatively rare derived allele, located a few hundred kb apart from a well-known positively selected mutation, increases in frequency tens of generations after the latter. For example, there are nine "independent" signals (signals 31 to 39 in Figures S4.3 and S4.4) within 5 Mb surrounding rs4988235, the selected variant responsible lactase

persistence. Among those, rs62160512 and rs143866305 show D' > 0.7 (in 1KG EUR phase 3) with rs4988235. Another example is the TLR1 locus; the authors found ten "independent" signals (signals 62 to 71, Figure S4.6) within 500 Kb of the selected gene. This includes rs3188469 that shows D' = 1 (in GBR) with rs5743618, the likely target of selection at this locus. I strongly suggest that the authors revisit their LD criteria to define independent selection signals, by conducting simulations of selective sweeps and leveraging LD levels in ancient genomes.

These are excellent comments, we agree with the critiques, and have addressed them in our revision.

Using an r^2 threshold of 0.05 recursively to prune variants and detect independent selection signals is highly effective in most cases. However, indeed there are exceptions in some regions, including the HLA region, the *TYR* locus, and *TLR1*, where multiple hits, all with an $r^2 < 0.05$, persist. This may indicate multiple genuine selective sweeps within the same region or a complex LD pattern possibly driven by a few sweeps, making it difficult to accurately estimate the exact number; in some cases, this could lead to over- or underestimation. To address this, we added an additional pruning step using a different threshold for D′ for the remaining SNPs after r^2-based pruning. A summary of these results is provided in Table S5.1.

6. Re-evaluating previous studies: I commend the authors for their systematic re-evaluation of previous scans for recent positive selection in ancient and modern West Eurasians. Nevertheless, the authors often make the strong claim that most previous signals not confirmed by their analyses are false positives. While this assertion may largely hold true, given the much larger sample size used in this study, I suggest that the authors provide more nuanced explanations: the variant QC filters applied may be too stringent (see point 1), and their GLMM method may miss true signals of fluctuating selection, such as those associated with OCA2 and TYK2. Prior selection studies have searched for selection that can start at different epochs within the last ten millennia, both on de novo or standing variation, allowing for the detection of fluctuating selection.

We have clarified this by adding the clause:

"**and in a few cases due to failing quality control or due to fluctuating selection in space or time**"

7. Evidence for polygenic selection: The authors report "12 traits with significant signals from all three tests after correction for number of traits tested". However, they should clearly state how many of these traits show replication when using Biobank Japan statistics and sibling-based statistics.

Now we clarify in the caption of Extended Data Figure 13 that:

"Years of schooling is the only trait that reaches Bonferroni-corrected significance threshold for all three tests."

We also mention in the caption of the Extended Data Figure 14 that:

"For direct genetic effects, educational attainment is the only trait with all three tests nominally significant (P < 0.05)."

Furthermore, they should clarify why Extended Data Figure 13 shows that the 95% confidence intervals for γ, γ_sign and r_s include zero for most of the traits, using GWAS statistics for both unrelated people and siblings.

We now state in the caption of Extended Data Figure 13 (previously 11) that:

"This comes at the cost of reduced power relative to using European GWAS directly, driven by the smaller sample sizes of East Asian GWAS (median 30% of European GWAS), differences in LD structure between populations, and divergence in genetic architecture (allele frequencies and causal effect sizes)[75–77]. Despite reduced power, Pearson's correlations between Z-scores for European and East Asian GWAS are 0.43, 0.75, and 0.81 for γ, γ_{sign}, and r_s, respectively, a positive correlation that must reflect real selection and would not be expected due to population structure."

We state in the caption of the revised Extended Data Figure 14 (previously 13) that:

"The fifth column shows the sample size, with the median sample size for family GWAS ~13,000 and for population GWAS ~31,000. Both are only a fraction of the UK Biobank cohort of ~490,000 individuals of non-Finnish European ancestry[286], explaining the reduced power and why confidence intervals are wide and often overlap zero."

Finally, the significance thresholds should be displayed in Extended Data Figure 11.

We added the Bonferroni-corrected significance threshold to the figure, which is now Extended Data Figure 13.

As previously shown by some authors of this study, signals of polygenic selection can be confounded by uncorrected population stratification. The authors address this issue by testing the association between PRS and time using -1 and 1 as weights instead of effect sizes in the PRS calculation. However, I wonder whether the results shown in Extended Data Figure 10 could be affected by stratification, which might explain the unexpectedly polygenic signal observed for skin pigmentation.

Residual structure in GWAS effect size estimates can indeed produce spurious trends in PGS over time that reflect ancestry shifts rather than true selection. This is clearly seen in the case of height, the phenotype analyzed in Berg et al. (2019) and Sohail et al. (2019), where ordinary linear regression without proper control for structure yields an inflated gamma signal ($P = 7 \times 10^{-}$

[159]). In contrast, the signal disappears entirely when using a linear mixed model (LMM) (P = 0.21). We have added forward-in-time simulations demonstrating that our polygenic selection statistic effectively removes the effects of residual population structure, resulting in very few false positives (Table S2.1). We also repeated all tests with the γ_{sign} statistic, which is less sensitive to residual stratification.

When selection statistics are confounded with population structure along the same axis as the GWAS, even LDSC may fail to correct for stratification (Berg et al. 2019; Sohail et al. 2019). Our test accounts for structure by incorporating a genetic relationship matrix as a random effect, unlike previous methods such as SDS, which lack any control for stratification. Indeed we show that some of the signals reported in those previous studies (including SDS) are likely artifacts of uncorrected structure (see Supplementary Information section 6). Moreover, our analyses are based on GWAS from the "white British" subset of the UK Biobank, a relatively homogeneous population analyzed using mixed models that robustly control for stratification (Berg et al. 2019; Sohail et al. 2019).

Finally, we replicated the signal of polygenic selection for the trait that today manifests as years of schooling, using two analyses that are essentially immune to confounding due to population structure. The first is a cross-ethnic replication, correlating estimated selection coefficients in West Eurasians to GWAS effect sizes measured in East Asians with uncorrelated population structure. The second is correlating estimated selection coefficients in West Eurasians to GWAS effect sizes measured in family-based studies which are not affected by population structure.

Pearson's correlations between Z-scores for European and East Asian GWAS are 0.43, 0.75, and 0.81 for γ, γ_{sign}, and r_s, respectively (Extended Data Figure 13), providing strong reassurance that our signals are unlikely to be driven by population structure. The Pearson's correlations between Z-scores from population GWAS and direct genetic effect GWAS are 0.51, 0.51, and 0.76 for γ, γ_{sign}, and r_s, respectively. When restricted to population GWAS signals with nominal significance ($P < 0.05$), the correlations increase to 0.84, 0.78, and 0.81, respectively, providing further reassurance that our signals are unlikely to be driven by environmental confounding.

We have added all of this content to the revised main text.

Notably, the authors state in the figure legend that the plot shows "the correlation of a trait with selection summary statistics (r_s)", whereas my understanding is that r_s represents the correlation between s and β.

The wording was confusing. By definition, genetic correlation is the correlation between the true, unobserved effect size vectors across traits (Bulik-Sullivan et al. 2015). Since these quantities cannot be directly observed, the standard practice is to estimate them from GWAS summary statistics. We have revised both the figure legend and the main text to be precise about this.

Minor comments:

- L. 2-3: The authors state that they developed a method "that leverages an opportunity not utilized in previous scans: testing for a consistent trend in allele frequency change over time". This is not the case, as Kerner et al. and Irving-Pease et al. estimated s from the full allele trajectory, using ABC and CLUES, respectively.

We agree and have modified the text to reflect this.

- L. 15-17: The authors interpret the significant decrease in the PRS for body fat percentage, waist circumference and waist-to-hip ratio as supportive of the "Thrifty Genotype" hypothesis (Neel JV, 1962). This is problematic, because there is evidence that the transition to food production was not a period of "plenty", but was characterized instead by nutritional deficiency, reflected by the prevalence of low stature and porotic hyperostosis in early European farmers (Cohen & Armelagos, 1984; Macintosh et al., PLoS One 2016; Theodorakopoulou et al., Arch Balk Med 2020; Marciniak et al., PNAS 2022). This observation disagrees with the authors' results, as the period when negative selection against higher body fat percentage was the strongest corresponds to the period when food production emerged in Europe, ~8,000 years ago (Figure 4).

The referee is misinterpreting the plots in Figure 4 as suggesting that the strongest selection against body fat occurred in the period when food production emerged in Europe ~8000 years ago. In fact, this apparent signal is an artifact of population structure, and the large drop from hunter-gatherers to farmers cannot be interpreted without accounting for ancestry differences. Extended Data Figure 9 shows the time-varying selection coefficient after controlling for population structure and indicates that selective pressure has been consistent from the Neolithic through the Bronze Age and into the Iron Age and later periods.

Note also that the hypothesis is not correctly formulated in L. 344-345 ("genetic adaptation to store fat in times of plenty, became deleterious after the transition to food-production"). The "Thrifty Genotype" hypothesis proposes the inverse: genetic adaptation to store fat in times of food scarcity, became deleterious after the transition to food-production.

We think we are citing the "Thrifty Genes" hypothesis correctly. From the Neel 1962 paper: "In this essay an hypothesis has been advanced which envisions diabetes mellitus as an untoward aspect of a "thriftiness" genotype which is less of an asset now than in the feast-or-famine days of hunting and gathering cultures." The referee is correct in identifying this as a potentially confusing issue, however, and to avoid this, have revised the main text as follows:

"Type 2 diabetes risk factors give compelling signals of negative selection. Thus, we observe negative selection on combinations of alleles that today increase body fat percentage (γ = -1.04 ± 0.13), waist circumference (γ = -0.91 ± 0.13), and waist-to-hip ratio (γ = -0.76 ± 0.13), as might be expected from the "Thrifty Genes" hypothesis[69] that a genetic adaptation to

store fat to promote survival during periods of scarcity in hunter-gatherers, became deleterious after the transition to food-production (Figure 4). Skeletal evidence for nutritional stresses in early European farmers—such as an increased prevalence of low stature and porotic hyperostosis[70–73]—could be viewed as in tension with this hypothesis, but is consistent with it if a farming lifestyle was associated with famines that occurred on too-infrequent a time scale (e.g. every few years) to be effectively buffered by higher fat."

- L. 54-55: When the authors state that "changes in frequency due to selection are often less than what can be expected from random genetic drift", they somehow imply that their method is able to detect selection even when its effects are weaker than those of genetic drift, but they provide no such evidence. They should clarify further why the GLMM approach can outperform other approaches, at equal sample size.

GLMM is not necessarily outperforming other approaches. For example, in a panmictic population with only two time points and no intermediate samples, the GLMM may not perform better than other methods, since much of the signal can be absorbed into the variance component.

In our revision, we emphasize that a strength of the GLMM is that it does not require explicit modeling of ancestry, allowing the inclusion of more samples and thereby increasing power. Instead of estimating the ancestral composition of each individual—which is difficult, introduces noise, and increases degrees of freedom—the GLMM accounts for relatedness through the genetic relationship matrix. This provides robust control for structure without prior demographic assumptions.

Sample heterogeneity, often seen as a limitation, can actually in this setting increase power; we have come to understand this more clearly in the last year of work, and we explain this in our revision. If selection pressures are broadly consistent across space and time, heterogeneity provides natural replicates of the evolutionary experiment, enabling the GLMM to integrate signals across datasets with fewer degrees of freedom. Mixed models are well suited to this type of variation, with heterogeneity modeled through random intercepts. While this assumption does not capture cases where selection differs strongly across space (e.g., the blue eye variant at *OCA2/HERC2*) or time (e.g., *TYK2*, *HLA-DRB1*, *HFE*; Figure 3 and Extended Data Figure 7), it allows efficient detection of selection when pressures are more uniform.

