Archaic introgression contributed to the pre-agriculture adaptation of vitamin B1 metabolism in East Asia

Highlights

- SLC19A2, SLC35F3, and SLC35F4 incurred positive selection at least 10,000 years ago.
- The putative adaptive haplotype at SLC35F4 is likely of Neanderthal origin.
- The putative adaptive haplotype at SLC35F3 is of archaic ancestry with unknown origin.
INTRODUCTION

Thiamine, also well known as vitamin B1, is a water-soluble vitamin absorbed through thiamine transporter proteins in the proximal gastrointestinal tract. It plays a crucial role in cell functioning, energy metabolism, and proper immune function. Thiamine has been shown to reduce the risk of several diseases, such as type 2 diabetes, cardiovascular disease, and aging-related disorders. Particularly, thiamine, when used as adjunctive therapy, has potential survival benefits in critically ill patients with COVID-19. Thiamine must be obtained from the diet, which can be problematic due to a high turnover rate and limited body stores. These stores of thiamine in the body become rapidly depleted with decreased intake and metabolic stress. So a regular supply is needed to maintain adequate thiamine levels. A recent study suggested that many populations worldwide are at risk of clinical or subclinical thiamine deficiencies, due to famine, reliance on thiamine-low staple crops, or food preparation practices, such as milling grains and washing milled rice. More specifically, severe thiamine deficiency is recognized as a problem affecting mostly infants in low-income communities of South and Southeast Asia. Thiamine deficiency is common in alcoholics or patients with high-calorie malnutrition in high-income countries. Thiamine deficiency is also common in subjects with some diseases, such as obesity and diabetes.

The functions of thiamine, the global prevalence of thiamine deficiency, and the usage of thiamine supplementation to eliminate thiamine deficiency have attracted attention. However, limited efforts have been made in studying the evolutionary history of genes involved in thiamine metabolism or the precise requirement for thiamine in people with different genetic backgrounds. The two most well-known thiamine transporter genes are SLC19A2 and SLC19A3. The region encompassing SLC19A2 has been reported to be under positive selection in East Asian populations, although the mechanism has not been characterized. Interestingly, SLC35F3, encoding a potential thiamine transporter, has also been extensively reported to be under positive selection in East Asians. A single nucleotide polymorphism (SNP) (rs12049593) in the intron of SLC35F3 is associated with protein-calorie malnutrition, and this SNP carries an allele from the Neanderthal population. However, rs12049593 does not show the signal of positive selection in East Asia. It remains unclear whether there is any connection between the SNP’s under positive selection and archaic introgression. In addition, SLC35F4, a paralog to SLC35F3, has been reported to be adaptively introgressed in East Asian populations. These genes have been reported to be under positive selection in separate studies but have not been examined together in a single study. Moreover, limited efforts have been taken to study the evolutionary history of genes involved in thiamine metabolism or the precise requirement for thiamine in people with different genetic backgrounds.

SUMMARY

Thiamine (vitamin B1) is an essential micronutrient. Genes involved in thiamine metabolisms, such as SLC19A2, SLC35F3, and SLC35F4, were assumed to be underlying positive selection in East Asians, but the detailed mechanism remains unknown. Here, we analyzed genome data of 3,823 individuals representing 223 global populations and identified the adaptive haplotypes at thiamine genes. Interestingly, the putative adaptive haplotype at SLC35F4 was of Neanderthal ancestry, while that at SLC35F3 was also likely of archaic origins. Leveraging new methods and available ancient DNA data, we further demonstrated that the beneficial haplotypes reached a high frequency at least 10,000 years ago and are maintained persistently in present-day East Asians. We argue that pathogens, rather than agriculture developed ~10,000 years ago in East Asia, were likely the initial driving force of the putative positive selection. Notably, the first American people did not carry the putative adaptive haplotype at SLC35F4.
been made to identify the genetic origin or to infer the evolutionary history of these genes. Some important questions remain debatable: What are the evolutionary connections of SLC19A2, SLC35F3, and SLC35F4? When did the positive selection occur on these genes? Did positive selection occur simultaneously or independently?

