Predicting Functional Consequences of Recent Natural Selection in Britain

Lin Poyraz,[1,2] Laura L. Colbran ,[1] and Iain Mathieson [1,*]

[1]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

[2]Department of Computational Biology, Cornell University, Ithaca, NY, USA

***Corresponding author:** E-mail: mathi@pennmedicine.upenn.edu

Associate editor: Evelyne Heyer

Abstract

Ancient DNA can directly reveal the contribution of natural selection to human genomic variation. However, while the analysis of ancient DNA has been successful at identifying genomic signals of selection, inferring the phenotypic consequences of that selection has been more difficult. Most trait-associated variants are noncoding, so we expect that a large proportion of the phenotypic effects of selection will also act through noncoding variation. Since we cannot measure gene expression directly in ancient individuals, we used an approach (Joint-Tissue Imputation [JTI]) developed to predict gene expression from genotype data. We tested for changes in the predicted expression of 17,384 protein coding genes over a time transect of 4,500 years using 91 present-day and 616 ancient individuals from Britain. We identified 28 genes at seven genomic loci with significant (false discovery rate [FDR] < 0.05) changes in predicted expression levels in this time period. We compared the results from our transcriptome-wide scan to a genome-wide scan based on estimating per-single nucleotide polymorphism (SNP) selection coefficients from time series data. At five previously identified loci, our approach allowed us to highlight small numbers of genes with evidence for significant shifts in expression from peaks that in some cases span tens of genes. At two novel loci (*SLC44A5* and *NUP85*), we identify selection on gene expression not captured by scans based on genomic signatures of selection. Finally, we show how classical selection statistics (iHS and SDS) can be combined with JTI models to incorporate functional information into scans that use present-day data alone. These results demonstrate the potential of this type of information to explore both the causes and consequences of natural selection.

Key words: **ancient DNA, gene expression, human evolution, time series.**

Introduction

Ancient DNA (aDNA) time series can provide direct evidence of natural selection on specific variants, avoiding confounding factors associated with inferring selection using modern data (Marciniak and Perry 2017; Dehasque et al. 2020; Mathieson 2020). However, in itself ancient DNA does not provide any information about the functional consequences of selection, limiting our ability to learn about phenotypes under selection and to identify effects of selection that may for example affect disease risk.

One problem is that, similar to genome-wide association studies, selection signals often span multiple genes due to linkage disequilibrium (LD), making it difficult to identify the loci targeted by selection. Indeed, long haplotypes due to selective sweeps make this problem even more challenging. One approach that has been promising in the genome-wide association study (GWAS) context is to incorporate functional information, for example expression quantitative loci (eQTL), which have been used to link significant GWAS hits to functional consequences in transcriptome-wide association studies (TWAS) (Wainberg et al. 2019). Similarly, while the results of GWAS can be used to link signals of selection to phenotypes, without information about the intermediate functional changes, it is difficult to interpret these links.

Changes in gene expression are expected to underlie many complex traits relevant to recent human evolution (Corradin et al. 2016), particularly as many signals overlap noncoding regions of the genome. We previously used predictive models of gene expression to detect changes between different ancient subsistence groups (Colbran et al. 2021) and to infer selection based on differences between present-day populations (Colbran et al. 2023). Here, we develop this idea to test for selection directly using changes in predicted expression over time inferred from ancient DNA times series. We used an approach (Joint Tissue Imputation [JTI]; Zhou et al. 2020) developed to predict gene expression from genotype data to predict the expression levels of ∼17,000 protein-coding genes in ancient (4,500–1,000 BP) and modern individuals from Britain. We then inferred significant shifts in expression levels in this 4,500 year time-transect based

Received: October 16, 2023. **Revised:** February 02, 2024. **Accepted:** March 01, 2024

Discoveries

on linear regression models of predicted gene expression against time.

This approach allows us to perform a gene-level test for selection on gene expression, identifying four novel signals of selection on gene expression resulting from small shifts in allele frequency that were not captured by genome-wide scans for selection. We are also able, for regions identified to have been under selection in this or other analyses, to identify which genes are likely to have changed their expression due to this selection—in several cases showing that selection at known loci (*LCT*, for example) has substantially affected the expression of several nearby genes. Our work demonstrates the utility in incorporating functional information into genome-wide scans for selection.

Results

Imputed Data Recovers Genome-Wide Selection Scan Results

We assembled a dataset of 91 present-day and 616 ancient (4,500–1,000 BP) individuals from Britain. Our approach assumes that the sample population is closed and homogeneous. We thus chose this population from Britain due to its small geographical spread, relatively continuous demographic history, and large aDNA sample size. Present-day individuals were from the GBR population of the 1000 Genomes project (1000 Genomes Consortium 2015). Ancient individuals had either been genotyped using the 1240k single nucleotide polymorphism (SNP) capture reagent, or shotgun sequenced and then genotyped at 1240k sites. This is the same dataset used in Mathieson and Terhorst (2022) (original sources Martiniano et al. 2016; Schiffels et al. 2016; Olalde et al. 2018; Brace et al. 2019; Margaryan et al. 2020; Patterson et al. 2022) with additional individuals from Gretzinger et al. (2022), and removing individuals with less than 0.1× coverage at 1240k sites. We calculated genotype likelihoods at 1240k sites, and then imputed diploid genotypes at 1240k sites using *beagle4* (Browning and Browning 2007). We then lifted over 1240k sites from hg19 to hg38 and imputed at ungenotyped sites using the NHLBI TOPMed imputation server (Fuchsberger et al. 2015; Das et al. 2016; Taliun et al. 2021).

As a point of comparison to the transcriptome-wide scan, we ran the SNP-based genome-wide selection scan described in Mathieson and Terhorst (2022) on the imputed diploid 1240k dataset. Briefly, this uses the *bmws* software to estimate time varying selection coefficients based on the time series of allele frequencies and reports *P*-values based on fitting a gamma distribution to the root mean squared selection coefficient, averaged in 20-SNP sliding windows. The results are largely consistent with those of Mathieson and Terhorst (2022), identifying strong evidence of selection at *LCT*, *DHCR7*, *SLC22A4*, *OAS1*, the HLA region and other loci (Fig. 1a). Despite the larger sample size and diploid (as opposed to pseudohaploid) coverage, the new analysis does not find substantially more signals of selection and the shared signals are not more significant. Although imputation of ancient DNA generally produces accurate genotype calls (Hui et al. 2020; Ausmees et al. 2022; Sousa da Mota et al. 2023), we noticed that at strongly selected sites, imputed allele frequencies were slightly biased towards present-day allele frequencies compared to pseudohaploid allele frequencies (supplementary Fig. S1, Supplementary Material online). This suggests that, while generally accurate, imputation might reduce power to detect selection because this bias has a greater effect on sites with large changes in frequency over time. In our case, this seems to offset the advantage from greater sample size. That said, the imputed data do not seem to perform worse than the pseudohaploid data and produce well-calibrated results (Fig. 1b). As imputation is necessary to perform the transcriptome-wide scan, we proceeded with the imputed dataset.

Transcriptome-Wide Scan Identifies Significant Changes in Gene Expression

We next carried out a transcriptome-wide scan for selection. We predicted expression of 17,388 protein coding genes for each ancient and present-day individual using *JTI* models trained on the genotypes and transcriptomes of 49 tissues from the GTEx project (Zhou et al. 2020). As it is difficult to determine the most relevant tissue for each gene and models across tissues are generally correlated with each other, we used the model for the tissue with the highest training R^2 for each individual gene (Colbran et al. 2023). For each gene, we fit ordinary linear regression models of expression against time to identify genes with non-neutral shifts in predicted expression levels. We did not include genetic ancestry principal components as covariates in this model, as the principal components of the ancient individuals clustered closely with the present-day individuals (supplementary Fig. S2, Supplementary Material online). Instead, we applied genomic control to account for any inflation in test statistics due to genetic drift or residual population structure. After filtering for imputation quality, 28 genes at seven loci had evidence (false discovery rate [FDR] < 0.05) for significant shifts in predicted expression (Fig. 1c, Table 1, supplementary Fig. S3, Table S2, Supplementary Material online). We confirmed that predictions for these genes were reflective of actual population-level trends by comparing predicted and observed expression across all 1kG populations (supplementary Fig. 4, Supplementary Material online). Despite a mismatch in tissues for most genes, we found that the order of population medians agreed more often than expected by chance ($p = 0.0053$).