Finally, the GLMM mitigates the variance inflation and power loss that occur in generalized linear models (GLMs) when principal components are included as covariates. Because PCs are often collinear with time, adding more PCs inflates the variance of the estimated selection coefficient, reducing power. By modeling structure through a random effect, the GLMM both controls for stratification and minimizes power loss (Extended Data Figures 16, 17).

We have clarified all these points in our revision.

- L. 191: The authors decided to focus on 36 candidate variants, out of 347 signals, based on the fact that these candidates were of "particular interest". These include 7 variants with

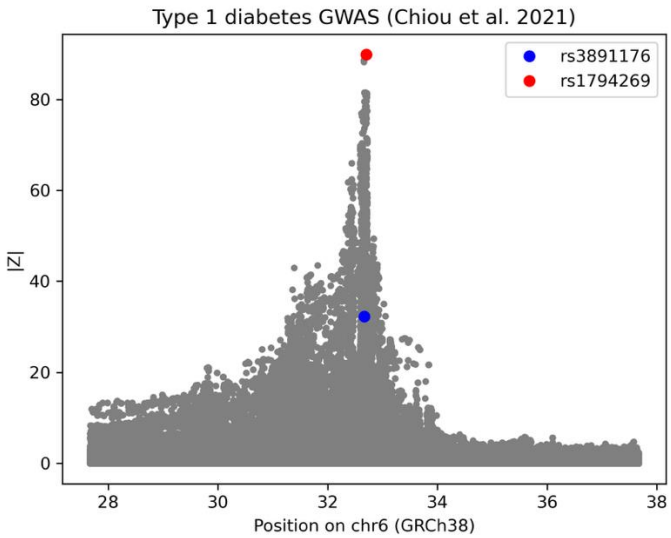
64% < π < 98%, which could be false positives. Instead, it would be informative to comment the more confident candidate variants (π > 99%) with, for example, the largest GWAS effect sizes on traits.

In follow-up work, we are doing exactly what the referee suggests, curating a set of large-effect variants and analyzing their evolutionary history.

However, for our main text revision, we chose to keep to our approach of showcasing 36 variants that are of particular interest because they address outstanding discussion in the literature, including a few that are probably signals but not genome-wide significant. The present paper is already very complicated, and we felt that this was a way to make the results accessible. The interested reader can refer to the full presentation of 479 genome-wide significant signals in the supplement, or the AGES browser we make available, in order to explore the data in other ways.

- L. 201: rs3891176 is also strongly associated with type 1 diabetes (P = 8.7 × 10-229; Chiou et al., Nature 2019)

rs3891176 is highly pleiotropic, with associations to 135 traits in the Pan-UKBB, including type 1 diabetes. Its strongest effect size is on celiac disease (z = 47.6, β = 2.1), where it is among the top associated variants and colocalizes with the selection signal. The lead variant for type 1 diabetes identified by Chiou et al. (2021, PMID: 34012112) is rs1794269 (z = 89.9, β = 1.60), whereas rs3891176 shows a weaker association (z = 32.3, β = 0.60). The two SNPs are only weakly correlated (D' = 0.18, r^2 = 0.01), providing little evidence that rs3891176 tags the main causal haplotype for type 1 diabetes at this locus (See the figure below). For this reason, we do not emphasize type 1 diabetes in the manuscript.



- L. 212: Please cite the phenotypes most associated with A and B alleles.

In Extended Data Figure 7, we present the full list of traits associated with either allele A or allele B after applying Bonferroni correction.

- L. 269: The authors do not test for balancing selection, which would result in a non-significant test. The trajectory of ΔF508 could be compatible with balancing selection.
Perhaps the authors can describe here the independent signal they found in the 5'UTR of CFTR (candidate 199).

The reviewer is correct that there could be balancing selection on ΔF508 which our test is not designed to detect, and we state this explicitly in our revised manuscript:

"The major risk allele for this recessive disease in Europeans[54,55] has been hypothesized to be an example of heterozygote advantage due to conferring resistance to cholera in carriers[56], keeping its frequency substantial despite its strong deleterious effects in homozygotes. However, we find no evidence of directional selection over the time frame that that cholera was likely endemic in West Eurasia (π<1%), with the earliest direct observation at ~2200 years ago in Great Britain and the earliest imputed one ~10100 years ago in Anatolia. Another explanation is needed for the allele's persistence, plausibly by balancing selection."

As the reviewer notes, we do identify independent candidate signals at the *CFTR* locus (tagged by intronic variant rs2237725) and at the nearby gene *ASZ1* (tagged by intronic variant rs7779160, an eQTL and sQTL for *CFTR* in GTEx) (Supplementary Information section 5), suggesting that this locus has been shaped by complex selection pressures. However, we do not include this discussion in the paragraph about CFTR ΔF508 in the main text, to keep the paper as readable as possible.

- L. 383-384: The significant genetic correlations between traits showing evidence of directional selection are puzzling. The authors may consider this is beyond the scope of this study, but one possibility is to follow the approach in Stern et al., Am J Hum Genet 2021, to disentangle polygenic adaptation on correlated traits, assuming a multivariate version of the Lande approximation and leveraging their GLMM-based s estimates.

This is an excellent suggestion for future work.

- L. 402-404: The authors state that "there have in fact been changes in selection pressures at a number of the loci […] (Extended Data Figure 6), including at HLA-DRB1, TYK2 and HFE". It is unclear why they have assumed fluctuating selection for some genes and not for others, how many genes were tested for fluctuating selection, and if they reported only the significant results. This should be clarified.

Detecting heterogeneous or fluctuating selection across space and time is a natural extension of our methodology, and we leave genome-wide scans of such signals to future studies. We manually tested for evidence of changes in selection pressure over time for *HLA-DRB1*, *TYK2*,

and *HFE* because these loci are clinically important, previously reported to be under selection, yet they do not emerge as significant signals in our framework. This outcome is consistent with the limitations of our current model, which assumes constant selective pressure across space and time. We modified the relevant sentence to explain this as follows:

"By sliding a 2000-year window through our time transect and re-estimating selection coefficients within each window not only for all genome-wide significant loci (Supplementary Information section 5) but also a few manually chosen loci not genome-wide significant but clinically important and previously reported to be selected (Extended Data Figure 7, Figure 3), we can see that there have in fact been changes in selection pressures at a number of the loci we analyze, including at *HLA-DRB1, TYK2* and *HFE*."

- L. 408: The authors state that "selection coefficients on the alleles must have shifted over time" because the estimated age of favored alleles in a selective sweep is lower than the date of origin of the mutation inferred from RELATE. However, this observation is true in absence of fluctuating selection, as RELATE assumes neutrality and will systematically overestimate the age of a selected mutation (Speidel et al., Nat Genet 2019).

Although RELATE is expected to overestimate the age of selected alleles, we focus only on alleles under selection and do not compare their ages to neutral cases. Our analysis emphasizes qualitative patterns within selected variants, where we observe that RELATE estimates can be one to two orders of magnitude older than the expected sweep age, while in other cases they align more closely with the sweep age. We added the following clarification to the Discussion:

"By comparing the estimated age of the mutation that contributed each selected allele[9], to the extrapolated time to reach fixation given its estimated *s*-value, we find that about one third of the favored mutations are ancestral alleles and another third are derived alleles with true ages an order of magnitude older than the expected sweep age, indicating that selection coefficients on these combined two-thirds of alleles must have shifted over time (Figure 2c). The remaining up to a third of the signals we detect are consistent with classic sweeps where the newly arising variants have been under consistent positive selection since they arose (in all these cases, the mutations arose recently enough that the sweeps are not complete)."

- L. 450-451: The linear mixed model was also already used in the context of genomic scans of natural selection (Yang et al., PNAS 2017).

Thank you for bringing this study to our attention. We have modified the text and added a reference to Yang et al. 2017.

- L. 460: The authors justify the choice of the GLMM method by the fact that, "in practice, it detected many loci". These may be false positives.

In the revised manuscript we added a forward-in-time simulation analysis to validate our methodology, which substantially increases confidence that the signals of directional selection we detect are overwhelmingly real. We deleted the clause "in practice, it detected many loci".

- L. 512-513: the authors have grouped individuals into clusters of individuals with similar ancestry and coming from similar times, to make analysis tractable. It would be useful to show in a Suppl. Figure how the clustering affects |X| for a few significant signals.

To calculate X without clustering, we would need to run it on the whole genome and calibrate it using enrichment patterns from GWAS studies. This is not computationally feasible, as now we mention in the online methods:

"Despite an 82-fold speed increase, running a GLMM on ~20,000 individuals for ~9.7 million variants was infeasible (~6500 CPU years)."

- L. 543: Please clarify if variants used to compute the GRM were filtered on MAF, and in which sample of individuals f_k is computed.

We created a GRM using all autosomal SNPs that passed quality control with MAF > 0 across all individuals. f_k refers to the allele frequency at these SNPs calculated across all individuals. We have edited the text to clarify these points.

- L. 554: Please replace "proces s" by "process".

We corrected this typo.

- L. 560-561: Please clarify why y_i was scaled by the standard error of PGS of the modern samples.

We edited the text as follows:

"Here, y_i is the polygenic score of the sample i, centered at zero and scaled by the standard error of PGS of the modern samples to make it interpretable as the strength of polygenic selection and comparable across different traits"

- L. 630: What makes the difference between the 56 and remaining 244 newly reported shotgun genomes that sum to 300?

For the former genomes, no data from them have previously unpublished, and we release the shotgun data for them in an anonymized form (without any linkage to skeletal codes or other metadata except a point estimate of their date and broad region within West Eurasia). The latter samples are represented in the previous literature by capture data, and in this study, we report shotgun sequences for them (without any anonymization).

- L. 663: Showing a PCA identifying modern and ancient genomes, as well as sampling regions, would be very useful, in addition to Figure S1.2.

We have added the PCA plot in Figure S2.7. Approximate geographic regions, grouped into five clusters together with abundance and date distributions, are shown in Extended Data Figures 1a and 1b.

- L. 642/683: The authors refer to either 5,227 or 5,382 newly reported ancient DNA libraries. Please clarify.

In the originally submitted version of the manuscript, we explained this in Online Table 3:

"Online Table 3: List of 5382 newly reported ancient DNA libraries. The majority (n=5227) are from 4471 never-before-reported ancient individuals; the rest (n=155) are from 74 individuals for which we increase data quality. Available as an Excel table."

In the current version where we have now increased sample size, and we updated this to:

"Supplementary Table 3: 18439 ancient DNA libraries with newly reported capture data. The majority (n=18007) are from 9487 never-before-reported ancient individuals; the rest (n=432) are from 181 individuals for which we increase data quality. Available as an Excel table."

- Figure 1c: Please report as a Table.

We believe the embedded table is integral to the intellectual content of this display item and that the information is best communicated in the current format. We respect the journal's formatting guidelines and editorial process, and we will make changes if required by the editor.

- Figure 2c: Please replace "Dervied" by "Derived".

This typo is corrected.

- Extended Data Figure 1: An admixture plot of the genomic transect used would be very useful.

