Tracing the Evolution of Human Gene Regulation and Its Association with Shifts in Environment

Laura L. Colbran[1,2,*], Maya R. Johnson[3,4], Iain Mathieson[2], and John A. Capra[1,5,6,7,8,*]

[1]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, USA
[2]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, USA
[3]School for Science and Math at Vanderbilt, Vanderbilt University, USA
[4]Department of Computer Science, Bryn Mawr College, Pennsylvania, USA
[5]Department of Biological Sciences, Vanderbilt University, USA
[6]Department of Biomedical Informatics, Vanderbilt University, USA
[7]Bakar Computational Health Sciences Institute, University of California, San Francisco, USA
[8]Department of Epidemiology and Biostatistics, University of California, San Francisco, USA

*Corresponding authors: E-mails: laura.colbran@pennmedicine.upenn.edu; tony@capralab.org.

Accepted: 16 October 2021

Abstract

As humans populated the world, they adapted to many varying environmental factors, including climate, diet, and pathogens. Because many of these adaptations were mediated by multiple noncoding variants with small effects on gene regulation, it has been difficult to link genomic signals of selection to specific genes, and to describe the regulatory response to selection. To overcome this challenge, we adapted PrediXcan, a machine learning method for imputing gene regulation from genotype data, to analyze low-coverage ancient human DNA (aDNA). First, we used simulated genomes to benchmark strategies for adapting PrediXcan to increase robustness to incomplete data. Applying the resulting models to 490 ancient Eurasians, we found that genes with the strongest divergent regulation among ancient populations with hunter-gatherer, pastoralist, and agricultural lifestyles are enriched for metabolic and immune functions. Next, we explored the contribution of divergent gene regulation to two traits with strong evidence of recent adaptation: dietary metabolism and skin pigmentation. We found enrichment for divergent regulation among genes proposed to be involved in diet-related local adaptation, and the predicted effects on regulation often suggest explanations for known signals of selection, for example, at *FADS1*, *GPX1*, and *LEPR*. In contrast, skin pigmentation genes show little regulatory change over a 38,000-year time series of 2,999 ancient Europeans, suggesting that adaptation mainly involved large-effect coding variants. This work demonstrates that combining aDNA with present-day genomes is informative about the biological differences among ancient populations, the role of gene regulation in adaptation, and the relationship between genetic diversity and complex traits.

Key words: human evolution, gene regulation, machine learning.

Introduction

In the last decade, the number of ancient DNA (aDNA) samples from anatomically modern humans has increased dramatically (Marciniak and Perry 2017). These samples span time periods from several hundred to tens of thousands of years ago and provide a rich data source for understanding genetic changes and adaptations that occurred as humans expanded across the globe. However, linking genetic differences in aDNA samples to phenotypes poses several challenges (Irving-Pease et al. 2021). First, although the samples are often paired with archaeological information, this is limited to what biological material has survived for thousands of years. Thus, most phenotypes of interest are not directly measurable. Second, due the complexity of many phenotypes and gaps in our knowledge of the genetic architecture of most traits, drawing conclusions about most phenotypes of interest based on genetic information alone is challenging (Li et al. 2020; Benton et al. 2021).

Significance

Humans adapted to diverse environmental pressures as they spread around the globe, but identifying the biological systems and genetic changes underlying these adaptations remains challenging. We adapted a machine learning framework to analyze newly available ancient DNA data and predict differences in the control of genes between ancient human populations. We found differences for many genes involved in metabolism and the immune system between ancient populations with different diets and lifestyles, indicating that changes in the regulation of genes in those systems contributed to recent human adaptations. In contrast, we did not observe consistent regulatory differences associated with changes in skin pigmentation, suggesting that changes to genes themselves drove adaptation in pigmentation.

To date, most studies have focused on comparing aDNA from different geographical regions to map migrations and their relationship to archaeological changes (Skoglund and Mathieson 2018). Shifts from a hunter-gatherer lifestyle to pastoral herding and agricultural farming have been of particular interest, because these changes had profound implications for multiple aspects of life. These include changes in day-to-day activities, population density, interactions with the environment, and substantial dietary shifts, such as increased reliance on domesticated grains (Goude and Fontugne 2016; Olsson and Paik 2016). These shifts likely modified selective pressures on populations as their lifestyles, diets, and pathogen exposures changed.

Genomic scans in present-day populations have identified many loci with evidence of positive selection (Voight et al. 2006; Grossman et al. 2013; Field et al. 2016; Rees et al. 2020). In some cases, selection can be linked to changes in the coding sequence of specific genes (Lamason et al. 2005; Grossman et al. 2013). In others, it can be linked to changes in gene regulation. For example, selection at the *FADS1* locus is linked to increased expression (Buckley et al. 2017; Ye et al. 2017; Mathieson and Mathieson 2018). However in most cases, the molecular basis of signals of selection remains poorly understood, even when a specific gene can be implicated. For example, the leptin receptor (*LEPR*) is surrounded by a haplotype that has experienced recent positive selection (Voight et al. 2006), and protein-coding changes in *LEPR* have been implicated in increased cold tolerance (Hancock et al. 2008). However, altered expression of this gene is also associated with altered appetite regulation and metabolism (Loos et al. 2006; Kentish et al. 2013). Due to the difficulty in measuring environmental variables and disentangling linkage disequilibrium (LD) patterns, it remains unclear whether selection is acting on coding variants, expression changes, or both, and which environmental variable is the source of the selective pressure (Luca et al. 2010). Even these examples are exceptional; most selection signals cannot even be confidently attributed to specific genes. Selection peaks often span many genes, with little indication of which might drive changes in fitness or the underlying molecular mechanisms. This motivated us to ask whether information about variants associated with gene expression, such as expression quantitative trait loci, could help to identify genes under selection—analogous to the way in which expression quantitative trait loci data can inform variant–gene–phenotype mapping in genome-wide and transcriptome-wide association studies.

We therefore developed an approach to identify genes whose regulation shifted in coordination with lifestyle changes in recent human history. These differences in regulation between ancient human groups in distinct environments suggest adaptation. To quantify gene regulation from aDNA samples, we adapted the PrediXcan-based approach we previously used to study gene regulation in archaic hominins (Gamazon et al. 2015; Colbran et al. 2019). Since available human aDNA have variable quality and coverage, we conducted simulations and control analyses to evaluate how models for imputing gene regulation perform when applied to low-coverage data, and how to ameliorate the effects of missing variants. These yielded heuristics for determining when regulation could be accurately modeled.

Guided by these simulations, we applied PrediXcan models for thousands of genes to hundreds of ancient humans representing populations from hunter-gatherer, pastoralist, and agriculturalist lifestyles. We found enrichment for metabolic and immune pathways among the genes most divergently regulated between lifestyle groups. This reflects both the altered metabolic requirements and immune pressures of lifestyle shifts and highlights specific genes and pathways involved. For example, divergent regulation of *LEPR* suggests that its functions in metabolism and appetite regulation were relevant for recent adaptation. We also analyzed the predicted regulation of 20 diet-related genes in genomic regions with evidence of recent local adaptation. Supporting the accuracy of our approach, we rediscover the *FADS* locus regulatory haplotype that has been previously shown to vary by lifestyle and is likely the target of selection. We also identified divergent regulation between aDNA samples for selected genes involved in response to selenium (*GPX1*) and carnitine (*SLC22A5*) levels.

Modeling gene regulation using aDNA also allows us to characterize the nature of selection on specific phenotypes. To illustrate this, we investigated changes in predicted regulation of genes involved in skin pigmentation—the phenotype that is most clearly under directional selection in these populations—using PrediXcan models trained on expression data from melanocytes. We find that skin pigmentation genes show no consistent change in regulation over time suggesting that, for this particular phenotype, evolutionary change was driven by coding variants rather than regulatory changes. Overall, this work provides an atlas of imputed regulation for hundreds of ancient humans across thousands of genes to facilitate future exploration of gene regulatory shifts in recent human evolution, and demonstrates the utility of combining molecular predictive models with ancient DNA to understand the evolution of complex traits.

Results

Gene Regulatory Patterns Can Be Imputed Using Low-Coverage aDNA Data

The genetically regulated component of gene expression can be predicted by machine learning models trained on gene expression. Previous approaches have applied these models to genome-wide common variant data from present-day humans (fig. 1*A*), for example, to perform transcriptome-wide association studies (Gamazon et al. 2015; Zhou et al. 2020; Zhu and Zhou 2020), and to high-coverage archaic hominin genomes (Colbran et al. 2019). Here, we adapt this approach to enable application to low-coverage genotype data from ancient human individuals, considering the unique attributes of these data. In particular, aDNA data vary in coverage, depth, and quality. This creates a trade-off between number of individuals available for analysis and the genotype quality.