We also replicated our results using a method called UTMOST which, like JTI, constructs tissue-specific, linear models of gene expression using expression data from multiple tissues (Alzheimer's Disease Genetics Consortium 2019). We carried out a transcriptome-wide scan based on UTMOST models trained on genotypes and

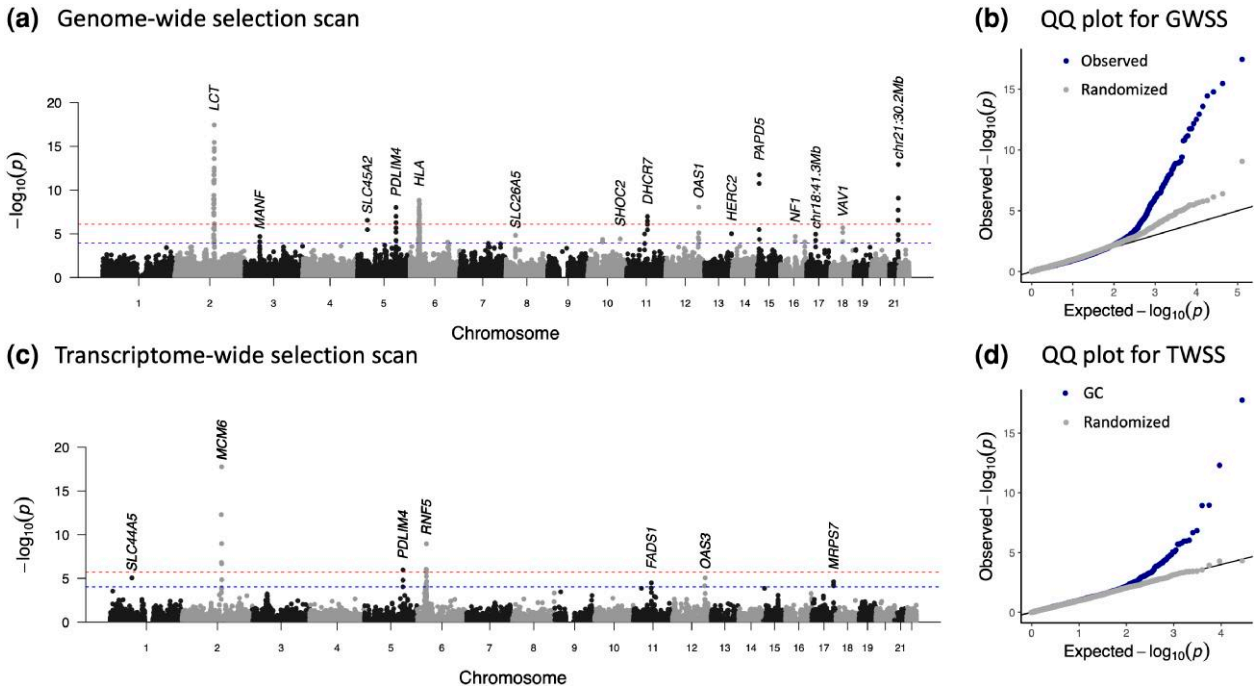

FIG. 1. Genome-wide and transcriptome-wide scans for selection. a) *P*-values for genome-wide selection. Each point represents a 20-SNP window. The blue line indicates FDR significance ($P < 10^{-4}$), and the red line indicates Bonferroni significance ($P < 10^{-6}$). FDR significant ($P < 10^{-4}$) windows are labeled with the nearest genes or known target of selection. b) QQ plot for genome-wide scan results in 20 SNP windows with points from *A* in blue, and results with dates of samples randomized in gray. c) *P*-values for transcriptome-wide selection scan. Each point represents a gene. Blue lines indicate FDR significance ($P < 10^{-4}$), and red lines indicate Bonferroni significance ($P < 10^{-6}$). The most significant gene at each locus is labeled. Five peaks were shared between the two scans: *FADS1*, *LCT*, HLA, *PDLIM4*, and *OAS1/3*. d) QQ plot for transcriptome-wide scan results with points from c) in blue and results with dates of samples randomized in gray.

transcriptomes from the GTEx project (Alzheimer's Disease Genetics Consortium 2019; Zhou et al. 2020). Sixteen out of 28 JTI significant genes replicated in the UTMOST scan, including all genes in Figs. 2 and 3 (supplementary Fig. S5, Table S1, Supplementary Material online). Of the 12 significant genes that did not replicate, 3 genes could not be modeled by UTMOST, and 1 gene was filtered out due to low imputation quality. There was broad agreement in the direction of effect predicted by the two methods (Pearson $R^2 = 0.797$, supplementary Fig. S6, Supplementary Material online). These results indicate that the JTI results are largely reproducible by other methods of predicting gene expression from genotype data.

We compared the results from the transcriptome-wide scan with the SNP-based genome-wide scan. Five peaks overlap between the two scans (Fig. 1). Similar to the genome-wide selection scan where peaks contain multiple SNPs (or windows of SNPs) in LD, the transcriptome-wide scan peaks span multiple genes due to both LD between eQTLs and coregulation of nearby genes (Wainberg et al. 2019).

In some cases where the peak is shared, the transcriptome-wide scan is able to identify the genes targeted by selection as those with the largest predicted change in expression. For instance, our approach identified *FADS1* as the only gene out of the six in the FADS region with evidence for significant changes in expression (Fig. 2a, Table 2). The genomic signal of selection in this region has been previously linked to the increased expression of *FADS1* (Ameur et al. 2012; Buckley et al. 2017; Mathieson and Mathieson 2018), which is corroborated by our results.

In other cases, the transcriptome-wide scan does not uniquely identify the targeted gene, but highlights a subset of genes in the region with significant changes. For example, at the *LCT* locus (Fig. 2b, Table 2) the transcriptome-wide scan identified 6 out of 11 genes with significant changes in expression. As expected, we predicted a significant increase in *LCT* expression (Fig. 2b), but we also predicted significant changes in five other genes. Indeed, the regulatory variants associated with adult *LCT* expression lie inside *MCM6* (Ségurel and Bon 2017), which showed the most significant shift in expression levels in this time period.

The genome-wide scan peak at the HLA region contains 148 genes, of which we predict 12 to have significant changes in expression (Fig. 2c, Table 2). The most significant signal was for *RNF5*, which is involved in the degradation of misfolded proteins and regulation of viral infection (Zeng et al. 2021; Li et al. 2023). Another signal of interest in this region is for decreased expression of *C4A*, the expression of which is associated with increased risk for schizophrenia (Yilmaz et al. 2021). These 12 genes might be priority candidates for the target of selection, but it remains possible that they are all hitchhiking and the real

Table 1. Genes with significant shifts in predicted expression (FDR < 0.05) characterized by the transcriptome-wide selection scan