PCA and ADMIXTURE plots are alternative approaches to summarizing population structure, and we chose PCA because it is assumption-free and conveys the major axes of genetic variation across the transect (Figure S2.7, Extended Data Figures 1a,b). We believe that the PCA representation together with our modeling framework most effectively captures the relevant structure, so have chosen to retain only a PCA-based representation of the variation.

- Extended Data Figure 3b: In the legend, the authors indicate that s = 0.01, but they show curves for variants under neutrality and negative selection as well.

We clarified this in the caption:

"Simulating neutral, background, and positive selection for a 200 kb window around a focal SNP, with derived allele frequency drawn uniformly from [0,1]. Under positive selection, the focal SNP has a selection coefficient of s = 0.01 for the favored allele. Under

background selection, 20% of mutations are drawn from a gamma distribution with mean -0.03 and shape parameter 0.206, and the remainder are neutral. The population size is constant at 20000 diploid individuals, mutation rate per base pair per generation is $2x10^{-8}$, and recombination rate is 1 cM per 1 Mb."

- Extended Data Figure 5: Please define horizontal lines. Clarify in the legend why modern genomes were excluded.

We have updated the captions of Figure 7 (previously Figure 5) and Figure 10 (previously Figure 8) in the revised version of the paper to address these concerns.

- Extended Data Figure 8: It may be interesting to indicate in the main text that s tends towards 0 for traits 10-12.

We agree. We added the following line:

"The signals are largely driven by selection before ~2000 BP, after which γ tends toward zero (Extended Data Figure 10)."

Referee #2 (Remarks to the Author):

This manuscript presents a new method for detecting directional selection from time-series data and reported "pervasive findings of directional selection", using an expanded European ancient DNA dataset. The authors additionally report new signals of polygenic selection.

Unfortunately, there are major concerns about the statistical machinery and interpretations. It remains unclear whether the pervasive findings of directional selection (both locus-specific or polygenic) are genuine or are artifacts of biases from failing to adequately correct for confounders and/or using a loose significance cutoff. As it stands, the paper makes strong claims without adequate support.

We agree about the importance of rigorous validation of our new methodology. In our revision, we added comprehensive forward-in-time simulations using SLiM and used them to extensively evaluate the robustness of our approach with respect to potential confounding factors. None of a range of simulated processes that are not directional selection—including stabilizing selection and purifying/background selection—can produce the correlation of selection signals to GWAS signals and HAF score that we observe in real data (Supplementary Information section 2, Figures 1b,d, and Extended Data Figures 4,5). In contrast, when we simulate directional selection, we observe exactly the empirically observed pattern, and show that we can use the enrichment to identify a stringent threshold for statistical significance. These analyses greatly increase confidence in our claims.

Major comments:

This manuscript used a GLMM which is motivated by the LMM commonly used in GWAS to correct for potential population structure. However, it is not immediately clear from theory why this should be well-calibrated and simulation verification is critically needed. The simulations as performed (L610) are inadequate as they are based entirely on the assumed model rather than probing the complexities of real population genetic scenarios.

It's essential to evaluate the following: (1) whether the GLMM works for a no-gene-flow case, with a well-calibrated p-value for a few Ne values (forward-in-time simulations); (2) whether the GLMM works for a realistic scenario of complex European history, population movement and gene-flow; (3) what is the impact of linked selection from background selection? What impact do these scenarios have on the distribution of test statistics?

I'm perplexed that the central output of the analysis is to model and report frequency changes as a function of time. It is well known – thanks in large part to the Reich lab -- that there have been major ancestry turnovers in European history. These have the potential to drive significant allele

frequency changes. For example, if the WHG, Farmers or Steppe populations had markedly different frequencies at any given locus, the model could interpret this as a rapid selective shift. Browsing Figure 3 suggests several cases where this may be the correct interpretation. In some cases, outlier variants between populations may indeed result from past selection, but the time and place of that selection would be far removed from what is reported by the model. I view it as essential to recast the model in a more biologically motivated manner.

I am troubled by how heavily the paper relies on the overlap with GWAS loci to set the significance threshold. Genome regions containing GWAS hits differ from random loci by many measures (eg gene constraint and conservation, gene density, measures of chromatin landscape, predicted background selection, etc). Without really understanding the properties of the GLMM very well, how sure are we that these features aren't correlated with the probability of large values of X? I'd worry particularly about background selection as it's entirely unclear if sigma_j can adequately capture this effect.

These are very important issues, which we indeed hadn't addressed adequately in our original submission, and which we have spent the last year addressing. We have now conducted comprehensive forward-in-time simulations to model a demography based on European history (Supplementary Information section 2). The genome contains three genomic subsets: coding, functional noncoding, and neutral regions. We tested various models with distinct selection mechanisms to simulate background selection and its effect on neutral diversity within each genomic compartment. Each model includes an experimental condition with directional selection and a corresponding negative control, differing only by the absence of directional selection. We show that our approach is fully robust to various confounding factors, including population structure, linked purifying selection, sampling bias, and old selection in the ancestral population (Supplementary Information section 2, Figures 1b,d, and Extended Data Figures 4,5). In the simulations, we observe enrichment in GWAS signals and HAF score in association with directional selection, but not a hint of such enrichment in simulations of all the other confounding factors we simulated. These simulations greatly increase confidence in our findings and we believe they address the important issues the referee raised.

Moreover, the authors don't really spell out exactly why they expect the overlap with GWAS variants to be a reliable metric for their purpose. This might make more sense to me if we thought these traits were the actual targets of selection, but instead the UKB traits provide an eclectic sample of a small part of the possible trait space. Presumably the authors have in mind some implicit hypothesis that there is enough pleiotropy between the measured traits and the targets of selection.

For the enrichment analysis in GWAS signals to be a valid way to calibrate genome-wide significance of our statistic for detecting natural selection, it is NOT necessary for us to be analyzing GWAS that comprehensively sample trait space. Even a single GWAS including a

large enough number of signals of association to build up a curve would be valid. The reason this is so is that alleles that are genuine targets of directional selection are by definition biologically functional, and functional alleles will also have an enhanced probability of showing signals of association in GWAS compared to random (and thus largely non-functional) alleles in the genome. It is true that the degree of enrichment will vary depending on the type of GWAS and the trait: for different traits, the strength of coupling of selection strength to effect sizes can be expected to differ dramatically. But this variation is expected to affect only the magnitude of the y-axis in the enrichment plot. It should not affect where on x-axis enrichment begins to occur and where it plateaus (that is, the point at which negligible false-positives are expected).

We use 452 Pan-UKBB GWAS for the enrichment analysis because they provide us with a large set of alleles that are highly likely to be functional and that we can use to build up a well-measured enrichment curve.

But there's also a second problem here: presumably most of these GWAS hits have quite small effects on specific traits--perhaps explaining 1/1000th of the variance. Why should these have selection coefficients of ~1%?

We believe that one of the reasons why our enrichment approach works is pleiotropy. For any given GWAS, it is true that the effect sizes of the most alleles are almost certainly small with selection coefficients associated with the trait being explored with GWAS that are too small for us to detect at our significance threshold. But the right way to think of our enrichment strategy is not on the single GWAS level. If the selection was for the trait being studied in that GWAS, most of the alleles would indeed be expected to have small associated selection coefficients and to be difficult to detect in our scan. But single-variant GLMM identifies selection at the variant level, without any knowledge about the actual trait under selection. Because of pleiotropy, a selected variant can appear in GWAS of many traits where it will show small effect sizes, but the selection that matters will be for a trait where the effect size is large. For example, the *LCT* variant rs4988235 is strongly selected but associated with 19 traits in PAN-UKBB, with effect sizes of only 1-2% for most traits except "Never have milk" (Pheno ID 101 in AGES; pheno code 1418 in UKBB), which has an effect size of 9%. The strong effect size traits presumably drive the selection.

Another issue to keep in mind in connection with the referee's comment is that there are vastly more genuine signals of directional selection than we have power to detect at genome-wide significance, so the ones with selection coefficients of >0.5% that we have power to detect are only the proverbial tip of the iceberg. As a concrete example of this, in Figure 1d and Extended Data Figure 5a, the strongest single-variant signals of selection are enriched for immune traits, but not for behavioral or brain traits. By contrast, Figure 4 shows clear polygenic signals for behavioral and brain traits. This is not contradictory: Extended Data Figure 6a shows that variants with weaker selection signals are enriched for these traits. Polygenic tests aggregate effects across many variants, so individually small contributions can sum to a strong signal of

selection, even if individual contributing variants do not appear as genome-wide significant in our scan.

And third, why would we see selection acting on specific complex trait variants and not on other variants that affect the same trait? For example, why does one particular balding variant show up in the data, while there is no polygenic signal for balding?

This is an excellent question. A speculation (which we think is outside of the scope of this study so we do not want to include it in the main text) is that the balding phenotype we analyzed is not the actual trait under polygenic directional selection, even though it is the only PAN-UKBB trait associated with this variant. The variant rs11803731, like many others under selection, is pleiotropic, and the true trait under selection may not be represented in our trait list. For example, rs11803731 explains 6% of the variance in hair morphology ($p = 1.5 \times 10^{-31}$) according to Medland et al. 2009 (PMID: 19896111), which is an unusually large effect size for a complex trait. However, hair morphology is not included in our dataset, and the relevant summary statistics are not available for direct testing.

Identifying the favored mutation and determining the precise trait under selection are both important but challenging problems. Exploring every individual case in detail is beyond the scope of this study. We view this as a critical direction for future research. Further insight into the exact trait(s) under selection is best sought in studies that involve molecular work.

I couldn't tell if HLA is included in this analysis. HLA could have a huge biasing effect in this type of analysis and it would be best to exclude it from all genome-wide analyses and only report HLA results in single-locus analyses.

We are deeply aware of the complications associated with the huge effect sizes in the HLA region. Our standard polygenic analyses are performed excluding the HLA region, and we specify this in the revise main text:

"We used three statistics to test for coordinated selection on alleles affecting the same trait (for our main analyses we excluded variants at the HLA locus as we were interested in polygenic signals, not signals driven by large effect variants which are common at HLA)."

This is also specied in the Online Methods, section "Polygenic score computation":

"We generated four variations of the PGS score by using the GWAS effect values (β_i) or only the sign of the effects (sign(β_i)) as weights, and by including or excluding the HLA region (the results quoted in the main text exclude the HLA region)."

For completeness, we also report the results of the same tests including the HLA region in Supplementary Table 5, and where there are differences these are plausibly explained by large effect sizes at HLA.

The paper claimed the signals arise from directional selection, but directional selection is

not the only mechanism to cause frequency shifts over time. Background selection and stabilizing selection can also change the distributions of allele trajectories of common variants. The effects of background selection cannot be correctly modeled as a simple varying Ne along the genome, as represented by the locus-specific noise term in the GLMM.

In our revised Discussion, we now address these comments.

Hayward and Sella (2022) demonstrate that underdominance due to stabilizing selection in the stochastic phase, following a deterministic directional selection phase triggered by a sudden shift in the trait optimum, leads to non-neutral changes in allele frequencies that gradually push variants toward fixation or loss. However, Equation 7 of the same paper shows that the selection coefficient (s) in the deterministic phase is proportional to the effect size (a), while in the stochastic phase it is proportional to a^2. If we assume a = 0.01, which is a plausible approximation based on typical GWAS effect sizes for common variants, this yields a selection coefficient in the order of s ≈ 1e-4, which is well below the statistical power of our method and is practically indistinguishable from neutrality using this approach.