To explore this trade-off and the feasibility of this approach on available aDNA data, we created simulated ancient genomes by removing variants from present-day individuals with whole-genome sequencing from the 1000 Genomes Project (1 kG) (1000 Genomes Project Consortium 2015). (fig. 1*B* and supplementary fig. 1, Supplementary Material online; See the supplementary materials, Supplementary Material online, for detailed discussion.) First, we found that PrediXcan models trained using common variants identified from present-day whole-genome sequencing data are robust to random patterns of missing data (Spearman $\rho > 0.75$ with up to 45% of variants missing; supplementary fig. 2*A*, Supplementary Material online). However, nearly all aDNA samples used here were genotyped by targeted capture of ~1,240k variants (1240k set), largely chosen to maximize overlap with existing genotyping arrays (Fu et al. 2015; Haak et al. 2015). Furthermore, many of the ancient samples have low genotyping coverage resulting in many missing variants (supplementary fig. 2*B*, Supplementary Material online). Thus, we next matched the missing data to patterns observed in aDNA and compared the performance of different prediction models applied to full genomes versus genomes with simulated missing data (supplementary fig. 1*B*, Supplementary Material online). These models' consistency decreased substantially when applied to genomes with missing data matched to that in ancient DNA samples (median Spearman $\rho = 0.39$; fig. 1*C*).

To address this, we trained prediction models using only variants from the 1240k set. The predictions of these models were correlated with those of the full models (median Spearman $\rho = 0.67$), as expected given the LD between variants in the 1240k set and those in the full models. We also identified a set of variants that were most frequently available in the highest quality ancient samples; this resulted in a set of the 600,000 most-informative variants from the 1240k set (top600k set). We then trained models using these variants targeted to the aDNA data (top600k) and evaluated their performance on full genomes and simulated ancient genomes (see Materials and Methods). Although predictions made by the 1240k and top600k models were largely consistent with those made by the Full models when applied to genomes with no missing data (median $\rho 0.82$ and 0.79, respectively), only the 1240k models maintained consistency when applied to incomplete genomes (fig. 1*D*). We therefore concluded that the 1240k trained models strike a balance between accuracy and sample size when applied to ancient data, and thus, we used these models for the rest of our analysis.

Imputing Gene Regulatory Differences between Ancient Human Populations

We collected ancient human samples with genetic data from a variety of sequencing and genotyping platforms (see Materials and Methods). Based on the analyses in the previous section, we ranked individuals by the number of sites successfully genotyped, and took the top quartile of individuals (>771,240 SNPs, or 0.74× coverage), restricting to individuals from Eurasia due to sample density and genetic similarity to the training data (fig. 2*A*). The samples ranged in date from 90 to 45,000 years before present (yBP), with the majority between 2,500 and 6,000 yBP (fig. 2*B*).

We then assigned individuals to a lifestyle (hunter-gatherer, pastoralist, or agricultural) by literature review of the associated archaeological culture based on information from the original aDNA publications. In general, hunter-gatherers were from sites: 1) dated to times before any evidence of domestication or 2) with evidence only for foraging and meat consumption and no domesticated plants or animals. Agriculturalists were from sites with evidence for domesticated grains and animals. Pastoralists can be difficult to distinguish from agriculturists, and here refers to individuals from often seminomadic societies focused on domesticated animals (primarily the Yamnaya and similar groups). In addition, in some cases, the lifestyle distinction was based on genetic similarity to other groups, so the categories used here are based on a combination of

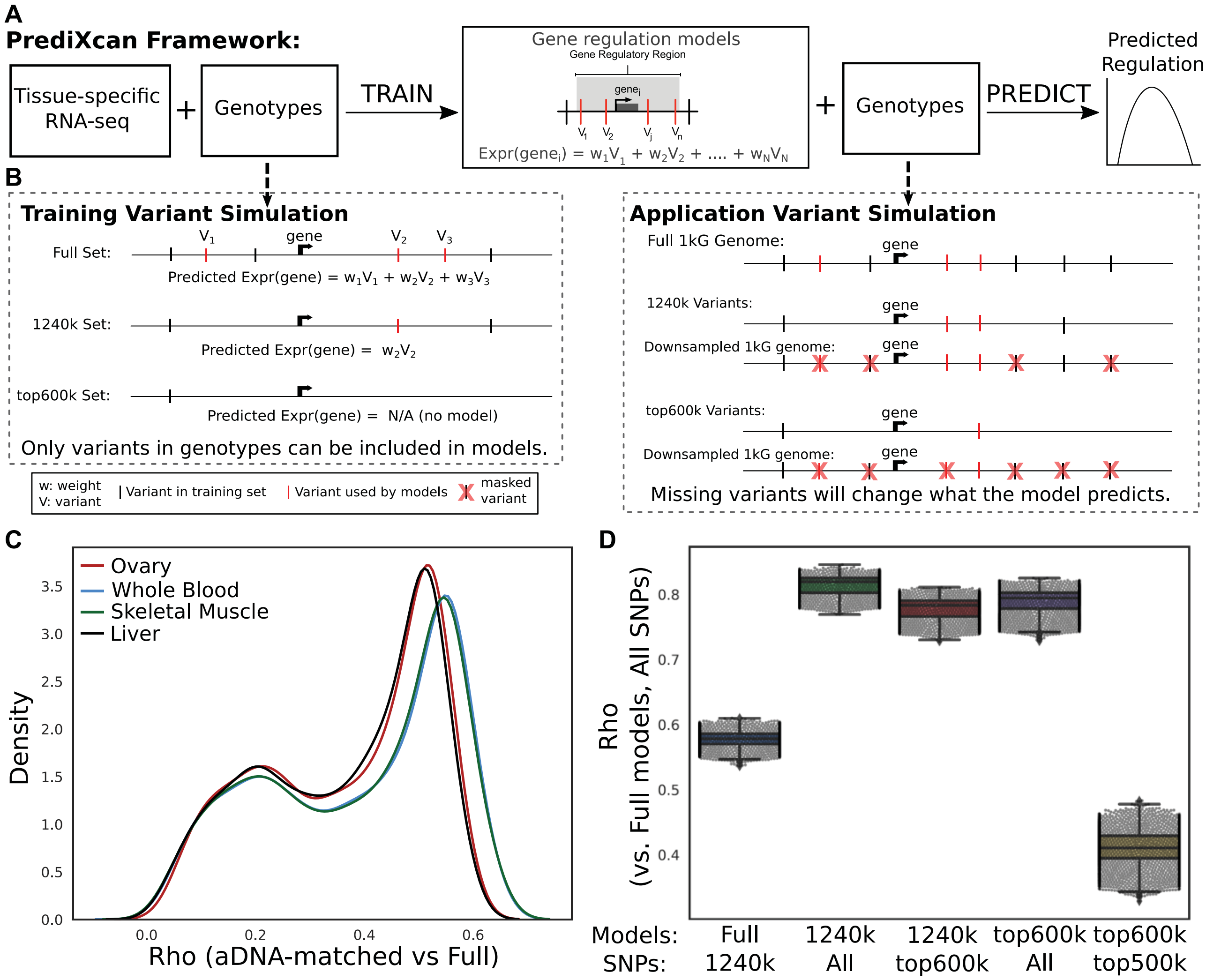

Fig. 1.—Gene regulatory prediction models can be trained for application to low-coverage ancient DNA. (*A*) Schematic of the framework for training and testing PrediXcan models. PrediXcan consists of statistical models for imputing genetic regulation of gene expression that are trained on genetic variants and normalized transcriptomes from diverse tissues collected as part of the GTEx Project. For each gene, PrediXcan considers genetic variants within 1 Mb of the gene (gray box) and uses elastic net regression to learn a combination of variants and weights to predict variance in its expression across individuals. Variants included in the final model are illustrated by red vertical lines. (*B*) To evaluate the potential for gene regulatory prediction using aDNA, we performed several analyses. First, we evaluated the effects of using three different variants for model training: full (all common variants in GTEx), 1240k (all variants in the aDNA 1240k capture set), and top600k (the 600k most representative variants from the 1240k capture set; see Materials and Methods). We also simulated the presence of missing data in the prediction phase by masking variants from genomes from the 1000 Genomes project such that only variants from each of the three sets (Full, 1240k, top600k) were available for use in prediction. (*C*) Distribution of Spearman ρ between predictions per individual in four tissues (Skeletal Muscle, Whole Blood, Liver, Ovary) when considering the complete genome versus 1240k-matched simulated ancient genomes. (*D*) ρ between predictions from a range of targeted models on down-sampled genomes to the Full PrediXcan models applied to all variants available for 1 kG individuals. Models were trained on different variant subsets (*x* axis, top row: all, 1240k, top600k) and applied to complete or downsampled 1 kG genomes (*x* axis, bottom row: all, 1240k, top600k). There is one point per individual sample.

genetics and archaeology. Because of these difficulties, we focused primarily on comparisons between hunter-gatherers and the other groups. This process resulted in 490 ancient Eurasian individuals with an assigned lifestyle and aDNA for further study (fig. 2*C*).