Chr	Gene	Tissue	R^2	*P*-value	Beta	GWSS peak
1	*SLC44A5*	Skin Sun Exposed Lower leg	0.56520	8.827e−06	−1.178e−04	Novel
2	*TMEM163*	Kidney Cortex	0.38790	5.012e−13	1.400e−04	*LCT*
2	*MAP3K19*	Testis	0.23290	1.490e−07	1.127e−04	*LCT*
2	*LCT*	Cells EBV-transformed lymphocytes	0.08194	1.078e−09	8.239e−05	*LCT*
2	*MCM6*	Esophagus Muscularis	0.19770	1.736e−18	−1.795e−04	*LCT*
2	*DARS*	Whole Blood	0.16510	2.137e−07	−9.886e−05	*LCT*
2	*CXCR4*	Adrenal Gland	0.08339	1.457e−05	−4.051e−05	*LCT*
5	*P4HA2*	Thyroid	0.38000	9.358e−05	1.027e−04	*PDLIM4*
5	*PDLIM4*	Brain Cortex	0.28600	1.098e−06	−4.210e−05	*PDLIM4*
5	*SLC22A5*	Cells Cultured fibroblasts	0.43790	1.605e−05	−1.283e−04	*PDLIM4*
6	*PPP1R18*	Artery Aorta	0.20470	4.166e−05	−2.423e−05	HLA
6	*TUBB*	Cells EBV-transformed lymphocytes	0.02940	1.635e−06	4.196e−05	HLA
6	*PSORS1C1*	Thyroid	0.65610	7.309e−05	−7.907e−05	HLA
6	*CDSN*	Skin Sun Exposed Lower leg	0.19040	4.200e−05	4.507e−05	HLA
6	*CCHCR1*	Spleen	0.55490	9.258e−07	8.014e−05	HLA
6	*APOM*	Testis	0.14170	1.935e−06	−2.599e−05	HLA
6	*C4A*	Brain Cerebellum	0.47000	2.017e−06	−8.152e−05	HLA
6	*ATF6B*	Brain Spinal cord cervical c-1	0.46570	6.076e−06	−5.952e−05	HLA
6	*RNF5*	Colon Transverse	0.29300	1.145e−09	−6.436e−05	HLA
6	*PBX2*	Whole Blood	0.08456	1.059e−06	−3.343e−05	HLA
6	*HLA-DMA*	Cells Cultured fibroblasts	0.50690	8.493e−05	5.442e−05	HLA
6	*HLA-DPA1*	Cells Cultured fibroblasts	0.60710	2.249e−05	−1.380e−04	HLA
11	*FADS1*	Brain Cerebellum	0.47180	3.310e−05	6.236e−05	*FADS1*
12	*FAM109A*	Kidney Cortex	0.11600	6.886e−05	−1.752e−05	OAS
12	*OAS3*	Cells Cultured fibroblasts	0.42350	8.779e−06	−1.264e−04	OAS
17	*NUP85*	Brain Cerebellar Hemisphere	0.62490	7.025e−05	1.011e−04	Novel
17	*GGA3*	Breast Mammary Tissue	0.21800	4.907e−05	3.052e−05	Novel
17	*MRPS7*	Brain Cerebellar Hemisphere	0.41430	2.420e−05	−3.548e−05	Novel

Tissue indicates which tissue model was used. Note that this does not mean that the gene did not have significant shifts in other tissues, just that this tissue had the highest JTI training R^2. R^2 indicates the JTI training R^2 for these tissues. *Beta* indicates the effect size of time on expression levels in the ordinary regression models. *GWSS Peak* indicates the significant (FDR < 0.05) genome-wide selection scan peak indicated in Fig. 1 to which each gene corresponds.

target is a coding variant or an expression change in a gene that is not significant in our analysis.

Finally, in some cases, the gene with the most significant change in expression is probably not the main target of selection. For example, at the *PDLIM4* region (Fig. 2d), we predict significant changes in expression in *PDLIM4* and *P4HA2*, but Huff et al. (2012) identified a coding variant in *SLC22A4* as the target of selection. Similarly, although the target of selection at the OAS locus is thought to be a splice variant in *OAS1* carried by a Neanderthal introgressed haplotype, human cells with the introgressed haplotype displayed reduced *OAS3* expression and no changes in expression of *OAS1* or *OAS2* in response to viral immune triggers (Sams et al. 2016). Our transcriptome-wide scan captured this signal for reduced expression in *OAS3*, with no significant changes in predicted *OAS1* or *OAS2* expression (Fig. 2e).

Selection on Gene Expression not Captured by SNP-Based Scans

Most of the genes identified by the transcriptome-wide scan fell under selection scan peaks in the genome-wide scan. However, we identified four genes at two loci with evidence for significant regulatory shifts that did not (Fig. 3, Table 1). *SLC44A5* is a member of the choline transporter-like family that is highly expressed in skin, testis, and esophagus. The significant predicted change in expression is due to small coordinated shifts in frequency across many alleles (Fig. 3). *SLC44A5* is one of relatively few genes with a population-biased eQTL (GTEx Consortium 2020). Specifically, rs4606268 has a much larger effect on *SLC44A5* expression in European ancestry individuals compared to those of African ancestry, consistent with rapid evolution of the regulation of this gene in European populations. *SLC44A5* is also generally more highly expressed in lymphoblastoid cell lines (LCLs) of European ancestry, compared to African or East Asian ancestry, and LCLs derived from Northern Europeans show lower expression compared to Southern Europeans. Positive selection at *SLC44A5* has previously been reported in East Asian populations (Yasumizu et al. 2020) and both JTI predictions and observed expression suggest low expression in East Asia (supplementary Fig. S4, Supplementary Material online). Since *SLC44A5* is highly expressed in skin, and skin pigmentation experienced strong selection in both Britain and East Asia, we hypothesize that selection on *SLC44A5* expression may also be related to skin pigmentation. Choline is closely related to folate, which is broken down by UV radiation and thought to drive selection for darker skin pigmentation in high-UV regions (Jablonski and Chaplin 2010). In mice, choline partially rescues the effects of low folate in development (Craciunescu et al. 2010), so one possibility is that selection