Figure 3 of the same paper illustrates this process, with the x-axis representing generations on a **logarithmic** scale. While the deterministic phase unfolds over about 100 generations, the stochastic phase spans 10,000 to 100,000 generations, confirming that this underdominance due to stabilizing selection is effectively indistinguishable from neutral evolution in the context of our study which is only on the order of 300-400 generations.

Buffalo and Coop (2019, 2020) show that in a panmictic evolving population, purifying selection can generate autocovariance in allele frequency changes for linked neutral loci over a short period (10–20 generations). This appears as positive covariance between allele frequency changes in adjacent time intervals, which rapidly decays to zero within 10–20 generations.

We lack the temporal resolution to detect such rapidly decaying signals, as 20 generations is in the order of the standard error of estimated dates in ancient DNA. Our study spans 300–400 generations, indicating that the signals we detect exhibit autocovariance on a much longer timescale. Therefore, the patterns we observe cannot be explained by the short-term purifying selection effects described in Buffalo and Coop (2019, 2020).

Moreover, the population analyzed in our study is not panmictic, as in Buffalo and Coop (2019). Instead, on the time scale of 10-20 generations over which the Buffalo and Coop phenomenon occurs, it resembles distinct replicates across geographic regions hardly connected by migration. Buffalo and Coop (2020) demonstrate that convergent correlations across replicates, that is, the correlation between allele frequency changes of two replicates over the same time interval, also decays rapidly after 10–20 generations of evolution from the founder population. Thus, short-term autocovariance in allele frequency changes is even less relevant for the ancient human time transect we study. This is consistent with the findings of Simon and Coop (2024), who analyzed two ancient human time transects and found no genome-wide signal of linked selection.

In addition to adding content in the Discussion about all these issues—building an intuition for readers for why our results make sense—in the revised manuscript, we added comprehensive forward-in-time simulations demonstrating that purifying selection and stabilizing selection do not generate false positives, and that our analysis is robust to their effects (Figures 1b,d, Extended Data Figures 4,5, Supplementary Information section 2).

The paper also reports potentially interesting results on polygenic adaptation. However, it is well known that these tests are highly sensitive to stratification. The quality of GWAS has certainly improved since the 2019 eLife papers on this topic, but this is still a critical concern. One potential issue with replication in a second population sample is that this strategy can be biased if the SNPs are initially ascertained in a stratified sample (https://elifesciences.org/articles/61548).

Family GWAS, on the other hand, is less likely to be biased by stratification (depending on how SNP ascertainment is performed). However, the authors were unable to replicate the signals of polygenic selection using family GWAS. I accept that the sib GWAS may be underpowered, but it may also be a warning sign that the signal is spurious.

One might be particularly worried that traits including education, intelligence, and household income show up so strongly in this analysis. The emerging consensus is both that these traits have relatively low true SNP heritability, and that much of the polygenic score may be driven by stratification. Added to general concerns about portability of PGS across environments (how portable are these PGS to the neolithic?) it's hard not to be at least a little skeptical about these results. Among many possible references:
https://pubmed.ncbi.nlm.nih.gov/38225408/
https://pmc.ncbi.nlm.nih.gov/articles/PMC9627830/
https://www.medrxiv.org/content/10.1101/2024.10.01.24314703v1

The referee raises a serious concern that drove many analyses we did over the course of this project as well as in the last year during our revisions. To address this:

- We applied LMM to the PGS to minimize the effects of population structure.
- We used two sets of PGS, including one based on the sign of effect sizes, which is more robust to portability issues.
- We conducted a genetic correlation test, which partially controls for stratification.
- We validated our results using effect sizes estimated from GWAS data in East Asia and observed signals that are essentially impossible to attribute to population structure (because population structure in West Eurasian and East Asia are largely uncorrelated).
- We also explored family-based GWAS, despite the fact that this approach is highly underpowered due to one to two orders of magnitude lower sample sizes. In the course of our revision we were able to access data from a better family-based GWAS compendium than the one we had available for our original submission (Tan et al. 2024 rather than

Howe et al. 2022). For the years of schooling phenotype, we now observe a replication as a single hypothesis test ($P < 0.05$) for all three tests of polygenic directional selection, and have added this to our revision.

The limited data release policy (L642) is unacceptable. Future researchers studying selection would certainly need access to more detailed meta-data including, but not limited to, geo-locations and basic archaeological context such as putative population group. Beyond enabling effective future use of the data, there's also a serious concern that if some of these data are never reported elsewhere, they could remain in limbo indefinitely, with no meaningful metadata and open-ended publication restrictions. And it's worth noting that the proposed data release policy is literally the opposite of the Fort Lauderdale principles since the point of the Fort Lauderdale agreement was to create a framework for full data release to allow resource users "to use the data in any creative way. There should be no restrictions on the use of the data… In some cases, this might best be done by discussion or coordination with the resource producers".
https://www.genome.gov/Pages/Research/WellcomeReport0303.pdf

In our revision, we have removed the reference to the Fort Lauderdale principles. We also state that there are no restrictions at all on the use of the data we publish. We are publishing the full dataset needed to replicate our analyses and to carry out diverse other analyses.

We do not agree that we are proposing a limited data release. Instead, we are releasing all the data we used for our analyses, with the full permission of third-party archaeologist and anthropologist sample custodians, obtained through hundreds of explicit permission letters covering each and every newly reported sample that we have on file and that we have shared with the editor as documentation of this permission. We think this is an innovative, Open Science approach that will help the entire community. It is similar to the anonymization process common in analyses of human genetic data, where the full data needed to support an analysis and critical re-examination of results are released, with other data reserved for future studies involving authors whose specialty relates to those data aspects.

After discussion with the editor and cognizant of the referee's comments, our revised Data Availability statement includes the following content:

"Aligned sequences for the newly reported data from 10036 ancient individuals are available through the European Nucleotide Archive under an accession number that will be made available upon final publication. This includes complete genetic data for all newly reported individuals alongside point estimates of their dates and geographic categorization into five regions of West Eurasia. Our release of these data without restrictions has the approval of third-party sample custodians who provided letters of permission for an approach of not being authors of the present study but instead of being authors of future studies that should be the references for analyses of their population history; those studies

will provide archaeological contextual information including links to the excavation skeletal codes. Imputed genomes for ancient and modern individuals are available at the Dataverse repository at an accession number that will made available upon final publication."

Minor comments:
It could be helpful to draw the p-value QQ-plot, and check if the inflation is global or just the tail. The latter would support the pervasive findings of directional selection, while the former would support poorly-calibrated statistical test and the selection coefficients cannot really be trusted;

In the newly added Supplementary Information Section 2, subsection "Non-normality of the null distribution in GLMM test statistics", we demonstrate that the null is non-normal. Because of this, a QQ-plot (which relies on an assumption that the null distribution is normal) is not an appropriate diagnostic for calibration, and inflation in the plot would not accurately reflect the reliability of our test statistics.

The LD-score plot that we present in Extended Data Figure 2 and Figure S3.1 provides further evidence that calibrating the selection statistics using a genomic control approach—which is related to inspection of a QQ plot—is not valid. It shows that the variants with low LD-score are associated with a different degree of inflation of statistics than variants with high LD-score. We understand why this occurs and discuss this in our Supplementary Information section 3, but what it means in practice is that we cannot use the bulk of the distribution to calibrate the tail of the distribution. However, we can obtain a well-calibrated statistical test using the methodology we employ (as we validate by simulations).

The word "Derived" is mis-spelled in the x-axis label in Figure 2c;

This typo is corrected.

The HAF-score results (Supplementary Information Section 3) are derived under a very simple case of constant population size, and here it could be varying population size and massive gene-flow. It is hard to justify the correctness of these results;

We have now carried out forward-in-time computer simulations showing how the HAF-scores behave for a realistic demographic history for Europe, and for a variety of scenarios of selection. We show that the rise in HAF scores that we observe is expected only in the case of directional selection (Figure 1d, Extended Data Figure 5, and Supplementary Information section 4), and is not expected for stabilizing or purifying selection. These results increase interpretability of our HAF-score findings, and further increase our confidence in the findings of the paper based on this independent line of evidence.

The authors have inferred allele frequencies of ancestral populations with qpAdm, can the authors check how much frequency change can be attributed to mixture composition change versus outliers which are likely true selection targets? (Supplementary Information Section 6)

We quantified this by showing that only about 2.16 ± 0.11% of allele frequency change can be attributed to directional selection. Nearly all observed variation is explained by other evolutionary processes such as gene flow and genetic drift, consistent with previous observations (Simon and Coop 2024).

The title rather overpromises: even if we take the 347+ hits as correct, it's not yet the case that ancient DNA now "realize(s) the promise" to "elucidate human adaptation", any more than finding hundreds of GWAS hits has elucidated the molecular pathways of complex traits such as schizophrenia.

We believe the title is justified, in a similar way that a title for an early schizophrenia GWAS paper as follows would be appropriate: "Hundreds of common variants associated to schizophrenia realize the promise of GWAS to elucidate psychiatric disease". A title like this does not make any claim that the study provides a comprehensive explanation for a biological process. Rather, it states that the study for the first time reports a large number of loci associated with a process – a number of loci substantial enough to permit quantitative study.

By annotating 9.7 million variable positions in the genome with selection coefficients estimated to around ±0.17%, we provide ancient DNA-based annotation of directional selection's effects at a genome scale, and reveal patterns that do indeed shed light on human adaptation.

Referee #3 (Remarks to the Author):

Summary comments

Using the largest ancient DNA dataset ever assembled for this purpose, the paper claims to have found evidence for an order of magnitude more signals of adaptive responses to natural selection in humans than have been discovered in previous studies. Notably, it does this primarily by claiming to be able to overwrite the existing statistical significance standards of this field via a new enrichment procedure. In my opinion, there is not enough evidence to justify this.

To briefly summarize, the authors used a generalized linear mixed model (GLMM) with a logistic link function to model allele frequency change as a function of time in a sample of over 8,000 ancient individuals. To account for variation in ancestry, they modeled the among-individual covariance structure using a genetic relatedness matrix (GRM) estimated from genome-wide data. The test statistic is heavily inflated relative to the null expectation. The authors argue that this inflation is due to an abundance of real signal and devise a procedure for determining a significance cutoff as a function of the enrichment of GWAS hits among selection signals. I did not find this approach very convincing. The authors also analyzed polygenic scores and reported some signals. I have concerns about whether the efforts to avoid population structure confounding have been successful.

We agree that a limitation of our original submission was insufficient validation of our new methodology. To make our identification of signals of selection convincing, we needed to provide compelling evidence that our approach does not produce false-positives in the face of a range of realistic confounding factors.

In our revision, we added an extensive set of forward-in-time simulations using SLiM and used them to evaluate the robustness of our approach with respect to potential confounding factors. None of a range of simulated processes that are not directional selection—including stabilizing selection and purifying/background selection—can produce the correlation of selection signals to GWAS signals and HAF score we observe in data (Supplementary Information section 2, Figures 1b,d, and Extended Data Figures 4,5). In contrast, when we simulate real directional selection, we observe exactly the empirically observed pattern, and show that we can use the enrichment to identify a valid threshold for genome-wide statistical significance. These new analyses greatly increase confidence in our claims.

Major comments on single locus outlier scan

The authors' GLMM approach is very similar to environmental association methods that control for a GRM, but it uses time as the predictor rather than an environmental variable (e.g., LFMM, Bayenv, etc.). In contrast, previous papers analyzing aDNA time trends have

used only the top principal components of this GRM. The authors argue that the PCA approach suffers from collinearity between the PCs and the time axis, which reduces its power, whereas the GLMM does not have this problem.