We then applied the 210,800 "1240k" gene regulation prediction models described in the previous section to the 490 ancient samples, as well as to 503 present-day Europeans from the 1000 Genomes Project (1000 Genomes Project Consortium 2015). This resulted in normalized expression predictions in different tissues (predicted regulation) for 14,873 unique genes. The observed expression level of a gene in a tissue in an individual is a combination of genetically regulated and environmental factors. The output of our prediction

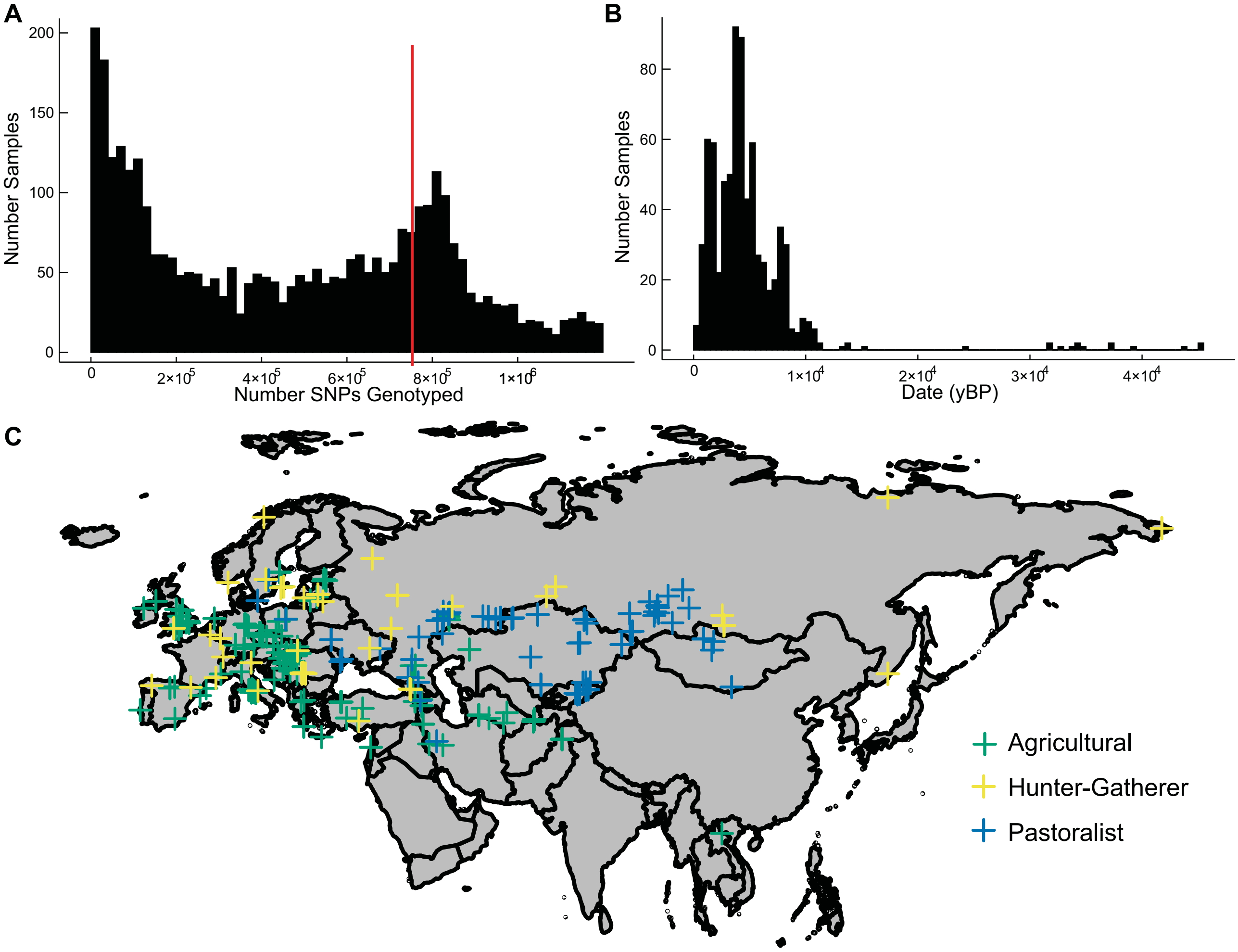

Fig. 2.—Attributes of ancient humans considered in this study. (*A*) Distribution of the number of variants with genotype call in the aDNA samples. The maximum is 1,233,013, the number of SNPs on the 1240k genotyping chip. We analyzed individuals in the fourth quartile (red line, 771,029 SNPs). (*B*) Distribution of the age of 490 Eurasian samples analyzed in yBP. (*C*) We assigned ancient Eurasians with sufficient genetic data to three lifestyles: green, agriculturalist; blue, pastoralist; yellow, hunter-gatherer.

model is not a direct proxy for the observed expression, but rather a quantification of the genetic component of gene regulation. Thus, differences in predicted regulation between individuals reflect potential differences in the inherited genetic component of expression, not environmentally driven differences.

Divergently Regulated Genes Are Enriched for Immune and Metabolic Functions

To survey high-level differences among ancient individuals from the three lifestyle groups, we identified divergently regulated genes in each tissue. Overall, 5,759 unique genes showed evidence of divergent regulation between lifestyles in at least one tissue (median 2 tissues; supplementary fig. 3*A*, Supplementary Material online), and an average of 9.8% of genes in each tissue were divergent (supplementary fig. 3*B*, Supplementary Material online). For example, *GPR84* in the adrenal gland was among the most different in predicted regulation between lifestyles (fig. 3*A*; predicted regulation of −0.0421 in agriculturalists vs. 0.197 in hunter-gatherers; corrected Kruskal–Wallis $P=2.78 \times 10^{-4}$). However, most divergent genes had relatively small changes in magnitude between groups (e.g., maximum 1.17 magnitude difference between hunter-gatherers and agriculturalists in Subcutaneous Adipose) and the majority of these differences are likely attributable to genetic drift, rather than the effects of selection. We therefore imposed a genomic control (see Materials and Methods) on the full distribution of 210,800 (genes × tissues) Kruskal–Wallis *P* values (fig. 3*A*), and report significant results based on the corrected distribution. We focused on the 500 genes with the most evidence of divergent regulation (corrected $P < 3.46 \times 10^{-3}$, FDR = 0.586), which are likely enriched for targets of selection.

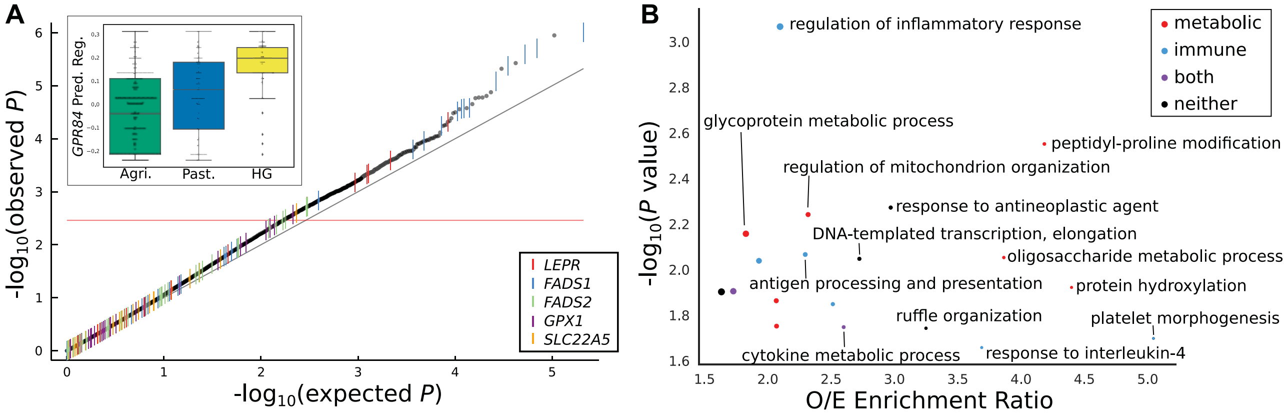

Fig. 3.—Immune and metabolic genes are among the most diverged between ancient lifestyle groups. (*A*) QQ plot for all gene regulation models in all tissues. Observed *P* values are calculated after GC correction. The 500 most divergently regulated genes have at least one model above the red line. Inset: predicted regulation of *GPR84* in adrenal gland. (*B*) The most-enriched GO terms among the 500 most diverged genes. Point size scales with number of diverged genes in each category (range 3–24).

We hypothesized that immune and metabolic traits were among those under the most selective pressure as populations transitioned between lifestyles. To identify systematic patterns in the 500 most divergently regulated genes, we conducted gene ontology (GO) overrepresentation analysis. The 20 most-enriched annotation terms (fig. 3*B*) included immune-related (e.g., antigen processing and presentation) as well as basic metabolic processes and cellular functions (e.g., glycoprotein metabolic process). In addition, we observed enrichment for several general functional annotations that appear to be driven by genes with pleiotropic immune system effects. For example, the eight genes driving the enrichment of the "DNA-templated transcription, elongation" term included *THOC5* (smallest $P = 2.6 \times 10^{-4}$), which also functions in immunity and response to stimuli through cytokine-mediated pathways (Tamura et al. 1999; Mancini et al. 2004), *ELP1* (smallest $P = 6.1 \times 10^{-4}$), which has functions in proinflammatory signaling (Cohen et al. 1998), and *AFF4* (smallest $P = 5.0 \times 10^{-4}$), a component of the super elongation complex, which is recruited in response to HIV-1 infection (He et al. 2010; Chou et al. 2013). To confirm that these trends are not influenced by the specific threshold we chose, we also ranked all results by *P* value and conducted a gene set enrichment analysis (see Materials and Methods). Again, the GO terms most enriched among significant *P* values included many immune and metabolic traits, several of which were also highlighted in the overrepresentation analysis (supplementary fig. 4, Supplementary Material online).