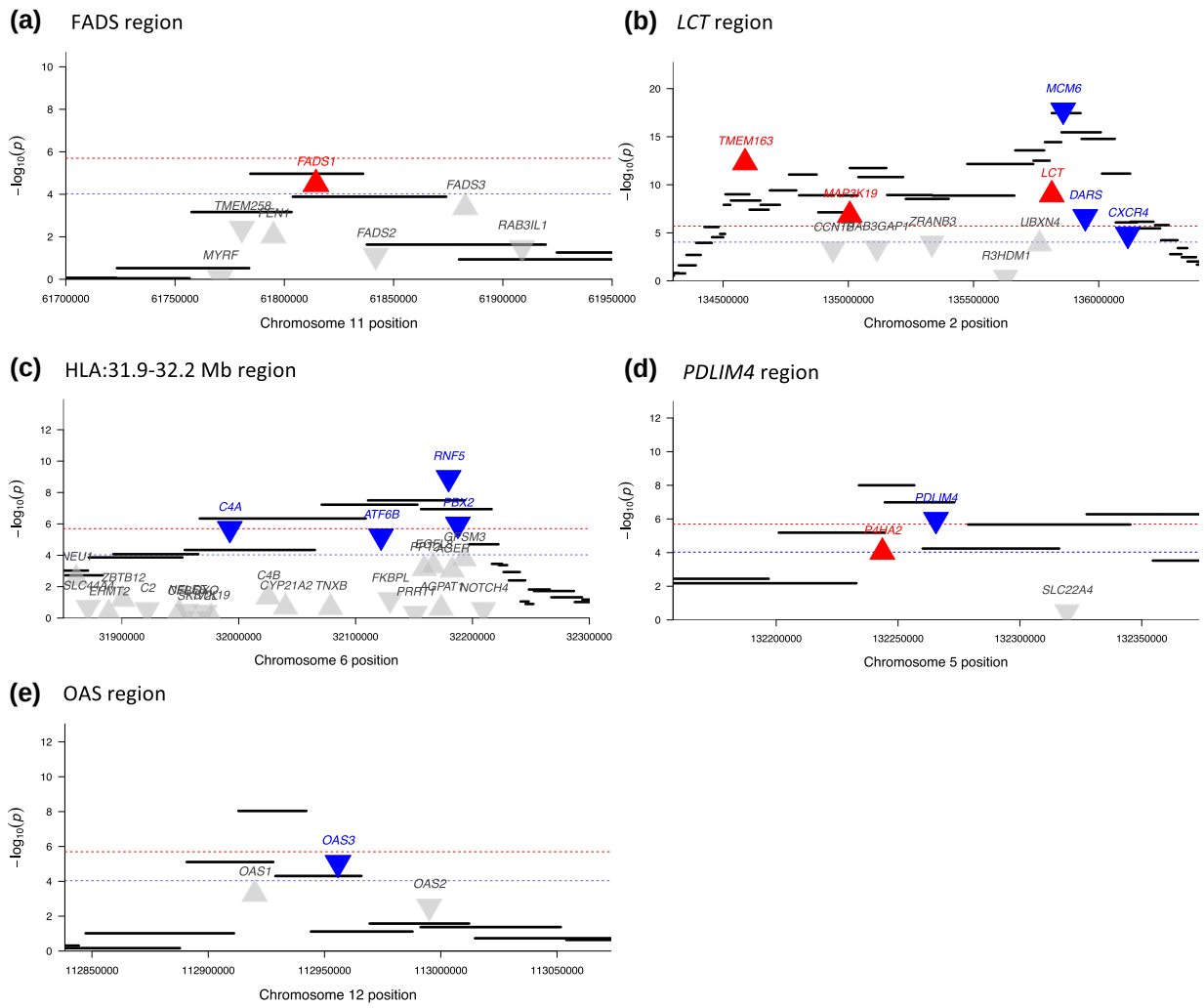

Fig. 2. Transcriptome-wide selection scan highlights genes with directional change in expression from genomic signals of selection. Each black bar represents a 20-SNP window in the genome-wide selection scan. Each triangle indicates a gene, with upturned and red indicating significant increased expression and downturned and blue indicating significant decreased expression in the transcriptome-wide scan. Blue lines indicate FDR significance ($P < 10^{-4}$) and red lines indicate Bonferroni significance ($P < 10^{-6}$) in the transcriptome-wide scan. Regions a) FADS, b) *LCT*, c) HLA, d) *PLDLIM4*, and e) OAS.

on *SLC44A5* acts to counteract the increased rate of folate degradation due to light skin pigmentation.

The second novel locus includes *NUP85*, *GGA3*, and *MRPS7*. It is likely that one of these genes is the target of selection, as all protein-coding genes within 100 kb of this signal were modeled. Of these three, *NUP85* is predicted to be upregulated in GBR compared to the other European 1kG populations, consistent with our inferred selection for increased expression on this gene in Britain, though this does not match the observed patterns in LCLs (supplementary Fig. S4, Supplementary Material online). *NUP85* encodes a part of the nucleoporin complex, which controls transport between the cytoplasm and the nucleus (Ling et al. 2022). It is involved in the recruitment and migration of immune cells through chemokine signaling (Toda et al. 2009), as well as the control of viral replication (Brass et al. 2008; Ling et al. 2022), suggesting pathogen-induced selective pressures.

Classical Selection Statistics can also be Combined with Functional Information

Although ancient DNA provides direct evidence of selection, its usefulness is limited by sample size, data quality and limited geographic and temporal availability. We therefore also explored a complementary approach of combining the JTI models with classical selection statistics based on present-day populations (Fig. 4). To do this, we generated gene-level selection statistics from SNP-level selection statistics based on the integrated haplotype score (iHS) and the singleton density score (SDS) (Voight et al. 2006; Field et al. 2016). The test statistic is a standardized weighted sum of the per-SNP selection statistics included in the predictive model for each gene weighted by effect size on normalized gene expression. As both SDS and iHS are already normalized, this weighted sum also has a standard normal distribution. We therefore calculated *P*-values based on Z scores and applied genomic control to account

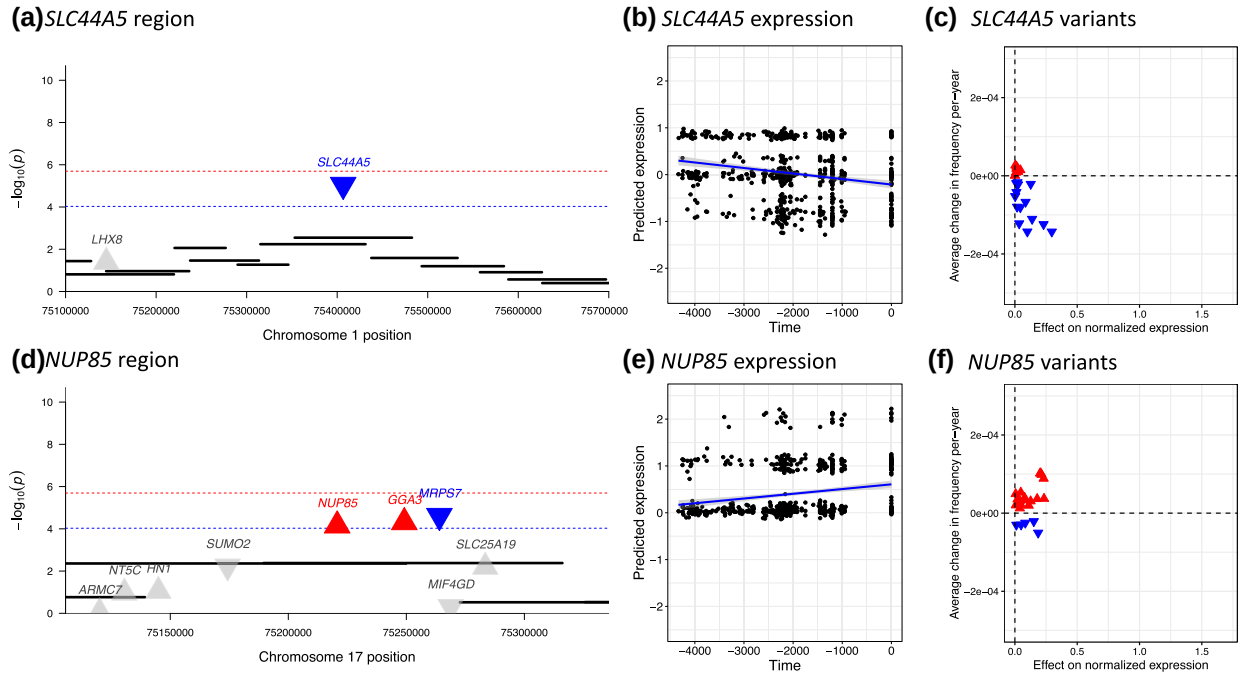

Fig. 3. Transcriptome-wide selection scan characterizes genes with directional change in expression not captured by genome-wide selection scans. a, d) Genome-wide scan for selection does not capture significant signal for selection at *SLC44A5*/*NUP85*. Each black bar represents a 20-SNP window in the genome-wide selection scan. Each triangle indicates a gene, with upturned and red indicating non-neutral increased expression and downturned and blue indicating non-neutral decreased expression in the transcriptome-wide scan. Blue lines indicate FDR significance ($P < 10^{-4}$), and red lines indicate Bonferroni significance ($P < 10^{-6}$) in the transcriptome-wide scan. b, e) *SLC44A5*/*NUP85* expression across time. The *x*-axis indicates time in years before present. The *y*-axis indicates predicted normalized expression level. Each point represents one individual. c, f) Allele frequency changes and effects of SNPs included in the prediction models for *SLC44A5*/*NUP85* across time. The *x*-axis indicates the effect of each variant on normalized expression as determined by the prediction models. The *y*-axis indicates average change in the frequency of each allele per year as calculated by a linear regression model of allele frequency against time. Each point represents an allele included in the prediction model for the gene. Red upturned triangles indicate alleles which have contributed to an increase in the expression level of the gene. Blue downturned triangles indicate alleles that decreased expression. Small but coordinated shifts in frequency across many alleles that were not captured by the genome-wide approach led to a decrease in the expression of *SLC44A5* and an increase in the expression of *NUP85*.

for residual correlation between JTI model SNPs. SDS and iHS were not available for each SNP included in our prediction models, so we removed genes for which less than half of the SNPs had scores available.