However, I think this represents a misunderstanding, or at least a misstatement, of the difference between the two methods. The GLMM method can be viewed as a model that includes all PCs, but with a prior on the magnitude of the contribution from each PC that is proportional to the eigenvalue associated with it, and a single common scale of variance shared across all PCs (see, e.g., https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0075707, or https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3002847). In this analysis, the GLMM primarily uses the scale of variation observed on "lower" PCs not correlated with time to determine how much variation there "should" be along the time axis. Therefore, the question is not whether including the kinship matrix in the GLMM results in the inclusion of covariates collinear with the time axis (it surely does), but whether the GLMM correctly estimates how much variation there should be along these axes in the absence of selection.

These are important comments, which we have addressed thoroughly in our revision, both through providing better intuition regarding the value the GLMM is adding, and also, and more importantly, through simulations that document its added value.

We now demonstrate, using simulations, the difference in statistical power between a GLMM and a simple GLM that uses PCs as covariates. We also show why the GLM is underpowered in such cases due to collinearity (Extended Data Figures 16 and 17; Supplementary Information section 2).

Additionally, we modify the GLMM compared to our original submission so that in our joint estimation of σ^2 and s, we constrain the σ^2 parameter to be no smaller than its mean value under the null for the corresponding MAF bin (i.e., from a mixed model where time is excluded as a covariate). This ensures that the model accounts for genetic drift at least at the genome-wide rate or higher, which is a technical improvement (qualitative results are unchanged).

The authors suggest they performed simulations of selection to compare the two approaches (e.g., Extended Data Figure 14). However, based on the descriptions in the Methods, it appears they simulated data from the GLMM itself (with a time trend to simulate selection) and then compared the GLMM that generated the data against the PCA-based method. If this is correct, it is not surprising that the GLMM performs better, as it is the data-generating model. If this is wrong, the simulations need to be explained more clearly.

We fully agree with this suggestion, and have implemented it in our revision. We conducted comprehensive forward-in-time simulations in SLiM to model a demography based on European history, and also simulated realistic variation in functional constraint across loci in the genome

(Supplementary Information section 2). We show why the GLM is underpowered in such cases due to collinearity (Extended Data Figures 16 and 17; Supplementary Information section 2).

It is clear, given the extreme inflation of the test statistic, that the GLMM does not fit the real data well. The question is whether this is 1) because it is a poor null model, or 2) because there is rampant adaptation, as the authors suggest.

It is reasonable to suspect *a priori* that our detection of a large number of genome-wide signals of selection where previous groups detected few, is most likely reflecting false-positives due to methodological artifacts. This is what we ourselves thought was likely when we began to observe these signals, and we required extraordinarily strong evidence to begin to believe these findings. As we demonstrate in our revision, multiple independent lines of evidence, now validated by simulation, indicate that we are detecting an order of magnitude more signals of selection than previous ancient DNA time transect studies (Figure 1, Extended Data Figures 4,5,6).

In the new Supplementary Information section 2, subsection "Non-normality of the null distribution in GLMM test statistics," we show through simulations that SNP ascertainment in the imputation reference panel, together with the non-linear logit transformation in generalized linear (mixed) models, generates an allele frequency-dependent bias and inflation. This is the main reason we do not use traditional methods to define the genome-wide significance threshold. Instead, we adopt an alternative approach, leveraging the GWAS data to estimate and define the significance level based on FDR. We show through our new simulations that this produces a valid calibration, with not a hint of enrichment in GWAS signals (or HAF score) unless there is genuine directional selection.

1) It is difficult to assess how poor the GLMM is as a model because the authors do not appear to have tested it in any real evolutionary simulations.

One possible issue with the GRM is the influence of background selection. The authors are cognizant of this potential confounder and try to address it, but I am not sure I am convinced. First, it is not clear that independently estimating sigma for each SNP is sufficient. The effects of background selection take multiple generations to accumulate and thus more strongly impact comparisons among more divergent groups than among less divergent groups.
Because the GLMM effectively uses the variance on the lower axes to estimate how much variance there should be on the time axis (see above), it could
underestimate the variance on the time axis in the presence of background selection.

I recognize that realistic simulations of the impact of background selection in this context would be challenging, but given that much of the GWAS enrichment signal appears to be confounded with variation in the strength of background selection (Extended Data Figure 3a; see also comments below), it seems an important and still unresolved issue.

We agree about the importance of more rigorous evaluation of our new methodology. It has taken us a year to resubmit the manuscript, in large part because we carried out the simulation series this referee and the other two referees suggested (and we agree) were important in order to establish sufficient confidence in our methodology. Our qualitative findings are unchanged, but we have increased confidence in our results.

In our revision, we added comprehensive forward-in-time simulations using SLiM and extensively evaluated the robustness of our approach with respect to potential confounding factors. We show that none of a range of simulated processes that are not directional selection—including stabilizing selection and purifying/background selection—can produce even a hint of the correlation of selection signals to GWAS signals we observe in data (Supplementary Information section 2, Figures 1b,d, and Extended Data Figures 4,5). In contrast, when we simulate real directional selection, we observe exactly the empirical pattern, and can use the enrichment to identify a valid threshold for genome-wide statistical significance. This is a major improvement to our study, and it makes the results even more compelling.

Also, for the record, I did not find the HAF score analysis helpful. HAF scores are a measure of frequency change and will necessarily be large given that these loci were identified by having a significant correlation with time. Therefore, there is no
new information. The simulations in the supplement about the distribution of HAF scores do not deal with this ascertainment bias and are not convincing.

We have now carried out forward-in-time computer simulations showing how the HAF-scores behave for a realistic demographic history for Europe, and for a variety of scenarios of selection. We show that the rise in HAF scores that we observe is expected only in the case of directional selection, and is not expected for stabilizing or purifying selection (Figures 1c,d, Extended Data Figure 5, and Supplementary Information sections 2,4). These results greatly increase interpretability of our HAF-score findings, and further increase our confidence in the findings of the paper based on this independent line of evidence.

2) The authors' preferred conclusion is that the test statistic is inflated because there is an abundance of true signals, and thus applying significance thresholds used in previous studies would be too conservative. Notably, if the authors apply a standard genomic control correction, followed by a standard genome-wide significance threshold of $p < 5e\text{-}8$, they find 48 signals, consistent with previous studies. Therefore, the order of magnitude increase in the number of signals reported follows entirely from the authors' conclusion that this threshold is too conservative.

This conclusion is justified by estimating a false discovery rate (FDR) as a function
of the enrichment of selection signals among GWAS hits from UKBB traits.
However, the authors need to clarify their FDR estimation procedure. The supplemental material contains some underexplained math, and it is unclear how this math relates to the curve-fitting

procedure used to determine the FDR. Is the fact that the FDR plateaus at approximately 1 actually data-driven, or does it simply follow from the assumption that once the enrichment curve levels off, the FDR must be close to 1, as suggested by the verbal model in the main text? Furthermore, is the idea that the FDR is approximately 1 once the enrichment levels off consistent with the observation that the enrichment is fairly modest (approximately 3.9)? More generally, does the total level of enrichment play any role? For example, when the authors repeat the enrichment analysis while controlling for local background selection effects (Extended Data Figure 3), the enrichment drops to only approximately 2.1. I find it difficult to reconcile such a modest enrichment with the strength of the claim it supposedly supports.

These are important questions, and our first submission did not provide sufficient confidence in the robustness of our FDR calibration procedure. In our revision, we have verified these approaches using forward-in-time simulations with SLiM, as shown in Figure 1b, Extended Data Figure 4, and Supplementary Information Section 2, subsection "Enrichment of GWAS hits for variants under directional selection."

The magnitude of the enrichment is not a concern, as we explain in our revision. It depends on the fraction of variants identified as GWAS hits. For example, if 1% of variants are GWAS hits (ignoring adjustments for minor allele frequency), the maximum absolute enrichment value cannot exceed 100. In our simulations, only a small fraction of variants are GWAS hits due to computational constraints. In real GWAS, a much larger portion of the genome (15–20%) are identified as GWAS hits, making the absolute enrichment magnitude appear smaller. Nevertheless, our study relies on the plateau effect of the enrichment pattern rather than the absolute magnitude to approximate FDR, and we show with our simulations that this is robust.

Another question is whether all future selection scans should adopt this GWAS enrichment method to determine significance thresholds. Do the results of this paper suggest we should expect an order of magnitude increase in significant hits in future selection scans? In the discussion, the authors compare their results to earlier work (from over ten years ago) that searched for evidence of adaptation on deeper timescales. They try to explain the discrepancy in the number of signals detected. However, it remains unclear whether those earlier scans would have found so few signals if they had used the same procedure to determine a significance threshold.

This study leverages a high-quality imputed ancient DNA sequences (14-fold larger than any previous study), combined with extensive GWAS data, analyzed using a powerful technique, and we believe that the increase in findings is, to a large extent, due to these factors.

Indeed, controlling for FWER is more conservative than FDR from a statistical perspective, and it is reasonable to assume that if other methods took the approach of controlling for FDR, they would detect more signals than previously reported. However, understanding how many more signals they would identify is beyond scope of this study.

We did manually re-evaluate all the signals found in the previous studies (Supplementary Information section 6), and show that despite their theoretically more conservative FWER-based threshold, about half of them do not replicate using our approach (only 18 of 38 non-HLA loci in the previous studies that we evaluate replicate at a posterior probability of >0.99). We discuss that some of the previous methods could have been sensitive to types of selection that our method was not able to detect, for example selection that fluctuated over time or space. At the same time, we show that a substantial fraction of instances in which we do not replicate previously reported signals are likely to be due to previous studies not adequately controlling for artifactual processes.

I also want to note a general conceptual disconnect in this enrichment procedure. The effect sizes these variants have on their GWAS traits are presumably quite small and therefore likely not the actual cause of the putative selection signal. So it is fairly hazy as to exactly why this is supposed to provide such strong evidence. Thus, while the enrichment is interesting, I don't think there is enough evidence to overwrite the existing statistical significance standards in this field.

In this study, we identify theoretical problems with applying standard procedures for declaring genome-wide significance to our test statistic (non-normality), propose a solution to address the issue, and confirm its validity using multiple lines of evidence and comprehensive forward-in-time SLiM simulations.

As we explain in our response to referee #2, we believe that one of the reasons why our enrichment approach works is pleiotropy. For any given GWAS, it is true that the effect sizes of most alleles are almost certainly small, and for this reason selection coefficients associated with the trait being explored with GWAS are probably too modest for us to detect. But the right way to think of our enrichment strategy is not on the single GWAS level. If the selection was for the trait being studied in that GWAS, most of the alleles would indeed be expected to have small associated selection coefficients and to be difficult to detect in our scan. But single-variant GLMM identifies selection at the variant level, without any knowledge of the actual trait under selection. Because of pleiotropy, a selected variant can appear in GWAS of many traits where it will show small effect sizes, but the selection that is being detected will be for a trait where the effect size is large. For example, the *LCT* variant rs4988235 is strongly selected but associated with 19 traits in PAN-UKBB, with effect sizes of only 1-2% for most traits except “Never have milk” (Pheno ID 101 in AGES; pheno code 1418 in UKBB), which has an effect size of 9%. The strong effect size traits presumably drive the selection.

Like the referee, we were skeptical of the large yield of significant signals we obtained. We spent years trying to disprove these results, but the evidence only got stronger. Our revised manuscript provides multiple lines of independent evidence showing that the signals of selection we identify are real. So, while the *a priori* assumption we had was that the findings were unlikely, we eventually built up evidence that was sufficiently strong as to make the findings compelling.