Many gene sets are likely to maintain similar regulatory patterns across populations, regardless of lifestyle, and these should not be enriched among the most divergently regulated genes. To test this, we quantified the enrichment of three such sets under strong functional constraint among the 500 most diverged genes between lifestyle groups across tissues: 1) genes that have experienced stabilizing selection on their levels of expression across many species (Chen et al. 2019), 2) genes responsible for core housekeeping functions (Eisenberg and Levanon 2013), and 3) genes that are intolerant to loss-of-function coding variation (LOF-intolerant) in present-day humans (Lek et al. 2016) (see Materials and Methods). As expected, LOF-intolerant genes and those under long-term stabilizing selection are not enriched (table 1). Surprisingly, housekeeping genes were slightly enriched (OR $=$ 1.33, $P = 0.0076$). By definition, housekeeping genes have ubiquitous expression across tissues, so this pattern could partially be explained by increased power to model changes in their regulation in multiple tissues. However, many housekeeping genes are also involved in basic cellular metabolism (Eisenberg and Levanon 2003), which could require fine tuning in response to changes in nutrient sources or other environmental shifts. We also tested for enrichment of genes that encode proteins that directly interact with viruses, since these genes are known to evolve rapidly (Enard et al. 2016), but we find no enrichment among the top 500 genes, suggesting that selection at these loci could be driven by coding rather than regulatory changes.

Several of the top divergently regulated genes underlying the GO functional enrichments have been implicated in local adaptation, for example, *EP300* (smallest $P = 1.7 \times 10^{-4}$) (Zheng et al. 2017) and several subunits of HLA-DQ (smallest $P = 4.1 \times 10^{-4}$ for *HLA-DQA2*) (Catassi and Catassi 2018; De Silvestri et al. 2018; Pierini and Lenz 2018). In the next two sections, we explore the connection between sequence signatures of recent adaptive evolution and divergent gene regulation with a focus on diet and skin pigmentation.

Table 1

Skin Pigmentation Genes with Nominally Significant Associations between Ancient Sample Age and Regulation

Gene	β (All) (95% CI)	*P* (All)	β (<15 ky) (95% CI)	*P* (<15 ky)
TYR	−2.05e-6 (−3.1e-6 to 9.67e-7)	0.00021	−4.7e-6 (−6.43e-6 to −2.97e-6)	1.08e-7
TRPM1	−9.93e-7 (−1.97e-6 to −2.05e-8)	0.045	−2.039e-6 (−3.59e-6 to −4.83e-7)	0.010
MITF	1.65e-6 (1.08e-7 to 3.2e-6)	0.036	2.80e-6 (3.34e-7 to 5.27e-6)	0.026
KIT	−3.62e-7 (−7.56e-7 to 3.16e-8)	0.071	−7.08e-7 (−1.33e-6 to −8.10e-8)	0.027

Note.—Betas and *P* values were calculated using a linear regression of the predicted regulation on the date, including the first ten ancestry principal components.

Changes in Gene Regulation Contributed to Adaptation to Diet between Ancient Lifestyles

Many regions of the human genome bear signatures of recent population-specific adaptive evolution. However, the phenotypic drivers and molecular mechanisms underlying these evolutionary signatures are largely unresolved. Since diet was one of the main factors that shifted with the change from hunting and gathering to farming, we hypothesized that gene regulatory changes between lifestyle groups might be the target of signals of selection at dietary genes.

We compared the predicted regulation of 20 diet-related genes in regions with evidence of population-specific local adaptation (Rees et al. 2020) between ancient human groups with different lifestyles (see Materials and Methods). Models for the 20 genes tested were enriched for lower *P* values ($P = 1.19 \times 10^{-14}$, K–S test), with four unique genes among the top 500 most diverged genes by group (supplementary table 4, Supplementary Material online).

FADS1 showed the most consistent evidence for divergent regulation between agriculturalists, pastoralists, and hunter-gatherers, with nominally significant divergence for 17 tissue models (supplementary table 2, Supplementary Material online). In each tissue, hunter-gatherers had significantly lower predicted *FADS1* levels than in agriculturalists or present-day Europeans, as would be expected from a diet containing higher levels of long-chain plasma unsaturated fatty acids (fig. 4*B*). We observed a similar trend among 32 ancient Africans, indicating that this trend is not specific to Eurasian populations (supplementary fig. 5*A*, Supplementary Material online). The variants driving these regulatory differences are in LD with the functional haplotype implicated in previous evolutionary studies (supplementary fig. 5*B* and supplementary table 3, Supplementary Material online) (Ameur et al. 2012). Overall, *FADS1* predicted regulation has increased over time in Eurasia (coefficient $P = 1.2 \times 10^{-10}$; supplementary fig. 6*A*, Supplementary Material online), which agrees with known allele frequency trajectories (Buckley et al. 2017; Ye et al. 2017; Mathieson and Mathieson 2018).

Another gene in the *FADS* gene cluster, *FADS2*, functions in the same pathway as *FADS1* and is also among the 500 most diverged genes. However it shows evidence for divergent regulation in fewer tissues than *FADS1* (supplementary table 4, Supplementary Material online), and the direction of effect is not consistent across tissues. Its presence therefore seems more likely to be due to overlap in regulatory variants with *FADS1* than to selection on *FADS2* regulation specifically. Our results further support the relevance of lifestyle differences between ancient populations in selection on the *FADS* locus and highlights the potential importance of regulatory changes of *FADS1* in human dietary adaptation.

Among the putative diet adaptation genes, *GPX1*, an antitoxin selenoprotein, and *SLC22A5*, a transporter responsible for recycling and uptake of carnitine (Console et al. 2018) (supplementary table 4 and supplementary fig. 7, Supplementary Material online) were also divergently regulated. The *GPX1* locus has experienced selective sweeps related to environmental selenium levels (White et al. 2015; Engelken et al. 2016), and has been implicated in response to viral infections (Guillin et al. 2019). Carnitine plays an important role in the transport of certain long-chain fatty acids to the mitochondria for energy production; thus, modulation of its regulation could suggest a difference in metabolism related to variation in the energy demands of different lifestyles. Both selenium and carnitine levels are likely to have differed in the primary diets of the ancient populations considered here (Flanagan et al. 2010; Mann 2018), suggesting that both as potential targets of local adaptation.

LEPR is another putatively adapted gene that has been suggested as the driver of nearby signatures of selection due to its function in appetite and cold tolerance (Voight et al. 2006; Hancock et al. 2008; Luca et al. 2010). *LEPR* was divergently regulated between lifestyle groups in the cerebellum (fig. 4*B*) (the only brain tissue with a model for *LEPR*), both adipose tissues, and several other tissues. It was consistently predicted to be downregulated in agriculturalists compared with the other two groups in each tissue (supplementary table 5, Supplementary Material online). Leptin is a hormone produced by adipose cells that suppresses appetite (Barrios-Correa et al. 2018), so this supports a possible connection between appetite regulation and the observed signatures of selection. This is particularly relevant to modern populations given the association of decreased *LEPR* function with obesity and metabolic disorders (Farooqi et al. 2007; Dehghani et al. 2018).

SLC22A5, *GPX1*, and *LEPR* also show changes regulation over the last 50,000 years in Eurasians when analyzed as a time course rather than a contrast between lifestyle groups (supplementary fig. 6*B–D*, Supplementary Material online).

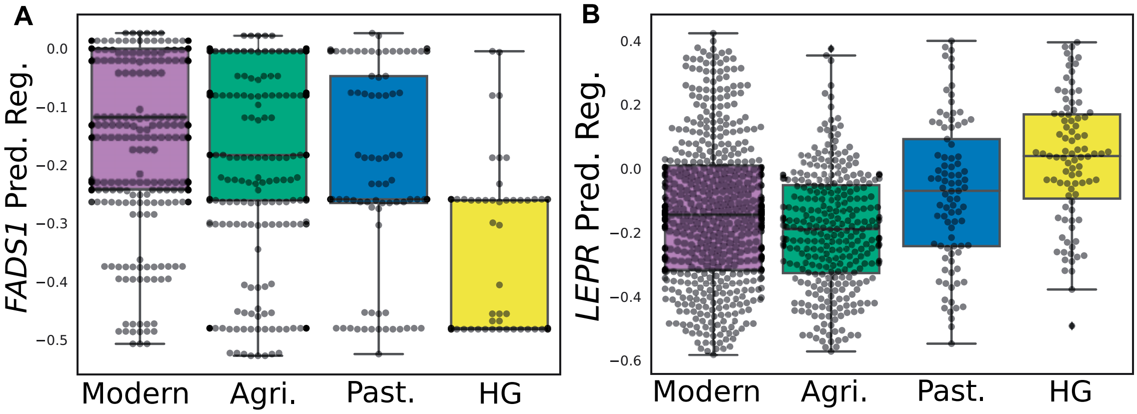

Fig. 4.—Ancient humans from different lifestyles had significant differences in regulation of key diet genes. (*A*) *FADS1* shows divergence in predicted regulation in Subcutaneous Adipose tissue between lifestyles (Kruskal–Wallis $P = 5.7 \times 10^{-24}$), as well as in eight other tissues. (*B*) *LEPR* regulation in Cerebellum is divergent across lifestyles (Kruskal–Wallis $P = 3.6 \times 10^{-17}$). Plotted with 503 present-day Europeans for comparison. Purple, present-day Europeans; green, agriculturalists; blue, pastoralists; yellow, hunter-gatherers.