In this study, we systematically analyzed seven protein-coding genes involved in thiamine metabolism. We refined the evidence of positive selection in SLC19A2, SLC35F3, and SLC35F4 in East Asia, identified the putative adaptive haplotypes, and estimated the frequency trajectories of these putative adaptive alleles to infer what might drive this positive selection using both the whole-genome sequencing data from present-day populations and large available ancient DNA data.

RESULTS

Genes involved in thiamine metabolism exhibited signals of positive selection in East Asia

The SLC19A2 gene encodes thiamine transporter 1 (THTR1). Our results showed that the genomic region encompassing the SLC19A2 gene exhibited the highest differentiation between East Asian and other continental populations (the empirical \( p = 3.8 \times 10^{-5} \)) (Figures 1A and S2). This genomic region also showed a signal of local adaptation in the test based on extended haplotype homozygosity (the empirical \( p = 0.023 \) for iHS\(^{17} \) and \( p = 0.004 \) for XP-EHH\(^{11} \)) (Figures S2 and S3). Within this genomic region, the SNP rs724048 exhibited the highest locus-specific branch length (LSBL)\(^{18} \) of 0.879 in the East Asian population (i.e. HAN) with the European population (i.e. Utah Residents (CEPH) with Northern and Western European Ancestry [CEU]) and African population (i.e. Yoruba in Ibadan, Nigeria [YRI]) as reference populations (Figure 1B). The derived allele G of rs724048 reached a high frequency in East Asia (e.g. 81.6% in Han Chinese in Beijing, [CHB]), was low in Europe (e.g. 1.5% in CEU), and was nearly absent in Africa (e.g. 0.9% in YRI) in the 1000 Genomes Project.\(^{19} \) It also reached a high frequency in Oceanian populations (e.g. 27.8% in PapuanHighlands, 50% in PapuanSepik, and 83.3% in Bougainville) in the Human Genome Diversity Project (HGDP)\(^{20} \) dataset (Figure 1C). We focused on SNPs with LSBL larger than the 0.1% threshold of the empirical distribution of the whole genome (i.e. 0.55) and these SNPs were considered extremely highly differentiated SNPs in this study. These 205 extremely highly differentiated SNPs spanned at least 489 kb (chr1: 169,087,494-169577333), which was larger than 400 kb inferred in the previous study.\(^{11} \)

The SLC35F3 gene encodes a potential thiamine transporter.\(^{12} \) The genomic region where SLC35F3 is located exhibited the second highest signal of genetic differentiation (the empirical \( p = 7.6 \times 10^{-5} \)) between East Asian and other continental populations (Figures 1A and S2), suggesting this gene might be subject to positive selection in East Asia. This genomic region also showed a signal of positive selection in the tests based on the extended haplotype homozygosity (the empirical \( p = 0.01 \) for iHS and \( p = 0.04 \) for XP-EHH) (Figures S2 and S4). Within this genomic region, the SNP rs201774594 exhibited the highest LSBL of 0.83 in the East Asian population (i.e. HAN) with the European population (i.e. CEU) and African population (i.e. YRI) as reference populations (Figure 1D). The derived allele A of rs201774594 reached a high frequency in East Asian populations (e.g. 89.8% in CHB), was low in European populations (e.g. 8.6% in CEU), and was nearly absent in African populations (e.g. 0.5% in YRI) in the 1000 Genomes Project. It also reached a high frequency in Oceanian populations (e.g. 50% in PapuanHighlands, 100% in PapuanSepik, and 91% in Bougainville) in the HGDP dataset (Figure 1E). This might be expected as the SLC35F3 gene has been previously reported to be subject to positive selection in East Asian populations.\(^{13,14} \)

SLC35F4 is a paralog of SLC35F3 with 92% sequence similarity. And little is known about its function. One previous study performed extensive coalescent simulations across a range of demographic models and identified 44 adaptive introgressed loci in East Asia.\(^{16} \) One locus overlapped with the genomic region where SLC35F4 is located. The genomic region where the SLC35F4 gene is located did not reach the top 1% genome-wide threshold of significance (the empirical \( p = 0.12 \)) in the LSBL test (Figure 1A). This is because this region did not contain extremely highly differentiated SNPs (Figures 1F and S5). If we relaxed the criteria, we focused on SNPs with LSBL larger than the 1% threshold (i.e. 0.36), which were considered highly differentiated SNPs. The proportion of highly differentiated SNPs in this region was significantly high (the empirical \( p = 0.013 \)) (Figure 1F), which might also indicate this region was subject to positive selection in East Asia. The derived allele A of rs17093943 reached a high frequency in East Asian populations (e.g. 45.1% in CHB), was low in European populations (e.g. 1% in CEU), and was absent in...
African populations (e.g. 0% in YRI). The frequency of the A allele was absent in Papuan Highlands and Papuan Sepik and was 9.1% in Bougainville in Oceania in the HGDP dataset (Figure 1G).