Major comments on polygenic score analysis

I have two main concerns about the polygenic score analyses. As always, the primary concern with this sort of analysis is that residual environmental confounding in the GWAS may have led to a bias in the distribution of polygenic scores, which can manifest as a signal of selection. This is particularly likely to be a problem when lenient p value thresholds are used to construct the polygenic score, as is the case here.

The authors are aware of these concerns and take steps to address them. However, it is unclear whether these steps have been successful, largely due to some ambiguity in the text regarding what was done.

The currently accepted best method for overcoming stratification bias in this kind of analysis is to use effect sizes from a study with a population structure that does not overlap with the test panel. This can be accomplished using either family-based effects or a population GWAS that is evolutionarily diverged. The authors attempt both, using effects estimated from siblings and from East Asians.

However, this method is only guaranteed to prevent bias if the unconfounded effect sizes are used both for ascertaining SNPs to be included in the polygenic score and for actually calculating the polygenic score itself. Using unconfounded effects only for the calculation but not for the ascertainment can still result in biased polygenic scores (see Figure 6 of Zaidi and Mathieson), because this ascertains a biased subset of SNPs. The text is unclear about the authors' exact approach, so it is difficult to judge whether the East Asian/sibling analyses are unconfounded.

This is an important point. For the same reasons described by the reviewer and also outlined by Zaidi and Mathieson (2020), SNP selection for each GWAS is performed independently and blind to other GWAS or selection signals. We have added the following sentence to clarify this.

“For each phenotype, we generated an independent set of SNPs using a two-step clumping and thresholding process. To avoid ascertainment bias [Zaidi and Mathieson 2020], SNP selection and effect size estimation for polygenic score calculation were both derived from the same GWAS, independent of other GWAS or selection signals.

Regarding the sibling effect analysis, I also want to note that the authors attribute the lack of meaningful insight to limited sample sizes in family-based GWAS (Extended Data Figure 13). However, they do not provide any power analysis to support this conclusion, and in fact several signals are no longer significant. While limited sample size may indeed be a factor, the authors should clearly state that multiple signals do not replicate.

In the previous version of the paper, we used the Howe et al. 2022 summary statistics, which Tan et al. 2024 later showed to be underpowered due to the inclusion of low-quality imputed variants. With the improved family-based GWAS from Tan et al. 2024, we now observe that the directional signal of selection for years of schooling is nominally significant across all three tests presented here.

We also revised the caption of Extended Data Figure 14 (previously 13) to address this comment:

"The fifth column shows the sample size, with the median sample size for family GWAS ~13,000 and for population GWAS ~31,000. Both are only a fraction of the UK Biobank cohort of ~490,000 individuals of non-Finnish European ancestry[286], explaining the reduced power and why confidence intervals are wide and often overlap zero. The Pearson's correlations between Z-scores from population GWAS and direct genetic effect GWAS are 0.51, 0.51, and 0.76 for γ, γ_{sign}, and r_s, respectively. When restricted to population GWAS signals with nominal significance ($P < 0.05$), the correlations increase to 0.84, 0.78, and 0.81, respectively. Among population GWAS in [68], educational attainment, ever-smoker, and myopia reach the Bonferroni-corrected significance ($P < 1.5 \times 10^{-3}$, correcting for 34 traits) for all three tests. For direct genetic effects, educational attainment is the only trait with all three tests nominally significant ($P < 0.05$)."

Minor comments:

- The first sentence of the abstract is simply untrue. For example, here are three papers that test for consistent allele frequency changes over time. There may be more these are just three that came immediately to mind.

Ju and Mathieson
https://www.pnas.org/doi/abs/10.1073/pnas.2009227118

Mathieson and Terhorst
https://genome.cshlp.org/content/32/11-12/2057.short

Fine and Steinruecken
https://www.biorxiv.org/content/10.1101/2024.05.10.593575v1

We agree, and have revised the abstract to reflect this correction.

- the title of SI section three is "HAF score analysis proves directional selection"
Above I said that I didn't find this section persuasive, but I also think the authors should remove the world "proves" and talk about "evidence" or something, especially in a scenario such

as this where we are relying on fraught statistical inferences about things we can never fully experimentally confirm.

We revised the title to “HAF score analysis **provides evidence for** directional selection”

Referees' comments:

Referee #1 (Remarks to the Author):

In the revised manuscript 2024-09-19432B, Akbari et al. have partly addressed several of my initial comments, by (i) exploring the power of their GLMM model and the expected proportion of GLMM-significant variants among trait-associated variants with forward-in-time simulations of neutral and selected mutations, assuming different models of selection under a semi-realistic demographic model, (ii) estimating the number of independent selected alleles by using different clumping strategies based on both r^2 and D'; (iii), showing that the null distribution of the GLMM model is not normal and the test statistics are inflated for low-frequency variants; and (iv) reporting correlations between polygenic selection test statistics using GWAS in Europeans and East Asians. Although conclusions in the revised manuscript are strengthened, several of my main concerns have not been satisfactorily addressed.

1. TPR, FPR and FDR of the GLMM. The reported forward-in-time simulations were carefully designed and the way purifying, stabilizing and directional selection were modelled is very reasonable. These simulations show unambiguously that no enrichment of trait-associated variants in significant GLMM variants is expected under stabilizing selection. However, because the authors investigated true positive rates using a different significance threshold than their empirical threshold ($|X| > 5.45$, corresponding to P-value < $1.95 \times 10\text{-}16$; Fig. S3.2), the simulations do not inform the actual accuracy of their method, under either background selection or pre-admixture selection. In Fig. S2.10, they assumed FPR = $1.0 \times 10\text{-}5$; for the case of pre-admixture selection, they considered a Bonferroni corrected P-value = $1.25 \times 10\text{-}4$. Therefore, I invite the authors to report the TPR and FPR of the GLMM and GLM models under models 1 to 3 considering $|X| > 5.45$ as the significance threshold (the enrichment plateaus at this value in both empirical and simulated data; Extended Data Fig. 4).

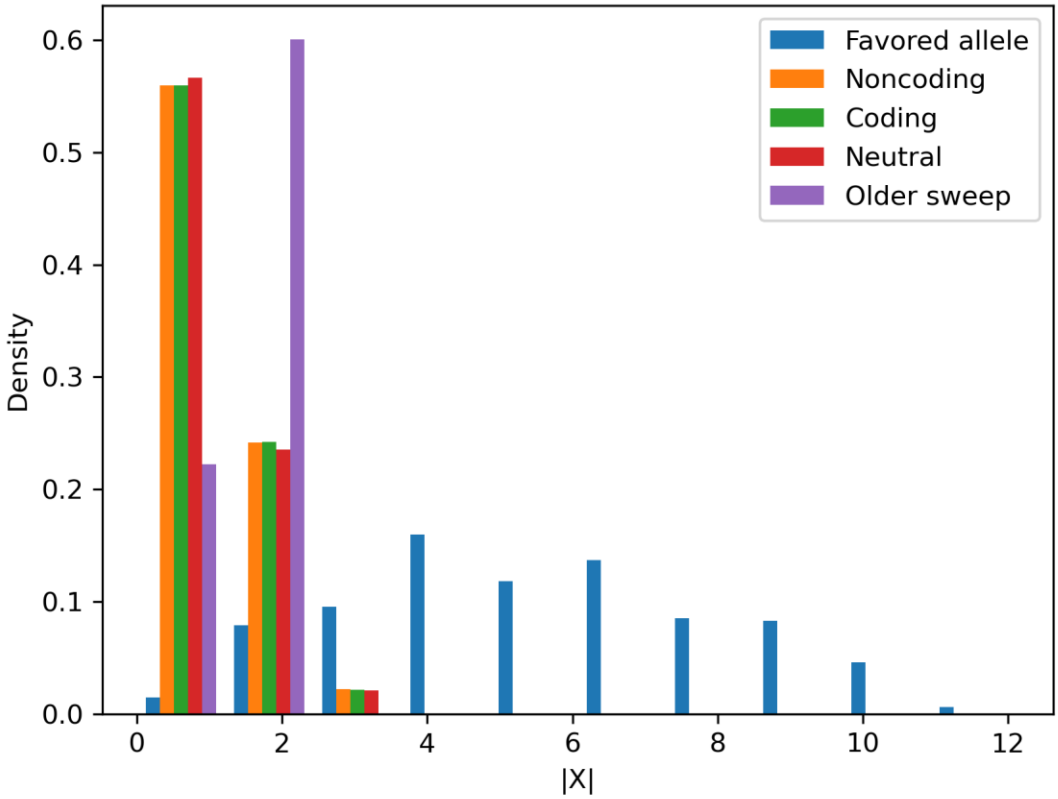
We have already demonstrated, using simulations of extreme sweep scenarios, that our GLMM does not detect pre-admixture selection. Accordingly, in Supplementary Information Section 2 (subsection: *Older sweeps in ancestral populations*), we added the following statement: "Under the significance threshold $|X| > 5.45$, the false-positive rate in this experiment is zero."

In addition, all three simulation models in Supplementary Information Section 2 incorporate background selection, and all power analyses accounted for background selection.

We conducted the experiment suggested by the reviewer, and the results are provided in Supplementary Information Section 2 (subsection: *Power analysis of GLMM and GLM and colinearity*), with TP values of 26.5%, 59.25%, 89%, and 22.25% for models 1, 2.1, 2.2, and 3, respectively, and FP equal to zero for all models. A summary of these results is presented in the new Supplementary Table S2.1, and the new additions are highlighted in yellow.

Plotting the X distribution for positively selected mutations, as well as mutations in coding, noncoding and neutral regions and mutations under pre-admixture selection would also be informative.

The requested plot is shown below. We used Model 2 without directional selection to estimate the distribution of |X| for noncoding, coding, and neutral regions, and Model 2.1 with directional selection to estimate the distribution of |X| for favored alleles. For older sweeps, we used 400 simulations from Supplementary Information section 2 (subsection: *Older sweeps in ancestral populations*), which model extreme cases of classic selective sweeps in ancestral populations. We determined that none of these variants are identified by our model as a signal of selection.



Notably, it is currently unclear whether the GLMM has lower FPR than GLM in the presence of population structure. This is important because GLMM and GLM show similar TPR, when the number of PCs is carefully chosen, FPR = 1.0 × 10-5 and MAF > 0.1 (Extended Data Fig. 16, Fig. S2.10b).

All simulations conducted for the power analysis explicitly include population structure. We have now added Supplementary Table S2.1 in Supplementary Information section 2 (subsection: *Power analysis of GLMM and GLM and co-linearity*), which shows that under the significance threshold |X| > 5.45, the GLMM outperforms all GLM models with lower FPR and higher TP, even when the number of principal components is optimally chosen for simulations. In real data, however, selecting an optimal number of principal components is not possible, as the studied

populations are much more complex than the simplified demographic models used in simulations, and GLM-based approaches face an unavoidable trade-off between residual population structure and collinearity that does not affect GLMM; thus, in realistic situations the GLMM has an even greater advantage.

Furthermore, it is also unclear which simulated model was used in Figure S2.11 and S2.12.

The following line has now been added to both figures:

"We used 20,000 simulations under Model 2 without directional selection on a randomly chosen 100-kb genomic region."

Finally, as initially requested by Reviewer 3, I invite the authors to provide more explanations about equations in pages 40-41, and detail the assumptions made.