Based on SDS scores, 34 genes had significant evidence for selection (FDR < 0.05). Five of these (*MCM6*, *TMEM163*, *P4HA2*, *CXCR4*, and *SLC22A5*) were significant in both the gene-level SDS and the transcriptome-wide selection scan (Fig. 4a), while 13 genes that were significant in the transcriptome-wide scan were filtered out of the gene-level SDS analysis due to missing SDS scores. Based on iHS scores, 48 genes had FDR < 0.05, of which 5 were also significant in the transcriptome-wide scan (*TMEM163*, *MAP3K19*, *C4A*, *APOM*, and *PPP1R18*; Fig. 4c). One gene that was significant in the transcriptome-wide scan was filtered out of the gene-level iHS analysis due to missing iHS scores.

Given that the SDS scan detects selection in the last ∼ 2,000 years, while iHS captures the last ∼ 10,000 years, we expected that the SDS results would be more similar to the transcriptome-wide scan. In terms of shared significant signals, 5/34 is not significantly different than 5/48. However, all 15 genes that were significant in the transcriptome-wide scan and had gene-level SDS available had the same predicted direction of change in both analyses, while the predicted changes in gene expression from the iHS analysis were relatively uncorrelated with those predicted by the transcriptome-wide scan ($\rho = 0.0438$).

The overlap between these results shows that these different analyses do identify some of the same signals, and that functional information can be used to enhance selection scans based on standard selection statistics. However, these analyses also highlight that the information obtained from these statistics is complementary to information obtained from ancient DNA time series. While ancient DNA allows direct observation of selection and precise estimates of timing, present-day samples can be much larger and therefore more powerful, though potentially more sensitive to artifacts and model mis-specification. Different statistics may also be sensitive to different types of selection, or to selection in different time periods and we do not know how much of the difference between these analyses is due to these different factors.

Discussion

In this study, we used JTI models to detect selection on gene expression over the last 4,500 years in Britain. We identified 28 genes (FDR < 0.05) with evidence for

Table 2. Genome-wide selection signal peaks and associated genes

Chr	GWSS Start	GWSS End	TWSS Significant Genes	# All Modeled Genes	# All Genes	# Significant Genes
2	134292811	136385296	*MCM6, LCT, DARS, CXCR4, TMEM163 MAP3K19*	11	12	6
3	50849763	51387494		2	2	0
5	33777346	34069589		2	4	0
5	132101060	132454548	*P4HA2, PDLIM4, SLC22A5*	5	5	3
6	28140757	33178885	*PPP1R18, TUBB, HLA-DMA, PSORS1C1, APOM, HLA-DPA1, PBX2, RNF5, APOM, CCHCR1, CDSN, ATF6B, C4A*	148	148	12
6	128796160	129062458		0	0	0
8	33218070	33438568		1	1	0
10	49643164	50348209		5	7	0
10	110778270	111086977		4	4	0
11	61684455	61903876	*FADS1*	7	7	1
11	71302258	71592390		4	7	0
12	110796571	113044151	*OAS3, FAM109A*	17	21	2
13	111558732	111786334		0	0	0
15	27951279	29045218		1	5	0
16	49972683	50325572		2	4	0
16	82983756	83190020		0	0	0
17	30984555	31423638		4	4	0
18	41362995	41666874		0	0	0
21	43286434	44295140		11	12	0

GWSS start and *GWSS end* indicate the selection peaks in the genome-wide scan for selection. Consecutive 20-SNP windows with less than 5 Mb distance in-between were merged into single signals. Three separate signals in the HLA region were merged to one signal. 0.1 Mb buffers were added to each selection signal to include all relevant genes. *TWSS significant genes* indicates the genes within the selection signal that have evidence for non-neutral regulatory shifts in the transcriptome-wide selection scan. *# All Genes* indicates the number of all protein coding genes within the selection signal, *# All Modeled Genes* indicates the number of all protein coding genes within the selection signal that were included in the TWSS after filtering for imputation quality *# Significant Genes* indicates the number of genes that achieved significance in the transcriptome-wide scan.

selection, of which 24 were also identified by a SNP-based genome-wide selection scan on the same data. The transcriptome-wide scan identified significant shifts in predicted expression of four genes that were not captured by scans based on SNP-based genomic signatures of selection that do not incorporate functional information. Though we focused here on the application to ancient DNA data, we also demonstrated how eQTL data can be incorporated into selection scans based on present-day data, with complementary results.

The results of the transcriptome-wide scan can be interpreted in multiple ways. First, where significant genes overlap with peaks from SNP-based genome-wide scans, the transcriptome-wide scan can be used to prioritize genes that are targets of selection. This is analogous to the way in which eQTL colocalization is helpful but not a complete solution to identifying causal genes at genome-wide association peaks. The most significant gene may not be the most important, or may not have a JTI model, or the target of selection may be a coding variant (e.g. at *SLC45A2*, where we find no significant genes in the transcriptome-wide scan). Nonetheless, we find several examples where the most significant gene in the transcriptome-wide scan is the targeted gene at a genome-wide scan peak. Second, the transcriptome-wide scan can identify genes (e.g. *SLC44A5*) that are not identified in the genome-wide scan because the selection is relatively polygenic. Finally, the transcriptome-wide scan can identify the effects of selection on genes that are not themselves the target of selection. For example, selection on the expression of *LCT* affects the expression of several other genes which may themselves have functional consequences. More generally, although the interpretation of selection scans tends to focus on a single causal gene at a locus, the transcriptome-wide scan makes it clear that the linked and coregulated genes can be important and the phenotypic changes that selection acts on reflect the composite effect of many genes. For example, we predict that selection on lactase persistence significantly changed the expression of at least five other genes and the fitness consequences of the selected allele would depend on the aggregate effects of these changes.

Our approach still has several technical limitations. First, it is tissue-agnostic. Because expression is typically correlated across tissues, we focused on the tissue with the highest R^2 in our scan. However, the tissue with the highest expression is not necessarily the one that is the target of selection. More tissue-specific predictions can be used to test specific hypotheses (such as melanocytes in the case of selection on skin pigmentation; Colbran et al. 2021), but in general eQTLs may be context-specific in which case this scan could miss signals of selection entirely. Second, even for most genes, the JTI models explain only a relatively small proportion of the variance in expression