The directions of change match the expectation given the changing prevalence of ancestries from the different lifestyle groups (i.e., those genes predicted to be highest in agriculturalists increase over time). Overall, these analyses suggest that recent regulatory changes made a substantial contribution to adaption to diet. More broadly, they demonstrate the potential for this method to explain observed signals of selection and to disentangle its effects on nearby genes.

Skin Pigmentation Evolution Was Not Driven by Changes in Gene Regulation in Melanocytes

We hypothesized that genes involved in complex phenotypes under selection in a population would exhibit systematic changes over time in their regulation. To test this, we focused on skin pigmentation, a trait that is known to have been under selection in humans in West Eurasia (Berg and Coop 2014; Wilde et al. 2014; Ju and Mathieson 2021) and for which many of the genes involved are well-understood (Sturm and Duffy 2012). We trained new PrediXcan models using genetic variants and gene expression in melanocytes from a diverse cohort (Zhang et al. 2018). We were able to model 17 genes known to be involved in the melanogenesis pathway (Sturm and Duffy 2012). Because skin pigmentation-associated variants changed in frequency over time, we applied these models to a time series of 2,999 ancient Europeans dated between 38,052 and 150 yBP, as well as 503 present-day Europeans from the 1000 Genomes Project and tested for systematic changes over time in predicted regulation.

Skin pigmentation genes are not enriched for differential regulation compared with all 6,923 genes modeled in melanocytes (K–S test $P = 0.53$; fig. 5*A*). Predicted regulation showed a nominally significant linear relationship with time for only four skin pigmentation genes (table 1), and only one (*TYR*) remained significant after genomic control.

We predict that *TYR*'s expression increased over time (fig. 5*B*) and is higher in non-African (particularly European) populations compared with African populations (fig. 5*C*), and in (more recent) agriculturalist populations compared with hunter-gatherers (fig. 5*D*). *TYR* encodes an enzyme important for one of the earliest steps of the melanogenesis pathway and LOF mutations cause albinism (Ghodsinejad Kalahroudi et al. 2014; Norman et al. 2017). It is therefore surprising that increased expression would be driven by selection for decreased pigmentation. One possibility is that increased expression due to gene regulatory variants compensates for the increase in frequency in Europeans of an activity-reducing coding variant (rs1042602) in *TYR* (Wilde et al. 2014). Selection on pigmentation could favor the coding variant, whereas the maintenance of other functions of the gene could require increased expression. Supporting this, rs1042602 has a positive weight in the fitted PrediXcan model showing that it in fact is associated with increased expression.

Finally, we were unable to build accurate melanocyte PrediXcan models for many known pigmentation genes. Some genes, including those with large-effect coding changes (Lamason et al. 2005; Soejima and Koda 2007), have relatively little *cis*-regulatory variation. Other known pigmentation genes with regulatory variation are not expressed in melanocytes. For example, despite its important role in melanocyte function, *KITLG* is expressed not in melanocytes, but in the dermal papillae and then transported to melanocytes (Botchkareva et al. 2001). Thus, it not modeled in our analysis. Overall, our results suggest that changes in gene regulation in melanocytes did not play a large role in the evolution of skin pigmentation in Europe. This is consistent with observations that selection signals for pigmentation-associated variants in Europe are mostly driven by a relatively small number of large-effect, coding variants

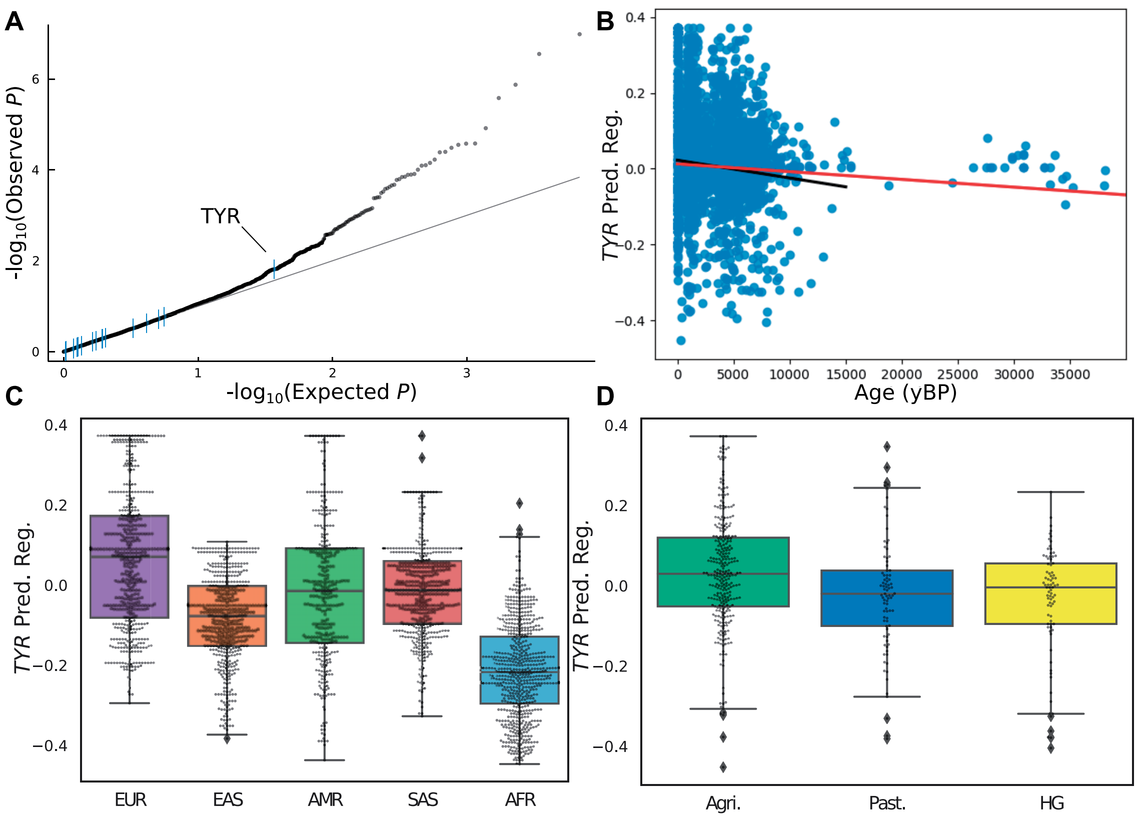

Fig. 5.—Most skin pigmentation genes show little change in regulation in the last 38,000 years in Europeans. (*A*) QQ plot with *P* values from linear regressions of date versus predicted regulation for all modeled genes in melanocytes (see Materials and Methods). The 17 skin pigmentation are highlighted in blue. (*B*) Predicted regulation of *TYR* increases over time in Europeans. The red line shows a regression calculated over all individuals, and the black regression line was calculated only over individuals <15,000 yBP. (*C*) *TYR* predicted regulation in present-day 1 kG populations, separated by continent of ancestry. (*D*) *TYR* predicted regulation in ancient Eurasians, split by lifestyle.

despite the polygenic nature of the phenotype (Ju and Mathieson 2021).

Discussion

In this study, we adapted the PrediXcan approach for modeling the genetic component of tissue-specific gene regulation and applied it to hundreds of low-coverage ancient DNA samples from individuals from three different lifestyles and to a ∼38,000-year transect of ancient Europeans. Our simulations and evaluations suggest that models of gene regulation for thousands of genes retain utility even when variant data are limited, as long as the models are trained for the specific application and their limitations properly taken into account. This is encouraging for the expansion of the PrediXcan approach to other contexts in which different variants were assayed than those used to train the original PrediXcan models. As more accurate methods are developed, it will be important to keep this aspect of their performance in mind.

Here, we found that over 5,000 genes showed evidence for divergent regulation among ancient hunter-gatherers, pastoralists, and agriculturalists in at least one tissue. The 500 genes most divergently regulated between lifestyles were enriched for metabolic and immune processes, indicating that altered gene regulation has shaped these functions during recent human evolution. Focusing on genes involved in diet, we find enrichment for divergent regulation in genes with nearby signals of recent selection, suggesting that changes in gene regulation may play a substantial role in adaptation to changes in diet.

Second, we trained new prediction models in melanocytes to analyze changes in the regulation of skin pigmentation genes in a time transect of ancient and present-day Europeans spanning 38,000 years. In contrast to genes associated with diet, we found that most genes we modeled show little to no systematic change in regulation over time, suggesting that selection on skin pigmentation mostly operated on a few large-effect coding variants. The exception, *TYR*, is predicted to have been upregulated over time, which is contrary (with respect to the trait) to the effects of a known coding variant in the gene and the predicted effects of gene expression on the trait itself (Chaki et al. 2011; Wilde et al. 2014).

However, the increased expression in Europeans may be a response to the increase in frequency of a coding variant (rs1042602) that decreases activity. These results underscore the wide variety of adaptive mechanisms in recent human evolution, and the ability of ancient DNA to illuminate these mechanisms. The other skin pigmentation genes that show nominal changes in predicted regulation over time, *MITF* and *TRPM1*, are closely linked to *TYR* in the melanogenesis pathway, with *MITF* regulating both *TYR* and *TRPM1* (D'Mello et al. 2016). Further analysis of the predicted perturbations of those relationships is needed to better understand the phenotypic consequences of these changes.