We have substantially revised and expanded Supplementary Information Section 4 (subsection: *Controlling for false discovery rate (FDR) by leveraging GWAS signals*) to clarify the underlying assumptions and logic of the procedure. The added text is highlighted in yellow.

2. Empirical consistency checks. The authors have evaluated whether their simulations reproduce some aspects of the true data, but these checks remain partial. The authors compared simulated and true PCA, and acknowledged that Irving-Pease et al.'s demographic model largely underestimates variation in ancestry proportions among Europeans. This raises some concerns regarding power comparisons between GLMM and GLM using these simulations. In particular, Fig. S2.11 shows that $1 - R^2$ is lower for GLM in simulations, relative to observed data, suggesting underestimation of GLM power in the simulations (due to higher correlation between sampling time and PCs, relative of empirical data).

We agree that the simulations provide only a partial representation of the complexity observed in real data. The demographic model of Irving-Pease et al. does not capture the full history of Europeans, leading to stronger correlations between sampling time and principal components in simulations than in real data. This simplification is expected to affect power estimates for GLM-based approaches.

There is an inherent trade-off in GLM analyses between residual population structure when too few principal components are included, and variance inflation due to collinearity when too many are used, which in turn reduces power. In this context, it is not surprising that the power of GLMs with larger numbers of principal components (for example, more than four) decays more rapidly in simulations than in real data, where the decline is slower due to weaker correlations. Although we do not expect the absolute magnitude of power loss observed in simulations to translate directly to empirical data, the variance inflation factors in the real data are greater than 4 when more than four PCs are included as covariates, indicating substantial collinearity and a consequential loss of power. This highlights a fundamental weakness of GLM-based approaches.

Importantly, the choice of an optimal number of principal components is well defined in simulations but inherently ambiguous in real data, where complex demographic histories make it challenging to balance residual structure against collinearity. To address this directly, we have added Supplementary Table S2.1, which shows that under the significance threshold $|X| > 5.45$, the GLMM consistently outperforms all GLM models in simulations. Given that real data exhibit substantially more complex population structure than the simulated models, we expect the relative advantage of the GLMM to be even stronger in empirical analyses.

Furthermore, I suggest the authors add observed heterozygosity in Fig. S2.4 (relative to intergenic regions, possibly).

We have normalized the heterozygosity values in this figure relative to neutral regions, updated the caption (highlighted in yellow), and report the corresponding empirical values in the caption.

and the observed β/EAF relationship for a trait under stabilizing selection in their empirical analyses in Fig. S2.6 (Koch et al., bioRχiv 2024).

In Koch et al. 2024, the authors convincingly explain the U-shaped (or smile-shaped) relationship between effect size and effect allele frequency shown in Figures 1, 2, and 4. We have now included the corresponding plots for height and cholesterol from the UK Biobank as panels (e) and (f) of Figure S2.6, shown in gray. We observe a similar pattern that closely matches the behavior seen in our simulations.

3. Pre-admixture selection. Instead of using forward-in-time simulations, the authors decided to simulate genotypes by sampling from a binomial distribution, assuming admixture proportions from Irving-Pease et al.'s model. However, simulating temporally varying s is easy to implement in SLiM, and would enable the authors to determine if the HAF statistic is inflated under pre-admixture selection, which is not currently investigated.

The theoretical dynamics of the HAF score, together with simulations in Ronen et al. 2025, show that the HAF score is expected to be larger than the neutral expectation for ongoing selective sweeps, as well as for partial selective sweeps that are no longer under selection for hundreds of generations after the sweep has stopped. Therefore, we expect the HAF score for variants that underwent a selective sweep in an ancestral population and are no longer under selection during our sampling period to also be inflated. However, with our current simulations, we show that our approach does not capture signals of selection in ancestral populations. In this case, inflation of the HAF score due to ancestral selection shifts the baseline HAF level but does not generate the pattern we observe for residual HAF as a function of the X score. Our HAF-score results provide unambiguous validation that our method is detecting genuine signals of directional selection during the period of the time transect, which is the only point we wish to make with this analysis.

4. s estimation. I suggest that the authors use the simulations already performed to evaluate if their selection coefficient estimates are unbiased. In several parts of the manuscript, the authors comment GLMM s estimates, but they have not evaluated estimation accuracy.

To address this concern, we have added Figure S2.10, which evaluates the accuracy and bias of selection coefficient estimation using the simulations already performed in Figure S2.7. We determined that the estimated selection coefficients track the true simulated values reasonably well. In addition, we have added a dedicated paragraph at the end of the section titled "Power analysis of GLMM and GLM and co-linearity" that explicitly discusses the performance of GLMM-based selection coefficient estimates.

5. LD clumping. The authors reported in a new Supplementary table the number of independent signals estimated considering different clumping strategies. Although some strategies result in a substantial reduction in the number of independent signals, and estimating such a number is a central objective of their study, they decided not to mention these results in the main text. The authors claim that their "approach to identifying distinct loci is conservative, because genuinely selected alleles in LD with nearby stronger ones will be missed" (l. 188-190), but they provide no evidence that their approach is indeed conservative. For example, by just taking a pair of nearby variants with very similar trajectories and s values in Fig. 3 (rs3891176 for locus 1 and rs7775397 for locus 19), I find with PLINK1.9 --ld that have $r^2 = 0.215411$ and D' = 0.737 in Europeans from 1KG (so one of the two variants should be filtered).

We changed the wording to avoid using "conservative," as indeed we do not provide quantitative evidence for it in the paper. We revised the text to: "Moreover, in our approach to identifying distinct loci, genuinely selected alleles in LD with nearby stronger ones will be missed."

Regarding the variants highlighted in Showcase Fig. 3 at loci 1 and 19, both are missense variants strongly associated with multiple traits, including autoimmune diseases such as celiac disease, and are well documented in the literature. Although they are in moderate LD with each other, conditional regression analyses, in which the genotype of one variant is included as a covariate when estimating the selection coefficient of the other, do not eliminate the signal. This indicates the two signals are likely independent. We highlight both variants because they represent independent, biologically relevant signals of selection and are clinically meaningful.

As stated in the caption of Fig. 3:

"The highlighted loci are not necessarily those with the strongest signals, and even include negative results. We highlight them here because of their biological interest and because they speak to long-standing debates."

We emphasize that inclusion of a variant in Showcase Fig. 3 does not imply that it is among the 479 independent signals reported in Supplementary Information Section 5. The showcase figure presents a hand-picked set of variants selected for illustrative and biological interest, whereas Supplementary Information Section 5 reports signals identified using a systematic, threshold-based approach independent of clinical or biological context, and this list does not include both of the loci the reviewer mentions.

Now we have edited the caption of the Figure 3 to further clarify this:

"The highlighted **variants** are not necessarily those with the strongest signals and **may** include negative results **or variants in partial linkage disequilibrium**. We highlight them here because of their biological interest and because they speak to long-standing debates."

Furthermore, their FDR approach assumes that variants are independent, which may not be true due to hitch-hiking effects, questioning the way they define genome-wide significance. Therefore, I strongly recommend that the authors investigate which clumping strategy provides the best estimate of the number of independent selected variants, based on the simulations already performed.

The reviewer is correct that estimating the number of independent signals of selection is challenging. A related problem—the exact number of causal variants above a significance threshold—has been widely studied in GWAS and has been solved to a substantial degree of satisfaction. But key assumptions underlying standard GWAS fine-mapping do not hold in our setting, because LD among variants of interest changes over time, which violates the assumptions of fine-mapping approaches, particularly in the case of multi-allelic fine-mapping.

The reviewer recommends using our simulations to provide better definition of the number of independent loci affected by selection, but in fact, in our simulations and particularly under Models 2 and 3, the variants effectively under selection are not well defined because selection acts through shifts in the optimal trait value, making the number of selected variants inherently ambiguous. Even in Model 2, where a single variant is explicitly designated as the target of selection, linked trait-associated variants in LD can shift the trait optimum and thereby induce an interaction between polygenic and single-variant selection mechanisms, resulting in more than one variant effectively experiencing directional selection. Because we do not explicitly model multiple causal variants under selection, the evaluation of an optimal LD pruning strategy depends critically on assumptions about the number of causal variants and cannot be straightforwardly determined from these simulations alone. Defining an optimal pruning strategy would require explicit simulation of multiple causal variants. For example, if one assumes exactly one causal variant per region, then an extremely stringent pruning strategy, such as $r^2 = 0.05$, $D' = 0.05$, and a 10 Mbp window, will trivially appear optimal by construction.

In light of the inherent difficulty of defining the number of independent loci, we have chosen not to carry out further work in this direction; future work may be able to provide a more precise answer. To acknowledge the fundamental ambiguity in the number of independent loci, however, we report the sensitivity of the results across clumping strategies in Table S5.1.

Minor comments.
X and HAF. The authors use HAF as an independent line of evidence, in addition to GLMM X. Can the authors compute the correlation between X and HAF in the different simulated models, without and with directional selection?

In the following table, we summarize the correlation between |X| and mean HAF-score across different simulated models, as requested by the reviewer.

Jackknife estimates of the correlation and standard error (SE) between |X| and the mean HAF score across different models, based on 400 simulations per scenario. These simulations are the same as those used in Extended Data Figure 5. Jackknife estimates were obtained by leaving out one simulation at a time.

Model	Directional selection	Correlation	SE
1	Without	-0.0374	0.0017
1	With	-0.0338	0.0022
2.1	Without	-0.0350	0.0017
2.1	With	-0.0282	0.0029
2.2	Without	-0.0378	0.0017
2.2	With	-0.0106	0.0034
3	Without	-0.0349	0.0017
3	With	-0.0318	0.0021

Across all models, the correlation between |X| and the mean HAF score is negative. The magnitude of this negative correlation is consistently reduced in the presence of directional selection. This pattern is consistent with Extended Data Figure 3c, which shows that stronger signals of directional selection are associated with higher HAF values and a corresponding attenuation of the negative relationship between HAF and |X|. Consistent with prior work showing that the HAF score is most sensitive under strong selection (Ronen et al. 2015; Akbari et al. 2018), Extended Data Figure 3c shows that the expected and observed mean HAF scores track closely up to $|X| \approx 4$, which includes approximately 99% of variants. As a result, directional selection induces a qualitatively small but statistically significant change in the correlation.

Importantly, our analyses (Figure 1c,d and Extended Data Figure 5) exploit the signal captured by the residual mean HAF score, rather than the raw mean HAF score. Residuals are computed as the observed mean HAF minus the expected value from a linear regression that corrects for multiple measures of background selection. This underscores the importance of controlling for background selection when studying directional selection and demonstrates that, although signals of directional selection are substantially masked by other evolutionary forces such as background selection, they remain detectable when analyzed with appropriate statistical models and data.

Simulations. Could the authors please clarify further how many 10-Mbp loci were simulated per model? Can the authors clarify how many 10-Mbp regions were used to compute PCs and the GRM?

We clarified this in the first paragraph of the Supplementary Information section titled "Enrichment of GWAS hits for variants under directional selection." The added text is highlighted in yellow.

Extended Data Fig. 3b. Please report HAF distribution under the semi-realistic demographic model.

This figure demonstrates the distribution of the HAF score under three evolutionary scenarios: background selection, neutrality, and positive selection. The qualitative differences among these scenarios are discussed in Supplementary Information Section 4 and in prior work (Ronen et al. 2015). The purpose of this plot is not to make claims about the absolute values of the HAF score under realistic demographic histories or across different evolutionary scenarios. Rather, it illustrates the relative behavior of the HAF score when positive or background selection is present compared with the neutral case. In this context, we do not see how reporting the HAF score distribution under a semi-realistic demographic model would add to the specific point being made by this figure.