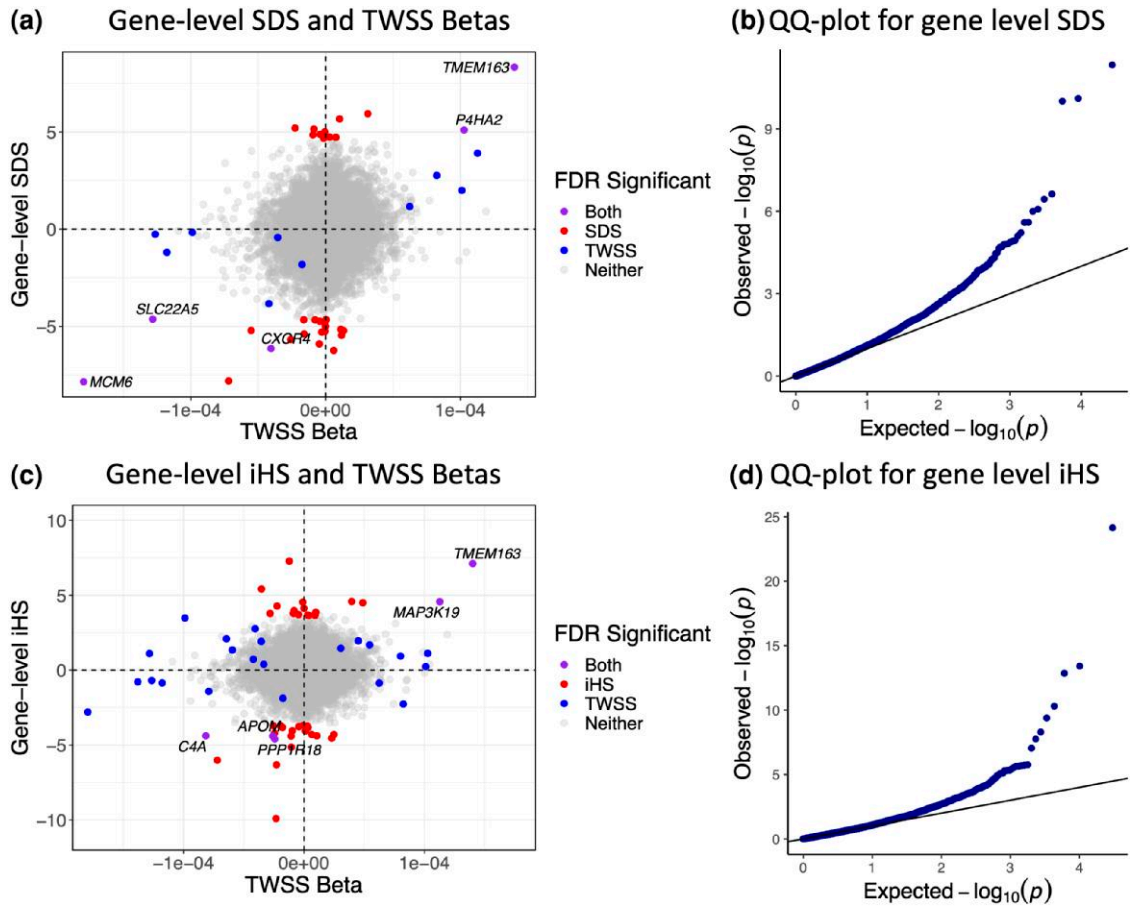

FIG. 4. iHS and SDS statistics combined with functional information reveal the gene-level consequences of selection. a, c) Gene-level SDS/iHS and average change in predicted gene expression per year. The *x*-axis indicates average change in predicted gene expression per year as measured by the beta of time in the transcriptome-wide selection scan and the *y*-axis indicates gene-level SDS/iHS values. Positive gene-level scores indicate an increase in gene expression resulting from selection, whereas negative scores indicate a decrease in expression. Each point represents a gene, with blue points indicating genes that reached FDR significance in the transcriptome-wide scan, red points indicating those that reached significance in the gene-level SDS/iHS analysis scan, gray points indicating genes that do not reach FDR significance in either scan, and purple indicating genes that achieve significance in both. There is greater concordance between the SDS analysis and the transcriptome-wide scan compared to the iHS scan, as expected by the similar time scales of the SDS and the transcriptome-wide scan analyses. b, d) QQ plot for gene-level SDS/iHS with genomic control imposed.

and include only *cis*-regulatory variants. Many genes have low (or even zero) training R^2, and we would have limited power to detect selection on those genes. Finally, since predicted expression levels are normalized, it is not possible to translate our effect size predictions into absolute expression levels, or to compare the magnitude of effects across genes. Further work would therefore be required to quantify the changes in expression at statistically significant genes, and understand the biological consequences.

Overall, this study demonstrates the potential of incorporating functional predictive models in the analysis of ancient DNA to explore the phenotypic drivers and consequences of selection. Our transcriptome-wide scan for selection provides a broad overview of the regulatory shifts associated with recent human evolution in Britain and shows how the TWAS workflow can be used to better understand the molecular basis and consequences of selection.

Methods

Data Collection and Imputation

We identified ancient individuals from Britain with genome-wide ancient DNA data, dated to within the past 4,500 years (Martiniano et al. 2016; Schiffels et al. 2016; Olalde et al. 2018; Brace et al. 2019; Margaryan et al. 2020; Gretzinger et al. 2022; Patterson et al. 2022). Most of these data had been generated using the 1240k capture reagent but some had been shotgun sequenced. We calculated genotype likelihoods at 1240k sites using a binomial model for read counts with a 1% error rate and a 5% deamination rate. We then imputed diploid genotypes at 1240k sites using *beagle4* (Browning and Browning 2007) with the 1000 Genomes reference panel (1000 Genomes Consortium 2015). We then lifted over the 1240k sites from hg19 to hg38, and imputed ungenotyped sites using the NIH TOPMed server (Fuchsberger

et al. 2015; Das et al. 2016; Taliun et al. 2021). Finally, we merged these genotypes with present-day individuals data from the GBR population of the 1,000 Genomes Project Phase 3 NYGC resequenced data using *bcftools* (Danecek et al. 2021; Byrska-Bishop et al. 2022). We removed genetic ancestry PCA outliers and individuals with less than 0.1× coverage at 1240k sites, retaining a total of 616 ancient and 91 present-day individuals.

Genome-Wide Selection Scan

We ran a genome-wide scan for selection based on selection coefficient estimation from time series aDNA data as described in Mathieson and Terhorst (2022). We used the imputed data at 1240k sites, lifted over to hg38. We started with 1,150,639 autosomal SNPs and filtered out all SNPs with MAF < 0.1, greater than 90% missingness and those with MAF $= 0$ in the ancient data leaving 409,232 SNPs. We inferred selection coefficients at each generation using a smoothing parameter $\lambda = 10^{4.5}$ and effective population size $N_e = 10^4$. We calculated root mean squared selection coefficients for 20-SNP sliding windows sliding in 10-SNP increments. We fit a gamma distribution to the window selection coefficients and computed *P*-values for each window.

Models for Predicted Gene Expression

In order to construct predictive models for gene expression, we used published JTI gene expression models, which leverage shared regulation across tissues (Zhou et al. 2020). These models were trained on common variants (MAF $>$ 0.05) for 49 tissues in version 8 of the Genotype Tissue Expression project (GTEx) (GTEx Consortium 2020). For each gene, we utilized the tissue with the highest training R^2 as described by Colbran et al. (2023). The median number of SNPs in each model was 12. The median R^2 for these models was 0.1938.

To test the robustness of our results with different modeling strategies, we also performed the analysis using predictive models from UTMOST (Alzheimer's Disease Genetics Consortium 2019), as trained for the JTI publication (Zhou et al. 2020). As for the JTI models, we chose the tissue with the highest R^2 for each gene. The median number of SNPs in each model was 15, and the median R^2 was 0.188.

Transcriptome-Wide Selection Scan

We constructed ordinary linear regression models of predicted expression against time for 17,833 protein-coding genes:

$$\text{Predicted expression} \sim \beta t, \tag{1}$$

where t indicates years before present and β indicates average change in predicted expression per year. We did not include genetic ancestry principal components as covariates in this model, as the principal components of the ancient individuals clustered closely with the modern individuals. We calculated imputation quality scores for each gene (R^2_{gene}) by taking a weighted average of the quality scores of each SNP included in the prediction model for each gene (R_i^2) with weights $|\beta_i$ equal to the absolute JTI effect size of the SNP on normalized gene expression:

$$R^2_{\text{gene}} = \frac{\sum |\beta_i| R_i^2}{\sum |\beta_i|}. \tag{2}$$

We filtered out the 20% genes with the lowest imputation quality, retaining 13,892 genes. We applied genomic control to the resulting *P*-values to account for genetic drift. We calculated an inflation factor, $\lambda = 1.791$ and divided all test statistics by λ to ensure that the median *P*-value was equal to the median *P*-value in the null χ^2 distribution (Devlin and Roeder 1999).

We also randomized the dates of the samples and re-ran the linear regression analysis to generate randomized *P*-values. We categorized genes that achieved FDR significance ($P < 0.0001$) as those with evidence for significant changes in predicted expression.

We repeated this analysis with the UTMOST models for 14,009 genes. After filtering, we retained 11,121 genes. The inflation factor was $\lambda = 1.755$. The cutoff for FDR significance was $P < 0.0001$.

Gene-Level iHS and SDS Analysis

We generated gene-level selection statistics from SNP-level classical selection statistics, the integrated haplotype score (iHS) and the singleton density score (SDS) (Voight et al. 2006; Field et al. 2016). We retrieved SDS scores calculated using data from 3,195 individuals from Britain in the UK10K dataset from Field et al. (2016). For the iHS analysis, we used data from 91 individuals from the GBR population in the 1000 Genomes Project (1000 Genomes Consortium 2015). We polarized the ancestral/derived alleles of the 1000G individuals with respect to the chimpanzee reference genome (GenBank accession: GCA_002880755.3). We then used selscan with the -norm flag to calculate normalized iHS scores (Szpiech and Hernandez 2014).