The genomic region at \texttt{SLC35F3} exhibited a signal of archaic introgression

We further investigated the genetic origin of the putative adaptive haplotype. We first focused on these extremely highly differentiated SNPs. For \texttt{SLC35F3}, these 118 extremely highly differentiated SNPs spanned more than 185 kb (chr1: 234219614-234404775) (Figure 2). For 76 of 118 SNPs, the putative adaptive alleles in East Asia were absent or rare (less than 0.5%) in the African population (i.e. YRI) and were putatively archaic introgressed alleles in all five East Asian populations from the 1000 Genomes Project. Variants carrying the putatively introgressed alleles were referred to as archaic informative variants in this study (STAR METHODS). This proportion of 64.4% (76/118) of the archaic informative variants accounting for the total extremely highly differentiated SNPs at this genomic region was significantly higher (Chi-square test, \( p < 2.2 \times 10^{-16} \)) than that of 1.5% (156/10,285) of the whole genome level. These results suggested the putative adaptive haplotype might be from some archaic introgression. Actually, this extremely highly differentiated 185 kb region overlapped with segments identified by ArchaicSeeker 2.0 (64 kb) and

Figure 1. Genes involved in thiamine metabolism exhibit signals of positive selection in East Asia

See also Figures S1, S2, S3, S4 and S5.

(A) The Manhattan plot of East-Asian specific index of divergence (LSBL) across the whole genome of comparisons among Han Chinese, European, and African populations. Each point indicates one fixed non-overlapping 100 kb window. The colored points indicate the windows overlapped with thiamine-related genes. The blue and red dashed lines indicate the 1% and 0.1% thresholds of the empirical distribution of the whole genome.

(B, D, and F) The signals of positive selection for \texttt{SLC19A2}, \texttt{SLC35F3}, and \texttt{SLC35F4}, respectively. Each point indicates one variant. The purple triangle indicates this is an archaic informative variant. The blue and red dashed lines indicate the 1% and 0.1% threshold of the whole genome, respectively.

(C, E, and G) The allele frequency distribution of rs724048 at the region encompassing \texttt{SLC19A2}, rs201774594 at \texttt{SLC35F3}, and rs17093943 at \texttt{SLC35F4} in the Human Genome Diversity Project dataset.
Sprime21 (167 kb) (Figure 2). Next, we compared these putative adaptive alleles with four known high-coverage archaic genomes (i.e. Altai Neanderthal, Vindija Neanderthal, Chagyrskaya Neanderthal, and Altai Denisovan). To avoid bias introduced by sequencing coverage, we applied the masks provided with the data, which filter sites with coverage depth below 10, mapping quality below 25 or ArchaicSeeker SPrime 01 0 2 0 3 0 (cM / Mb).

**Figure 2.** The SLC35F3 genomic region showed a signal of archaic introgression

(A) Each vertical line indicates one variant. A purple vertical line indicates an archaic informative variant. A black vertical line indicates an extremely highly differentiated variant. The purple point indicates the variant rs12049593.

(B) Each horizontal line indicates the segment exhibited a signal of archaic introgression in East Asia using different methods. The methods were annotated above the lines.

(C) The haplotype for four high-coverage archaic genomes. From top to bottom, the populations are Altai Denisovan, Altai Neanderthal, Chagyrskaya Neanderthal, and Vindija Neanderthal.

(D) Each horizontal line indicates one haplotype and is organized by population. The gray color indicates the allele is an ancestral state and the brown color line indicates a derived state.

(E) The recombination rate at this region is based on HapMap phase II data. The blue dashed line indicates the genome-wide recombination rate average.

(F) The x axis gives the GRCh37 coordinate. Protein-coding genes are presented above the x axis. The arrow indicates the direction of this gene.