There are a several caveats to consider when interpreting these PrediXcan results. Previous work has demonstrated that, although there are some decreases in accuracy, the approach maintains utility when applied to non-European present-day populations and to archaic hominins (Colbran et al. 2019; Petty et al. 2019). Furthermore, the ancient Eurasian individuals considered here are less diverged from the GTEx cohort used for training than in these previous applications. However, due to the low coverage of the aDNA data and the focus on commonly assayed variants, there are many regulatory effects that these models do not capture. These effects would reduce our power to detect divergent regulation though they are unlikely to create false positives. In addition, the models do not capture the effects of environment (both direct and indirect) on gene expression. Therefore, although differences in predicted regulation do not necessarily indicate a change in transcript expression levels, they do the identify change in the genetic architecture of a gene's regulation. Our approach is therefore complementary to experimental assays of the regulatory effects of ancient genomic variants in present-day human cells (Weiss et al. 2021), and such approaches could be used to test our computational predictions. Another major limitation is that we are only able to draw conclusions about genes with sufficient expression and nearby present-day common variation. We also have not developed a formal test for selection on gene regulation. Although we have in some cases been able to link regulatory variation to signals of selection based on genomic data, many of the differences we observe were likely the result of genetic drift. Developing tests for selection on gene regulation that consider aDNA remains an important area for future work. Finally, our analysis does not distinguish between cases where gene regulation was divergent among the ancestral populations that differentially contributed to lifestyle groups, and cases where it changed within groups after they adopted a specific lifestyle. This is also true of the time series analysis; although all diet genes tested showed significant changes over time, this pattern is likely attributable to changes in prevalence of lifestyle groups (and relevant ancestries) rather than shifts within a continuous population.

Despite these limitations, we demonstrate the utility of considering regulatory effects of variants in combination in ancient individuals. In particular, the frequent occurrence of metabolic and immune genes among the most divergently regulated genes between ancient lifestyles underscores the contribution of gene regulation to adaptation to the substantial changes in lifestyle that the shift from nomadic hunting and gathering to stationary farming had on humans. Our targeted analysis of diet genes with evidence of results adaptive evolution further suggests that adapting to diets with different nutrient and fat compositions required population-level shifts in the regulation of many metabolic genes. In contrast, the lack of consistent gene regulatory changes in skin pigmentation genes suggests that adaptation in this trait was mainly mediated by coding variants.

Lifestyle and sun exposure are not the only variables that differ among the ancient humans with genetic information, and more diverse aDNA data are rapidly becoming available. Therefore, extending this analysis to ancient individuals across other evolutionary shifts will promising. It will also be informative to expand studies into non-European populations, both ancient and present-day, to learn when gene regulatory shifts are unique to specific populations or shared.

Overall, this study demonstrates the power of focusing evolutionary analyses on combinations of variants with established relationships to molecular phenotypes. Our approach is well-positioned to use the increasing availability of present-day and ancient genome data to provide both mechanistic explanations of selection signals and to generate hypothesis about phenotypic differences between ancient and present-day groups. Although this study focused on gene regulatory shifts in response to changes in lifestyle and temporal shifts in regulation of skin pigmentation genes, similar methods could be applied in many other questions and sets of ancient samples. Given the importance of gene regulation in recent evolution, this is a necessary step in identifying and interpreting candidate regions that have been shaped by recent human evolution. Further analyses using this approach will contribute to understanding the genome's response to large-scale environmental changes and the influence of these changes on humans today.

Materials and Methods

Ancient Genotype and Lifestyle Data

For the lifestyle analyses, we obtained ancient human genotypes from a set compiled and analyzed by the Allen Ancient DNA Resource (v42.4; accessed March 1, 2020), then lifted them over the Genome Build hg38 using liftOverPlink. We filtered out samples that did not pass their QC procedure and ranked remaining samples by genotype count (i.e., the number of variants with a genotype call), and removed all samples outside the top quartile. We also filtered samples by their continent of origin, and primarily focused on 490 ancient Eurasians to whom we could assign a lifestyle. The

FADS1 analysis additionally considered 32 ancient Africans. For a present-day comparison, we used genomes for 503 European samples from the 1000 Genomes Project (1000 Genomes Project Consortium 2015). Missing variants were assumed to be homozygous reference.

We manually assigned ancient samples to lifestyle groups by literature review based on archaeological information about the site and previous research about the associated culture. More specifically, we used lifestyles as assigned by the original publication of the sample where available. We then propagated those lifestyle labels to other samples based on the associated culture (again, as assigned by the original publication), then conducted a further literature review to match any unassigned cultures to a lifestyle based on similarity to those already matched. Samples were removed from consideration when there was not enough lifestyle-related evidence to make a call. The distinction between pastoral and agricultural groups was often difficult, and when there was ambiguity the groups were preferentially assigned to the agricultural category (supplementary file 1, Supplementary Material online).

Adapting PrediXcan for aDNA

Model Training

PrediXcan models scripts were adapted from the PredictDB Pipeline (https://github.com/hakyimlab/PredictDBPipeline; last accessed November 19, 2018). We first filtered the expression data to identify genes expressed in each tissue, then normalized and corrected for covariates such as sex and ancestry (specific procedure for each training set described below). Models were trained using an elastic net algorithm as implemented by the R glmnet library (alpha = 0.5 and nk_folds = 10), and considered all input variants that were within 1 Mb upstream or downstream of the gene in question. Fitting a model was only attempted for genes with at least two variants in that window with multiple alleles present in the training data.

Final Models for aDNA-Based Gene Regulation Prediction

The set of models used to evaluate performance in differing scenarios and for all lifestyle analyses were trained on whole-genome sequencing and RNA-seq data from GTEx v8 for 49 tissues. The genotypes included variants with a minor allele frequency >0.05 and in Hardy–Weinberg equilibrium ($P > 0.05$), and were LD pruned ($r^2 = 0.9$). The expression data were normalized by GTEx, which involved the following: genes were selected based on expression thresholds of >0.1 TPM in at least 20% of samples and at least six reads in at least 20% of samples, then expression values were normalized between samples. For each gene, expression values were normalized across samples using an inverse normal transform. Expression was then corrected for covariates including sex, sequencing protocol, sequencing platform, the first five genotyping PCs, and 15–60 PEER factors, depending on the sample size of the tissue. For each tissue, we considered only models that explained a significant amount of variance (FDR < 0.05, $r^2 > 0.01$). For the lifestyle analyses, we focused on models trained using the ~1,240,000 variants that were genotyped by first enriching for the targeted variants (1240k set) (Fu et al. 2015; Haak et al. 2015), and further required that each 1240k-trained model maintain high correlations with the original full GTEx model ($r > 0.5$) over all 2,504 1 kG individuals. All LD calculations for variants in all 1 kG Populations were made using LDLink (Machiela and Chanock 2015).

The set of models used to study skin pigmentation were trained on genotype and RNA-seq data collected from melanocytes from 106 male skin samples (Zhang et al. 2018). We imputed all genotypes to 1000 Genomes using the NIH TOPMed server (Das et al. 2016) with the following settings: ref: 1 kG Phase 3 v5; pop = other/mixed; rsq filter 0.001; phasing = eagle v2.4. We filtered genes to those with measured expression in at least ten samples, with RSEM >0.5 and >6 reads, then each gene was inverse quantile normalized to a standard normal distribution across samples. We then corrected for ancestry using the first three principal components and ten PEER factors (not sex, as all samples are male). We trained the PrediXcan models using only ~1,240,000 SNPs that were genotyped by first enriching for those targeted SNPs (1240k set) (Fu et al. 2015; Haak et al. 2015), and included any gene for which the model was able to explain a nominally significant amount of variance in the observed expression ($P < 0.05$). We focused on a set of 17 genes (Sturm and Duffy 2012) involved in skin pigmentation for which we were able to build models.

We abbreviate the 49 GTEx tissues considered as follows: Adipose—Subcutaneous, ADPS; Adipose—Visceral Omentum, ABPV; Adrenal Gland, ADRNLG; Artery—Aorta, ARTA; Artery—Coronary, ARTC; Artery—Tibial, ARTT; Brain—Amygdala, BRNAMY; Brain—Anterior Cingulate Cortex, BRNACC; Brain—Caudate, BRNCDT; Brain—Cerebellar Hemisphere, BRNCHB; Brain—Cerebellum, BRNCHA; Brain—Cortex, BRNCTX; Brain—Frontal Cortex, BRNFCTX; Brain—Hippocampus, BRNHPP; Brain—Hypothalamus, BRNHPT; Brain—Nucleus Accumbens basal ganglia, BRNNCC; Brain—putamen basal ganglia, BRNPTM; Brain–Spinal Cord Cervical C-1, BRNSPN; Brain–Substantia Nigra, BRNSN; Breast, BREAST; Cells—Transformed Fibroblasts, FIBS; Colon—Sigmoid, CLNS; Colon—Transverse, CLNT; Esophagus—Gastroesophageal Junction, ESPGJ; Esophagus—Mucosa, ESPMC; Esophagus—Muscularis, ESPMS; Heart—Atrial Appendage, HRTAA; Heart—Left Ventricle, HRTLV; Kidney Cortex, KDNY; Liver, LIVER; Lung, LUNG; Minor Salivary Gland, MNRSG; Cells-EBV-transformed Lymphocytes, LYMPH; Ovary, OVARY; Pancreas, PNCS; Pituitary, PTTY; Prostate, PRSTT; Skeletal

Muscle, MSCSK; Skin—Not sun-exposed, SKINNS; Skin—sun-exposed, SKINS; Small Intestine, SMINT; Spleen, SPLEEN; Stomach, STMCH; Testis, TESTIS; Thyroid, THYROID; Tibial Nerve, NERVET; Uterus, UTERUS; Vagina, VAGINA; Whole Blood, WHLBLD.