Figure legends. The authors refer either to a proportion of SNPs associated in GWAS or an enrichment, when describing the y axes of plots reporting these enrichments. Please clarify and homogenise descriptions across panels and figures.

We replaced the term "proportion" with "enrichment" in Figure 1 and Extended Data Figure 3.

Figure 1 legend. Please remove 'perfectly' (l. 950).

We removed the word "perfectly".

Figure 2b and Extended Data Fig. 1. Please replace 'coeificient' by 'coefficient'.

We fixed the typo in both figures.

Extended Data Fig. 15. The authors may caution that these genetic correlations can be inflated due to assortative mating.

We have added this caveat to the Extended Data Figure legend.

Extended Data Fig. 16a. The curve for GLM (1) is not visible.

The curve for GLM (1) was obscured by the other curves. We added a small x-axis offset for each model to prevent overlap and ensure all curves are visible.

Referee #1 (Remarks on code availability):

I was able to download the code and run a toy example. Documentation remains relatively limited.

We expanded the documentation, added more detailed instructions, and included additional examples to improve usability for the community.

Referee #2 (Remarks to the Author):

This paper reports a series of very interesting results, and an extraordinary data set that will be of great interest to the field. The analysis is highly complex and there are details where one could still ask for more simulations or rationale, but at this point I am convinced that the major claims are likely correct. In summary, I am now supportive of publication in principle, and I congratulate the authors on what will be an important and widely-discussed paper.

I do have several specific comments that I would like to see addressed before acceptance:

1. The presentation of the GWAS FDR is absolutely critical to the paper, yet poorly described. I'd like to see the authors rewrite this.

We agree and have rewritten this section. We have substantially revised and expanded Supplementary Information Section 4 (subsection: *Controlling for false discovery rate (FDR) by leveraging GWAS signals*) to clearly articulate the inferential goal, assumptions, and empirical logic of the procedure. All new and modified text is highlighted in yellow.

-- The supplementary section on FDR control (around p. 40) is poorly explained and difficult to follow. I encourage the authors to improve the clarity of the mathematical derivations and provide a more transparent explanation of the logic behind the procedure. Is there some unstated assumption here? Maybe that the FDR is required to decrease all the way down to zero with increasing values of the test statistic? Please clarify that in the main text.

We agree the original presentation lacked clarity. In the revised text, we make the underlying logic explicit. In particular, we clearly state the ordering assumption that variants with larger absolute values of the selection statistic are increasingly enriched for true signals of directional selection, implying a monotonic decrease in the false discovery rate as the statistic increases. We clarify that the observed plateau in enrichment identifies a regime in which the contribution of false positives is negligible. If false positives were present at a non-negligible rate at high statistic values, enrichment would be expected to continue increasing rather than stabilize. This behavior is demonstrated both in forward-in-time simulations and in the empirical data.

-- Beyond the math for this section, I'd like to see the authors articulate more clearly what exactly they think they are testing, and what is the rationale for what they are doing. They had some of this in the Response to Reviewers, but it's so central to the paper that I think it should be in the main text.

We agree that this point is central and have clarified it in the revised text. We now articulate more clearly what is being tested and the rationale for the procedure, and we have incorporated this framing into the main text and into the revised Supplementary Information Section 4.

I agreed with some of what they wrote in the Response, but I would actually frame it a bit differently. They are including (nearly) all SNPs that are genome-wide significant in any of a

huge number of traits. This encompasses around 12.5% of all SNPs. Importantly, they are NOT attempting in any way to colocalize the selection signals with GWAS hits. So we should really think of this as a type of genome annotation that spans ~12% of the genome, and which is highly enriched for likely functional variants, as well as SNPs that are in even pretty weak LD with functional variants. Note that the SNPs in this set must surely have higher LD than all SNPs genome-wide. So this is a very blunt annotation, and the authors could perhaps have done about as well by using other types of functional annotations: eg perhaps some combination of highly functional genes + conserved noncoding regions. That's ok in my view, but I think it would be very helpful to the reader to frame it in this way, rather than using language that may be misunderstood by readers to suggest that the selection signals are actually colocalizing with specific GWAS hits.

As described in Supplementary Information Section 4, our goal is not to assess colocalization between selection signals and specific GWAS hits, nor to perform fine-mapping or assign biological interpretation to individual overlaps. Instead, we use genome-wide significant GWAS associations across many traits to define a broad external genomic annotation enriched for functional variants and variants in linkage disequilibrium with them. This GWAS-derived annotation encompasses approximately 12.5% of the variants analyzed here.

The role of this annotation is to examine how the proportion of variants overlapping this functional proxy changes as a function of the absolute value of the selection statistic. At low statistic values, enrichment is close to the genome-wide baseline, whereas as the statistic increases, enrichment rises and then reaches a clear plateau. This empirical behavior, rather than the specificity of GWAS loci, underlies our use of enrichment to empirically calibrate genome-wide statistical significance in a setting where standard null assumptions do not hold.

All that said, the 5-fold enrichment that the authors report here is very impressive, and this is the main observation that helps me to overcome any other concerns that I may have with the analysis pipeline. If, for example, they had found a 1.2-fold enrichment, this could also have been a highly significant overlap, but I would have had much more concern about possible confounders.

We agree that the 5-fold enrichment is a compelling finding. Given the broad genomic scope and coarse resolution of the annotation, the magnitude and stability of this enrichment support the interpretation that the extreme tail of the statistic is dominated by true signals of directional selection rather than residual confounding.

2. There should be straightforward ways to estimate how ancestry proportions change over time in the aDNA panel, and I encourage the authors to perform such checks. In particular, they could test whether variants strongly associated with ancestries that increase or decrease over time show systematically lower p-values. If so, this would indicate that the ancestry correction is still incomplete. In that case, the authors should explicitly note this limitation in the main text.

We agree that residual confounding by ancestry is an important concern in ancient DNA time-series analyses. However, the specific diagnostic suggested here is not valid in this setting.

Variants that are strongly associated with ancestry components whose proportions change over time are not expected to be neutral, even under correct ancestry correction, when selection acts on standing variation or along ancestry-associated axes. In such cases, genuine targets of selection will necessarily be correlated with ancestry early in the time transect, and this correlation does not imply incomplete correction. Consequently, testing whether ancestry-associated variants show systematically lower p-values is not a reliable gauge of whether population structure has been properly handled. Indeed, many well-established targets of selection with independent support from both ancient DNA and modern scans, including *LCT, SH2B3/ATXN2, FADS1/2, SLC22A4/5, SLC45A2, BNC2, DUOX2, KITLG, TLR1, TYR, ABO*, and *CYP1A2*, show strong correlations with one of the major ancestry components (Steppe, WHG, or EEF) in our three-way model, with ancestry-association p-values below $5x10^{-8}$.

Addressing population structure was a central challenge of this study, and our goal was to correct for it rigorously without sacrificing excessive power. To this end, we use a GLMM framework in which the variance component σ^2 is estimated independently for each variant and is constrained to be no smaller than the genome-wide average conditional on minor allele frequency. This ensures that population structure is corrected for at least at the genome-wide average level, and often more strongly, at each locus. We assess the adequacy of this correction using forward-in-time simulations with realistic demography, negative-control scenarios without directional selection, and independent biological validation through GWAS enrichment and HAF behavior, none of which shows evidence that residual confounding produces false positives.

3. I suggest reconsidering the section “LD score and pseudo-intercept”. In linear mixed model-based GWAS, LD score regression is generally not appropriate because the LMM changes the expected mean of the chi-squared statistics, making the intercept hard to interpret. The proposed “pseudo-intercept”, defined using SNPs above an LD-score threshold, does not appear to solve this underlying issue. The statement in the supplement (p. 41) that agreement with the pseudo-intercept provides independent support for the approach is therefore not convincing, and I recommend removing or revising this section.

We agree with the point raised by the reviewer and removed the following statement:

“Encouragingly, this value aligns closely with the pseudo-intercept of the LD score regression (2.33±0.21), providing independent support that this is a reasonable approach.”

Referees' comments:

Referee #1 (Remarks to the Author):

In the revised manuscript (2024-09-19432C), Akbari et al. have now largely addressed my previous comments. I had emphasised that it would be highly informative to determine which LD clumping strategy provides the most accurate estimate of the number of independent loci under directional selection. The authors argue that this question is too difficult to address. While I respectfully disagree, I do accept that the manuscript in its current form already presents numerous novel results and methodological innovations and represents as such one of the most important advances to date in the study of human evolutionary history.

Because the exact number of independent loci is inherently difficult to estimate, owing to (i) uncertainty in the significance threshold, (ii) residual LD among candidate loci, and (iii) the presence of genuinely independent, closely linked variants, I recommend that the authors adopt more cautious language when reporting the number of independent loci.

More specifically, I propose the following minimal changes in the main text:
- L. 10: “By applying this to 15836 West Eurasians who lived the past 18000 years and 6438 contemporary people, we find several hundred genome-wide significant signals with >99% probability of selection, an order of magnitude more than reported in previous studies.”

We have revised the Abstract, and this statement is no longer included.

- L. 192: “Moreover, although residual LD may exist among candidate loci (Supplementary Information section 5), genuinely selected alleles in LD with nearby stronger ones can also be missed.”

We have revised the corresponding text to address the reviewer’s concern. In lines 145-147 we now state:
“… residual LD exists among candidate loci which at some loci may cause some overestimation of signals (Supplementary Information section 5), while truly selected alleles in LD with nearby stronger ones will be undercounted.”

- L. 504: “Although the exact number of independent loci cannot be estimated with high confidence, we project that there are at least 3800 independent signals of directional selection”

We have incorporated the reviewer’s suggested revision into the main text. In lines 412-413 we now state:
“Although the exact number of independent loci cannot be estimated with high confidence, we project that there are at least 3800 independent signals of directional selection…”

Please also find below a few other minor changes:

- L. 123: “The plateau occurs at the same place when we control for “background selection”, the loss of neutral variation linked to deleterious mutations removed by purifying selection, although the enrichment in GWAS variants is reduced (~2.4-times, Extended Data Figure 3a).”

We have incorporated the reviewer’s suggested revision into the main text. In lines 80-83 we now state:
“The plateau occurs at the same place when we control for “background selection”, the loss of neutral variation linked to deleterious mutations removed by purifying selection, although the enrichment in GWAS variants is reduced (~2.4-times; Extended Data Figure 3a).”

- L. 178-179: Please consider citing Sikora et al., Nature 2025

We thank the reviewer for bringing this to our attention. We have added Sikora et al. 2025 to line 131.

- L. 379: “after correction for [the] number of GWAS (traits?) examined”

We have added “the” to correct the grammar. We intentionally retain the term “GWAS” because our analysis includes multiple distinct GWAS of the same underlying trait, often based on different definitions or cohorts. Referring only to traits would undercount the number of association analyses performed. Since some GWAS are correlated, correcting for the total number of GWAS examined is an appropriate and conservative multiple-testing adjustment.

- L. 542: “Three theoretical reasons indicate that [it] cannot be explaining our observations”

We have incorporated the reviewer’s suggested revision into the main text. In lines 445-446 we state:
“Three theoretical reasons indicate that **this** cannot be explaining our observations.”

Referee #1 (Remarks on code availability):

The code is now sufficiently documented.

We are glad the referee agrees our code availability and documentation is now sufficient.