We calculated gene-level selection statistics by taking the sum of the SDS/iHS values of each SNP included in the predictive model for each gene multiplied by its effect size on normalized gene expression (β_i). SDS and iHS were not available for every SNP included in our prediction models, so we removed genes for which less than half of the SNPs had scores available. We retained a total of 13,588 genes for the SDS analysis, and 15,123 genes for the iHS analysis. As the SNP-level SDS and iHS scores were normalized, we re-normalized the gene-level selection scores by dividing by the square root of the sum of the squared effect sizes:

$$\text{SDS}_{\text{gene}} = \frac{\sum \beta_i \text{SDS}_i}{\sqrt{\sum \beta_i^2}} \tag{3}$$

$$\text{iHS}_{\text{gene}} = \frac{\sum \beta_i \text{iHS}_i}{\sqrt{\sum \beta_i^2}}, \tag{4}$$

where the sums are over all SNPs i in the model for each gene. We treated the normalized gene-level statistics as Z-scores to generate P-values, to which we applied genomic control (Devlin and Roeder 1999).

Comparison to Observed Expression

To assess how similar the predicted patterns of expression are to observed patterns, we used data from the MAGE study (Taylor et al. 2023), which performed RNA-seq analysis in LCLs from individuals in all populations from 1kG. For each gene showing significant evidence for selection, we calculated the median observed read count and median predicted expression in each population. We used a Spearman correlation across populations to calculate the agreement between the two for each gene. To test whether observed and predicted expression agreed more often than expected by chance, we summed the ρ across genes, then calculated an empirical P-value by shuffling the expression of each gene, maintaining the between-population relationships.

Supplementary Material

Supplementary material is available at *Molecular Biology and Evolution* online.

Funding

This project was supported by the National Human Genome Research Institute training grant T32HG009495 to the University of Pennsylvania (L.L.C.), the National Institute of General Medical Sciences R35GM133708 (I.M.), and a Summer Experience Grant from the Cornell University College of Arts and Sciences (L.P.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or other funders.

Conflict of Interest

None declared.

Data Availability

All data used are publicly available from original sources cited in text. Summary statistics for the transcriptome-wide scan are included as supplementary Table S1, Supplementary Material online. Code for generating figures and running analyses is available at: https://github.com/linpoyraz/predicting_functional_britain.

References

1000 Genomes Consortium. A global reference for human genetic variation. *Nature*. 2015:**526**(7571):68–74. https://doi.org/10.1038/nature15393.

Alzheimer's Disease Genetics Consortium. A statistical framework for cross-tissue transcriptome-wide association analysis. *Nat Genet*. 2019:**51**(3):568–576. https://doi.org/10.1038/s41588-019-0345-7.

Ameur A, Enroth S, Johansson A, Zaboli G, Igl W, Johansson ACV, Rivas MA, Daly MJ, Schmitz G, Hicks AA, et al. Genetic adaptation of fatty-acid metabolism: a human-specific haplotype increasing the biosynthesis of long-chain omega-3 and omega-6 fatty acids. *Am J Hum Genet*. 2012:**90**(5):809–820. https://doi.org/10.1016/j.ajhg.2012.03.014.

Ausmees K, Sanchez-Quinto F, Jakobsson M, Nettelblad C. An empirical evaluation of genotype imputation of ancient dna. *G3 (Bethesda)*. 2022:**12**(6):jkac089. https://doi.org/10.1093/g3journal/jkac089.

Brace S, Diekmann Y, Booth TJ, van Dorp L, Faltyskova Z, Rohland N, Mallick S, Olalde I, Ferry M, Michel M, et al. Ancient genomes indicate population replacement in Early Neolithic Britain. *Nat Ecol Evol*. 2019:**3**(5):765–771. https://doi.org/10.1038/s41559-019-0871-9.

Brass AL, Dykxhoorn DM, Benita Y, Yan N, Engelman A, Xavier RJ, Lieberman J, Elledge SJ. Identification of host proteins required for HIV infection through a functional genomic screen. *Science*. 2008:**319**(5865):921–926. https://doi.org/10.1126/science.1152725.

Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet*. 2007:**81**(5):1084–1097. https://doi.org/10.1086/521987.

Buckley MT, Racimo F, Allentoft ME, Jensen MK, Jonsson A, Huang H, Hormozdiari F, Sikora M, Marnetto D, Eskin E, et al. Selection in Europeans on fatty acid desaturases associated with dietary changes. *Mol Biol Evol*. 2017:**34**(6):1307–1318. https://doi.org/10.1093/molbev/msx103.

Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, Corvelo A, Clarke WE, Musunuri R, Nagulapalli K, et al. High-coverage whole-genome sequencing of the expanded 1000 genomes project cohort including 602 trios. *Cell*. 2022:**185**(18):3426–3440.e19. https://doi.org/10.1016/j.cell.2022.08.004.

Colbran LL, Johnson MR, Mathieson I, Capra JA. Tracing the evolution of human gene regulation and its association with shifts in environment. *Genome Biol Evol*. 2021:**13**(11):evab237. https://doi.org/10.1093/gbe/evab237.

Colbran LL, Ramos-Almodovar FC, Mathieson I. A gene-level test for directional selection on gene expression. *Genetics*. 2023:**224**(2):iyad060. https://doi.org/10.1093/genetics/iyad060.

Corradin O, Cohen AJ, Luppino JM, Bayles IM, Schumacher FR, Scacheri PC. Modeling disease risk through analysis of physical interactions between genetic variants within chromatin regulatory circuitry. *Nat Genet*. 2016:**48**(11):1313–1320. https://doi.org/10.1038/ng.3674.

Craciunescu CN, Johnson AR, Zeisel SH. Dietary choline reverses some, but not all, effects of folate deficiency on neurogenesis and apoptosis in fetal mouse brain. *J Nutr*. 2010:**140**(6):1162–1166. https://doi.org/10.3945/jn.110.122044.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. Twelve years of SAMtools and BCFtools. *GigaScience*. 2021:**10**(2):giab008. https://doi.org/10.1093/gigascience/giab008.

Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, Vrieze SI, Chew EY, Levy S, McGue M, et al. Next-generation genotype imputation service and methods. *Nat Genet*. 2016:**48**(10):1284–1287. https://doi.org/10.1038/ng.3656.

Dehasque M, Ávila Arcos MC, Díez-del Molino D, Fumagalli M, Guschanski K, Lorenzen ED, Malaspinas A-S, Marques-Bonet T, Martin MD, Murray G, et al. Inference of natural selection from ancient DNA. *Evol Lett*. 2020:**4**(2):94–108. https://doi.org/10.1002/evl3.165.

Devlin B, Roeder K. Genomic control for association studies. *Biometrics*. 1999:**55**(4):997–1004. https://doi.org/10.1111/biom.1999.55.issue-4.

Field Y, Boyle EA, Telis N, Gao Z, Gaulton KJ, Golan D, Yengo L, Rocheleau G, Froguel P, McCarthy MI, et al. Detection of human

adaptation during the past 2000 years. *Science*. 2016:**354**(6313):760–764. https://doi.org/10.1126/science.aag0776.

Fuchsberger C, Abecasis GR, Hinds DA. minimac2: faster genotype imputation. *Bioinformatics*. 2015:**31**(5):782–784. https://doi.org/10.1093/bioinformatics/btu704.

Gretzinger J, Sayer D, Justeau P, Altena E, Pala M, Dulias K, Edwards CJ, Jodoin S, Lacher L, Sabin S, et al. The Anglo-Saxon migration and the formation of the early English gene pool. *Nature*. 2022:**610**(7930):112–119. https://doi.org/10.1038/s41586-022-05247-2.

GTEx Consortium. The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science*. 2020:**369**(6509):1318–1330. https://doi.org/10.1126/science.aaz1776.