Sprime21 (167 kb) (Figure 2). Next, we compared these putative adaptive alleles with four known high-coverage archaic genomes (i.e. Altai Neanderthal, Vindija Neanderthal, Chagyrskaya Neanderthal, and Altai Denisovan). To avoid bias introduced by sequencing coverage, we applied the masks provided with the data, which filter sites with coverage depth below 10, mapping quality below 25 or...
that are within tandem repeats or indels, or that have poor mappability.\textsuperscript{28} Totally, 58 of 76 SNPs passed the filter, and the pass rate of these extremely highly differentiated SNPs with putatively introgressed alleles at this genomic region was similar (Chi-square test, \( p = 1 \)) to that of the whole genome level. Only one (1.7\%) putative adaptive allele could match to Vindija Neanderthal genome (Figure 2). This 185 kb (chr1: 234219614-234404775) extremely highly differentiated region was part of the whole 356 kb (chr1: 234031116-234386959) region exhibiting a signal of archaic introgression and overlapping with the SLC35F3 gene. The upstream 188 kb (chr1: 234031116-234219000) showed a distinct pattern. There were 41 SNPs that passed the strict mask filter, and 85.4\% of the putative adaptive alleles could match to Altai Neanderthal genome, and 95.1\% could match to Vindija Neanderthal genome (Figure 2), suggesting this upstream archaic haplotype may be from a Neanderthal-related population. Further efforts are needed to elucidate whether the putative adaptive introgression haplotype was also from Neanderthal populations.

The putative adaptive haplotype at SLC35F4 was from a Neanderthal-related population

There were 60 highly differentiated SNPs at SLC35F4. For 46 of 60 highly differentiated SNPs, the putative adaptive alleles were absent in the African population (i.e. YRI) and were putatively archaic introgressed alleles in East Asian populations (Figure 3). The proportion of 76.7\% (46/60) of the archaic informative variants accounting for the total highly differentiated SNPs at this genomic region was significantly higher (Chi-square test, \( p < 2.2 \times 10^{-16} \)) than that 1.6\% (1630/100,054) of the whole genome. This suggested the putative adaptive haplotype might be from the archaic hominin group. Both ArchaicSeeker 2 and Sprime suggested that the region showed a signal of archaic introgression spanned at least 200 kb (chr14: 58243621-58449955) (Figure 3). We further compared the putative archaic adaptive alleles in this 200 kb with four high-coverage archaic genomes. We applied a mask for the sequence. There were 48 of all the 79 SNPs carrying putative archaic alleles that passed the strict mask. This was similar (Chi-square test, \( p = 0.92 \)) to that whole genome. 79.2\% could match to Altai Neanderthal genome and 95.8\% could match to Vindija Neanderthal genome (Figure 3), suggesting the putative adaptive haplotype possibly originated from Neanderthal ancestors.

We also constructed a neighbor-joining phylogenetic tree for all of the haplotypes covering the entire 200 kb region based on the estimation of nucleotide distance. We used all the haplotypes from four East Asian populations (i.e. CDX, CHB, CHS, JPT), four European populations (i.e. CEU, GBR, TSI, IBS), and one African population (i.e. YRI) from the 1000 Genomes Project phase 3. Before constructing the phylogenetic tree, we first merged the resulting haplotypes into core haplotypes that differed by <165 nucleotides (i.e. \( 1/1000 \) base pairs in the region were allowed to differ between two haplotypes). This resulted in a set of seven modern human core haplotypes. We constructed a neighbor-joining tree for the seven modern human core haplotypes together with Altai Neanderthal, Chagyrskaya Neanderthal, Vindija Neanderthal, Altai Denisovan, Chimpanzee, and Ancestral sequences for Homo sapiens (Figure S6). One haplotype was present at a high frequency (40.1\%) in East Asia and was low (1.1\%) in European populations and absent in African populations (Figure S6). This haplotype first clustered with the Neanderthal populations.

Moreover, we also test whether the putative adaptive haplotype was the result of incomplete lineage sorting (ILS). We first calculated the regional recombination rate according to the HapMap genetic map.\textsuperscript{29} In this region, the regional recombination rate was 1.4 \( \times 10^{-8} \) per generation per bp. According to a previous study,\textsuperscript{30} the probability of observing a haplotype of 200 kb due to incomplete lineage sorting was about 0. Collectively, all of the above results suggested this putative adaptive haplotype at the SLC35F4 gene was inherited from a Neanderthal-related population.