Evaluating Strategies for Applying PrediXcan to aDNA

To evaluate the performance of different strategies for training PrediXcan regulation prediction models and applying them to aDNA, we carried out several simulations. In the random simulations, for each percentage missing threshold, we randomly selected 20 European individuals from 1 kG (1000 Genomes Project Consortium 2015), then randomly removed that percentage of genotype calls from their genomes before applying PrediXcan models to the simulated genomes (supplementary fig. 1, Supplementary Material online). For each downsampled genome, we calculated a Spearman correlation between the predicted regulation of each gene in four tissues for the downsampled versus the full genome. Thus, each box in supplementary figure 2*A*, Supplementary Material online, has 80 (20 × 4) points. We then calculated the Spearman correlation between the median correlation between downsampled and full model predictions for each threshold and the percentage of variants missing at that threshold.

We also simulated missing data by matching patterns of missing variants from aDNA samples (supplementary fig. 1*B*, Supplementary Material online). We used 3,383 ancient human samples compiled and made available by the Allen Ancient DNA Resource on March 1, 2020 (v42.4). We selected three random Europeans from 1 kG, then for each ancient sample, we created three matching masked genomes that were missing exactly the same variants. For each masked genome, we calculated the Spearman correlation between the predicted regulation of each gene in all four tissues for the masked versus the full genome (i.e., one correlation per individual).

We also evaluated three different sets of variants for training PrediXcan models. The “full set” consisted of all variable sites identified in GTEx v8 (this included both single nucleotide variants and short indels in hg38 coordinates). The “1240k set” was formed by intersecting the full set with the variants genotyped on the 1240k chip, which totaled 714,959 variants after lifting them over to hg38. Lastly, we assembled the “top600k set” of variants, which is a subset of the 1240k set with high “support.” We calculated the “support” for each variant over N aDNA samples as $\sum_{n=1}^{N} NumVars_n$, where *NumVars* is the number of variants successfully called in sample n. In other words, support for a variant is the number of samples in which that variant was successfully genotyped, weighted by the quality (i.e., number of genotyped variants) of the sample. A variant can therefore obtain a high support either by being genotyped in many low-quality samples, or in fewer high-quality samples. We ranked the variants by their support. We identified the top 600k variants, and for the purposes of simulating the behavior of models when applied to incomplete data, we also considered the top 500k variants with the highest support (“top500k”; $N = 499{,}666$). For each set of variants, we trained a set of models and created a set of 1 kG genomes masked to only include those variants (fig. 1*A*). We assessed the performance of combinations of models and genomes by calculating the correlation of predictions made by each model-genome pair with predictions made by the Full models on the Full 1 kG Genomes (i.e., one correlation was calculated per individual 1 kG sample for each pair).

Identifying Divergent Gene Regulation between Ancient Lifestyles

To identify genes with evidence for divergence in predicted gene regulation between the three lifestyle groups, we applied a Kruskal–Wallis test for the predictions of each gene model over individuals from each group. We initially accounted for multiple testing with a Bonferroni correction within each tissue. The 5,759 genes passing this correction in at least one tissue are said to show evidence for divergent regulation; however in many cases this divergence is small and expected to be due to genetic drift. To further isolate the genes that are the most likely to be diverged due to selection rather than drift, we used genomic control to correct for population stratification by calculating the genomic inflation factor λ and recalculating the raw *P* values based on the expected value of χ^2/λ (Devlin and Roeder 1999). To focus our discussion on the genes with the strongest evidence for divergence, we sorted all models by GC-corrected *P* value and identified the top 500 unique genes (corresponding to 1,236 models), which corresponded to those with at least one model with a GC-corrected $P < 3.46 \times 10^{-3}$ and FDR = 0.586.

For select genes of interest (*FADS1*, *LEPR*, *SLC22A5*, and *GPX1*), we conducted an additional time series analysis, using all 763 Eurasians in the top quartile by coverage. We then calculated a linear regression for all samples of predicted regulation versus sample date, including the first ten principal components to control for broad-scale ancestry changes. We then did the same over just those samples <15,000 yBP since the number of older samples is somewhat limited.

Gene Set Enrichment among Diverged Genes

To conduct functional enrichment analyses on the top 500 most diverged genes, we tested for GO annotation overrepresentation using WebGestalt with default parameters (Liao et al. 2019) Specifically, we compared the biological process GO terms among the 500 most diverged genes versus all genes with a model in at least one tissue. To confirm that observed trends were not due to the particular threshold we chose, we also conducted a gene set enrichment analysis. Specifically, we ranked genes by the KW *P* value, choosing the smallest one for genes modeled in multiple tissues, then took the $\log_{10}(P)$ and used WebGestalt's GSEA implementation to identify the 20 most enriched and depleted biological process GO terms.

We also tested for enrichment of several other gene sets of interest among the top 500 diverged genes: 1) genes whose expression in particular tissues is under stabilizing selection across 17 mammalian species (Chen et al. 2019); 2) genes that are intolerant to LOF variants in their protein products (called if the upper bound of the 95% confidence interval of the observed/expected ratio is lower than 0.35) (Lek et al. 2016); 3) housekeeping genes that show consistent expression across tissues (Eisenberg and Levanon 2013); and 4) a set genes encoding virus interacting proteins (Enard et al. 2016). We calculated an odds ratio for each, and used a Fisher's exact test to determine significance. For the genes under stabilizing selection on gene expression, we considered only those tested in that study before calculating statistics.

Skin Pigmentation Time Series Data and Analysis

We obtained ancient human genome data from the Allen Ancient DNA Resource (v44.3; accessed February 8, 2021). We filtered for individual human samples from Europe (west of 59° East), and in the case of duplicate individuals chose the sample with the highest average coverage. We filled in missing dosages using the mean dosage across the other samples. This resulted in 2,999 ancient Europeans, to which we added 503 European samples from the 1000 Genomes Project (1000 Genomes Project Consortium 2015) to construct a time series ranging from 38,052 yBP to present (31 samples were older than 15,000 yBP).

To identify genes which showed a systematic change in regulation over time, we obtained predicted regulation values for each gene in each individual using the melanocyte PrediXcan models. We then regressed the predicted regulation on the date of the sample using a linear regression framework, including the first ten principal components to correct for ancestry. We further controlled for population stratification using genomic control (Devlin and Roeder 1999), and identified the skin pigmentation genes for which the effect size of date was significant (corrected $P < 0.05$). We additionally compared the predicted regulation of *TYR* in all 2,504 individuals from the 1000 Genomes Project (1000 Genomes Project Consortium 2015), separated by continental ancestry.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We would like to thank Eric Gamazon for discussions and help with data acquisition, as well as Dan Zhou for help in debugging the PrediXcan training pipeline. We would also like to thank current and former members of the Capra and Mathieson labs for useful discussions of the project and for providing comments on the figures and manuscript. L.L.C. was funded by National Institutes of Health (NIH) Grant Numbers T32GM080178 to Vanderbilt University and T32HG009495 to the University of Pennsylvania. I.M. was funded by NIH award R35GM133708. J.A.C. was funded by NIH awards R01GM115836 and R35GM127087, and the Burroughs Wellcome Fund. This work was conducted in part using the resources of the Advanced Computing Center for Research and Education at Vanderbilt University, Nashville, TN. The research is solely the responsibility of the authors and does not necessarily represent the official views of Vanderbilt University Medical Center, the National Institutes of Health, or other funding agencies.

Author Contributions

L.L.C., I.M., and J.A.C. designed the experiments and wrote the manuscript. M.R.J. designed the simulation experiments and conducted pilot simulation analyses. L.L.C., conducted all other experiments. All authors edited and approved the final manuscript.

Data Availability

All data and scripts are available on Github at https://github.com/colbrall/ancient_human_predixcan and https://github.com/colbrall/skin_pigmentation_regulation.

Literature Cited

1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. Nature 526(7571):68–74.

Ameur A, et al. 2012. 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. 90(5):809–820.

Barrios-Correa AA, Estrada JA, Contreras I. 2018. Leptin signaling in the control of metabolism and appetite: lessons from animal models. J Mol Neurosci. 66(3):390–402.

Benton ML, et al. 2021. The influence of evolutionary history on human health and disease. Nat Rev Genet. 22(5):269–283.

Berg JJ, Coop G. 2014. A population genetic signal of polygenic adaptation. PLoS Genet. 10(8):e1004412.

Botchkareva NV, Khlgatian M, Longley BJ, Botchkarev VA, Gilchrest BA. 2001. Scf/c-kit signaling is required for cyclic regeneration of the hair pigmentation unit. FASEB J. 15(3):645–658.

Buckley MT, et al. 2017. Selection in Europeans on fatty acid desaturases associated with dietary changes. Mol Biol Evol. 34(6):1307–1318.