Huff CD, Witherspoon DJ, Zhang Y, Gatenbee C, Denson LA, Kugathasan S, Hakonarson H, Whiting A, Davis CT, Wu W, et al. Crohn's disease and genetic hitchhiking at IBD5. *Mol Biol Evol*. 2012:**29**(1):101–111. https://doi.org/10.1093/molbev/msr151.

Hui R, D'Atanasio E, Cassidy LM, Scheib CL, Kivisild T. Evaluating genotype imputation pipeline for ultra-low coverage ancient genomes. *Sci Rep*. 2020:**10**(1):18542. https://doi.org/10.1038/s41598-020-75387-w.

Jablonski NG, Chaplin G. Colloquium paper: human skin pigmentation as an adaptation to UV radiation. *Proc Natl Acad Sci U S A*. 2010:**107**(supplement_2):8962–8968. https://doi.org/10.1073/pnas.0914628107.

Li Z, Hao P, Zhao Z, Gao W, Huan C, Li L, Chen X, Wang H, Jin N, Luo Z-Q, et al. The E3 ligase RNF5 restricts SARS-CoV-2 replication by targeting its envelope protein for degradation. *Signal Transduct Target Ther*. 2023:**8**(1):53. https://doi.org/10.1038/s41392-023-01335-5.

Ling Y-H, Wang H, Han M-Q, Wang D, Hu Y-X, Zhou K, Li Y. Nucleoporin 85 interacts with influenza A virus PB1 and PB2 to promote its replication by facilitating nuclear import of ribonucleoprotein. *Front Microbiol*. 2022:**13**:895779. https://doi.org/10.3389/fmicb.2022.895779.

Marciniak S, Perry GH. Harnessing ancient genomes to study the history of human adaptation. *Nat Rev Genet*. 2017:**18**(11):659–674. https://doi.org/10.1038/nrg.2017.65.

Margaryan A, Lawson DJ, Sikora M, Racimo F, Rasmussen S, Moltke I, Cassidy LM, Jørsboe E, Ingason A, Pedersen MW, et al. Population genomics of the Viking world. *Nature*. 2020:**585**(7825):390–396. https://doi.org/10.1038/s41586-020-2688-8.

Martiniano R, Caffell A, Holst M, Hunter-Mann K, Montgomery J, Müldner G, McLaughlin RL, Teasdale MD, van Rheenen W, Veldink JH, et al. Genomic signals of migration and continuity in Britain before the Anglo-Saxons. *Nat Commun*. 2016:**7**(1):10326. https://doi.org/10.1038/ncomms10326.

Mathieson I. Limited evidence for selection at the FADS locus in Native American populations. *Mol Biol Evol*. 2020:**37**(7):2029–2033. https://doi.org/10.1093/molbev/msaa064.

Mathieson S, Mathieson I. Fads1 and the timing of human adaptation to agriculture. *Mol Biol Evol*. 2018:**35**(12):2957–2970. https://doi.org/10.1093/molbev/msy180.

Mathieson I, Terhorst J. Direct detection of natural selection in Bronze Age Britain. *Genome Res*. 2022:**32**(11-12):2057–2067. https://doi.org/10.1101/gr.276862.122.

Olalde I, Brace S, Allentoft ME, Armit I, Kristiansen K, Booth T, Rohland N, Mallick S, Szécsényi-Nagy A, Mittnik A, et al. The Beaker phenomenon and the genomic transformation of northwest Europe. *Nature*. 2018:**555**(7695):190–196. https://doi.org/10.1038/nature25738.

Patterson N, Isakov M, Booth T, Büster L, Fischer C-E, Olalde I, Ringbauer H, Akbari A, Cheronet O, Bleasdale M, et al. Large-scale migration into Britain during the Middle to Late Bronze Age. *Nature*. 2022:**601**(7894):588–594. https://doi.org/10.1038/s41586-021-04287-4.

Sams AJ, Dumaine A, Nédélec Y, Yotova V, Alfieri C, Tanner JE, Messer PW, Barreiro LB. Adaptively introgressed Neandertal haplotype at the OAS locus functionally impacts innate immune responses in humans. *Genome Biol*. 2016:**17**(1):246. https://doi.org/10.1186/s13059-016-1098-6.

Schiffels S, Haak W, Paajanen P, Llamas B, Popescu E, Loe L, Clarke R, Lyons A, Mortimer R, Sayer D, et al. Iron age and Anglo-Saxon genomes from East England reveal British migration history. *Nat Commun*. 2016:**7**(1):10408. https://doi.org/10.1038/ncomms10408.

Ségurel L, Bon C. On the evolution of lactase persistence in humans. *Annu Rev Genomics Hum Genet*. 2017:**18**(1):297–319. https://doi.org/10.1146/genom.2017.18.issue-1.

Sousa da Mota B, Rubinacci S, Cruz Davalos DI, Amorim CEG, Sikora M, Johannsen NN, Szmyt MH, Wlodarczak P, Szczepanek A, Przybyla MM, et al. Imputation of ancient human genomes. *Nat Commun*. 2023:**14**(1):3660. https://doi.org/10.1038/s41467-023-39202-0.

Szpiech ZA, Hernandez RD. selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Mol Biol Evol*. 2014:**31**(10):2824–2827. https://doi.org/10.1093/molbev/msu211.

Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, Taliun SAG, Corvelo A, Gogarten SM, Kang HM, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed program. *Nature*. 2021:**590**(7845):290–299. https://doi.org/10.1038/s41586-021-03205-y.

Taylor DJ, Chhetri SB, Tassia MG, Biddanda A, Battle A, McCoy RC. Sources of gene expression variation in a globally diverse human cohort. bioRxiv. 2023, preprint: not peer reviewed.

Toda E, Terashima Y, Sato T, Hirose K, Kanegasaki S, Matsushima K. FROUNT is a common regulator of CCR2 and CCR5 signaling to control directional migration. *J Immun*. 2009:**183**(10):6387–6394. https://doi.org/10.4049/jimmunol.0803469.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. *PLoS Biol*. 2006:**4**(3):e72. https://doi.org/10.1371/journal.pbio.0040072.

Wainberg M, Sinnott-Armstrong N, Mancuso N, Barbeira AN, Knowles DA, Golan D, Ermel R, Ruusalepp A, Quertermous T, Hao K, et al. Opportunities and challenges for transcriptome-wide association studies. *Nat Genet*. 2019:**51**(4):592–599. https://doi.org/10.1038/s41588-019-0385-z.

Yasumizu Y, Sakaue S, Konuma T, Suzuki K, Matsuda K, Murakami Y, Kubo M, Palamara PF, Kamatani Y, Okada Y. Genome-wide natural selection signatures are linked to genetic risk of modern phenotypes in the Japanese population. *Mol Biol Evol*. 2020:**37**(5):1306–1316. https://doi.org/10.1093/molbev/msaa005.

Yilmaz M, Yalcin E, Presumey J, Aw E, Ma M, Whelan CW, Stevens B, McCarroll SA, Carroll MC. Overexpression of schizophrenia susceptibility factor human complement C4A promotes excessive synaptic loss and behavioral changes in mice. *Nat Neurosci*. 2021:**24**(2):214–224. https://doi.org/10.1038/s41593-020-00763-8.

Zeng Y, Xu S, Wei Y, Zhang X, Wang Q, Jia Y, Wang W, Han L, Chen Z, Wang Z, et al. The PB1 protein of influenza A virus inhibits the innate immune response by targeting MAVS for NBR1-mediated selective autophagic degradation. *PLoS Pathog*. 2021:**17**(2):e1009300. https://doi.org/10.1371/journal.ppat.1009300.

Zhou D, Jiang Y, Zhong X, Cox NJ, Liu C, Gamazon ER. A unified framework for joint-tissue transcriptome-wide association and Mendelian randomization analysis. *Nat Genet*. 2020:**52**(11):1239–1246. https://doi.org/10.1038/s41588-020-0706-2.