Catassi C, Catassi GN. 2018. The puzzling relationship between human leukocyte antigen HLA genes and celiac disease. Saudi J Gastroenterol. 24(5):257–258.

Chaki M, et al. 2011. Molecular and functional studies of tyrosinase variants among Indian oculocutaneous albinism type 1 patients. J Invest Dermatol. 131(1):260–262.

Chen J, et al. 2019. A quantitative framework for characterizing the evolutionary history of mammalian gene expression. Genome Res. 29(1):53–63.

Chou S, et al. 2013. HIV-1 Tat recruits transcription elongation factors dispersed along a flexible AFF4 scaffold. Proc Natl Acad Sci U S A. 110(2):E123–E131.

Cohen L, Henzel WJ, Baeuerle PA. 1998. IKAP is a scaffold protein of the IkappaB kinase complex. Nature 395(6699):292–296.

Colbran LL, et al. 2019. Inferred divergent gene regulation in archaic hominins reveals potential phenotypic differences. Nat Ecol Evol. 3(11):1598–1606. [CrossRef][*10.1038/s41559-019-0996-x*]

Console L, Scalise M, Tonazzi A, Giangregorio N, Indiveri C. 2018. Characterization of Exosomal SLC22A5 (OCTN2) carnitine transporter. Sci Rep. 8(1):3758.

Das S, et al. 2016. Next-generation genotype imputation service and methods. Nat Genet. 48(10):1284–1287.

De Silvestri A, et al. 2018. HLA-DQ genetics in children with celiac disease: a meta-analysis suggesting a two-step genetic screening procedure starting with HLA-DQ β chains. Pediatr Res. 83(3):564–572.

Dehghani MR, et al. 2018. Potential role of gender specific effect of leptin receptor deficiency in an extended consanguineous family with severe early-onset obesity. Eur J Med Genet. 61(8):465–467.

Devlin B, Roeder K. 1999. Genomic control for association studies. Biometrics 55(4):997–1004.

D'Mello SAN, Finlay GJ, Baguley BC, Askarian-Amiri ME. 2016. Signaling pathways in melanogenesis. Int J Mol Sci. 17(7):1144.

Eisenberg E, Levanon EY. 2003. Human housekeeping genes are compact. Trends Genet. 19(7):362–365.

Eisenberg E, Levanon EY. 2013. Human housekeeping genes, revisited. Trends Genet. 29(10):569–574.

Enard D, Cai L, Gwennap C, Petrov DA. 2016. Viruses are a dominant driver of protein adaptation in mammals. eLife 5:e12469.

Engelken J, et al. 2016. Signatures of evolutionary adaptation in quantitative trait loci influencing trace element homeostasis in liver. Mol Biol Evol. 33(3):738–754.

Farooqi IS, et al. 2007. Clinical and molecular genetic spectrum of congenital deficiency of the leptin receptor. N Engl J Med. 356(3):237–247.

Field Y, et al. 2016. Detection of human adaptation during the past 2000 years. Science 354(6313):760–764.

Flanagan JL, Simmons PA, Vehige J, Willcox MDP, Garrett Q. 2010. Role of carnitine in disease. Nutr Metab 7(1):30.

Fu Q, et al. 2015. An early modern human from Romania with a recent Neanderthal ancestor. Nature 524(7564):216–219.

Gamazon ER, et al. 2015. A gene-based association method for mapping traits using reference transcriptome data. Nat Genet. 47(9):1091–1098.

Ghodsinejad Kalahroudi V, et al. 2014. Two novel tyrosinase (TYR) gene mutations with pathogenic impact on oculocutaneous albinism type 1 (OCA1). PLoS One 9(9):e106656.

Goude G, Fontugne M. 2016. Carbon and nitrogen isotopic variability in bone collagen during the Neolithic period: influence of environmental factors and diet. J Archaeol Sci. 70:117–131.

Grossman SR, et al. 2013. Identifying recent adaptations in large-scale genomic data. Cell 152(4):703–713.

Guillin OM, Vindry C, Ohlmann T, Chavatte L. 2019. Selenium, selenoproteins and viral infection. Nutrients 11(9):2101.

Haak W, et al. 2015. Massive migration from the steppe was a source for Indo-European languages in Europe. Nature 522(7555):207–211.

Hancock AM, et al. 2008. Adaptations to climate in candidate genes for common metabolic disorders. PLoS Genet. 4(2):e32.

He N, et al. 2010. HIV-1 Tat and host AFF4 recruit two transcription elongation factors into a bifunctional complex for coordinated activation of HIV-1 transcription. Mol Cell. 38(3):428–438.

Irving-Pease EK, Muktupavela R, Dannemann M, Racimo F. 2021. What can ancient DNA tell us about complex trait evolution? Front Genet. 12:703541.

Ju D, Mathieson I. 2021. The evolution of skin pigmentation-associated variation in West Eurasia. Proc Natl Acad Sci U S A. 118(1):e2009227118.

Kentish SJ, Wittert GA, Blackshaw LA, Page AJ. 2013. A chronic high fat diet alters the homologous and heterologous control of appetite regulating peptide receptor expression. Peptides 46:150–158.

Lamason RL, et al. 2005. SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. Science 310(5755):1782–1786.

Lek M, et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans. Nature 536(7616):285–291.

Li R, Chen Y, Ritchie MD, Moore JH. 2020. Electronic health records and polygenic risk scores for predicting disease risk. Nat Rev Genet. 21(8):493–502.

Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. 2019. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. Nucleic Acids Res. 47(W1):W199–W205.

Loos RJF, et al. 2006. Polymorphisms in the leptin and leptin receptor genes in relation to resting metabolic rate and respiratory quotient in the Québec Family Study. Int J Obes (Lond). 30(1):183–190.

Luca F, Perry GH, Di Rienzo A. 2010. Evolutionary adaptations to dietary changes. Annu Rev Nutr. 30:291–314.

Machiela MJ, Chanock SJ. 2015. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics 31(21):3555–3557.

Mancini A, Koch A, Whetton AD, Tamura T. 2004. The M-CSF receptor substrate and interacting protein FMIP is governed in its subcellular localization by protein kinase C-mediated phosphorylation, and thereby potentiates M-CSF-mediated differentiation. Oncogene 23(39):6581–6589.

Mann NJ. 2018. A brief history of meat in the human diet and current health implications. Meat Sci. 144:169–179.

Marciniak S, Perry GH. 2017. Harnessing ancient genomes to study the history of human adaptation. Nat Rev Genet. 18(11):659–674.

Mathieson S, Mathieson I. 2018. FADS1 and the timing of human adaptation to agriculture. Mol Biol Evol. 35(12):2957–2970.

Norman CS, et al. 2017. Identification of a functionally significant tri-allelic genotype in the Tyrosinase gene (TYR) causing hypomorphic oculocutaneous albinism (OCA1B). Sci Rep. 7(1):4415.

Olsson O, Paik C. 2016. Long-run cultural divergence: evidence from the Neolithic Revolution. J Dev Econ. 122:197–213.

Petty LE, et al. 2019. Functionally oriented analysis of cardiometabolic traits in a trans-ethnic sample. Hum Mol Genet. 28(7):1212–1213.

Pierini F, Lenz TL. 2018. Divergent allele advantage at human MHC genes: signatures of past and ongoing selection. Mol Biol Evol. 35(9):2145–2158.

Rees JS, Castellano S, Andrés AM. 2020. The genomics of human local adaptation. Trends Genet. 36(6):415–428.

Skoglund P, Mathieson I. 2018. Ancient human genomics: the first decade. Annu Rev Genom Hum Genet. 198(April):1–824.

Soejima M, Koda Y. 2007. Population differences of two coding SNPs in pigmentation-related genes SLC24A5 and SLC45A2. Int J Legal Med. 121(1):36–39.

Sturm RA, Duffy DL. 2012. Human pigmentation genes under environmental selection. Genome Biol. 13(9):248.

Tamura T, et al. 1999. FMIP, a novel Fms-interacting protein, affects granulocyte/macrophage differentiation. Oncogene 18(47):6488–6495.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. PLoS Biol. 4(3):e72.

Weiss CV, et al. 2021. The cis-regulatory effects of modern human-specific variants. eLife 10:e63713.

White L, et al. 2015. Genetic adaptation to levels of dietary selenium in recent human history. Mol Biol Evol. 32(6):1507–1518.

Wilde S, et al. 2014. Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y. Proc Natl Acad Sci U S A. 111(13):4832–4837.

Ye K, Gao F, Wang D, Bar-Yosef O, Keinan A. 2017. Dietary adaptation of FADS genes in Europe varied across time and geography. Nat Ecol Evol. 1(7):167.

Zhang T, et al. 2018. Cell-type specific eQTL of primary melanocytes facilitates identification of melanoma susceptibility genes. Genome Res. 28:1621–1635.

Zheng W-S, et al. 2017. EP300 contributes to high-altitude adaptation in Tibetans by regulating nitric oxide production. Zool Res. 38(3):163–170.

Zhou D, et al. 2020. A unified framework for joint-tissue transcriptome-wide association and Mendelian randomization analysis. Nat Genet. 52(11):1239–1246.

Zhu H, Zhou X. 2020. Transcriptome-wide association studies: a view from Mendelian randomization. Quant Biol.

Associate editor: Emilia Huerta-Sanchez