The first American people did not carry the putative adaptive haplotype at SLC35F4

We used rs17093943 to tag this putative adaptive haplotype at SLC35F4 and investigate the frequency distribution of the putative haplotype in populations from different regions and different periods. We first investigated the global distribution of this putative adaptive haplotype in present-day worldwide populations in the 1000 Genomes Project, HGDP, and Simons Genome Diversity Project (SGDP)\textsuperscript{31} datasets. Intriguingly, the putative adaptive haplotype had a high frequency (~40\%) in present-day East Asians, was absent in present-day Siberians and Americans, and was low in Europeans (Figures 1G and S6). We further analyzed the frequency of this putative adaptive haplotype in ancient samples from America using the 1240K dataset (V44.3). We found no carrier of the putative adaptive allele among 54 haplotypes available from America before 4000 years ago. Actually, of all the 438 haplotypes available in ancient samples from pre-Columbian America, only Saqqaq,\textsuperscript{32} who lived in Greenland 3800 years ago, carried two archaic
alleles. Saqqaq showed a closer relationship with East Asians (Figure S6), suggesting there might be gene flow from East Asian-related ancestry into these samples. The results suggested that the first people to set foot in America did not carry this putative adaptive haplotype. The TMRCA of the putative adaptive haplotype was 31,097 (95% confidence interval [CI]: 28,087-34,169) years in East Asians. One explanation was that

Figure 3. The putative adaptive haplotype at SLC35F4 was from a Neanderthal-related population
See also Figure S6.
(A) Each vertical line indicates one variant. The purple indicates an archaic informative variant. The black dashed line indicates a highly differentiated variant. The black point indicates the variant rs1932242.
(B) Each horizontal line indicates the segments exhibit a signal of archaic introgression in East Asia using different methods. The method was annotated above the segment.
(C) The haplotype for four high-coverage archaic genomes. From top to bottom, the populations are Altai Denisovan, Altai Neanderthal, Chagyrskaya Neanderthal, and Vindija Neanderthal.
(D) Each horizontal line indicates one haplotype and is organized by population. The gray color indicates this is an ancestral state and the brown color indicates this is a derived state.
(E) The recombination rate at this region according to HapMap phase II data. The blue dashed line indicates the genome-wide recombination rate average.
(F) Protein-coding genes are presented above the x axis. The arrow indicates the direction of this gene.
the ancestors of Americans and East Asians might have split more than 30,000 years ago. Indeed, recent studies suggested that the peopling of America began at least 10,000 earlier than was generally suspected, i.e., more than 30,000 years ago.33,34

Historical fluctuations of the putative adaptive haplotype frequency were inferred using whole-genome sequencing data from present-day people

Collectively, the above results suggest these thiamine metabolism-related genes may be subject to positive selection in East Asia. We attempted to determine when the putative positive selection occurred to infer what might have driven it. We first inferred the allele frequency trajectory of these putative adaptive alleles over time using whole-genome sequencing data from present-day people. First, we reconstructed the genealogical history at loci putatively under selection at these genes based on all the haplotypes from the CHB population using RELATE35 and CLUES.36 We began with testing the evidence of positive selection for each locus at these genes using RELATE. Next, we used CLUES to infer the allele frequency trajectory for loci with evidence of positive selection (Figure 4A). CLUES used an importance sampling approach that reweights each sample genealogical history (obtained from RELATE) according to its likelihood under selection and infers a maximum-likelihood selection coefficient and allele frequency trajectory. For each gene, the causal variant of the putative positive selection remains unknown, so we chose the SNP with the strongest signal (i.e., the lowest P) to tag the putative adaptive haplotype.

For SLC19A2, the lowest p-value (0.002) was observed at rs66737090. The frequency of the derive C allele reached 1.3% about 47,600 years ago, increased to 10.1% about 39,025 years ago and to 50.3% about 14,325 years ago, and reached up to 60% 10,000 years ago. The allele frequency remained stable high (~70%) in East Asian populations during the past 10,000 years (Figure 4A). For SLC35F3, the SNP rs17093835 has the lowest p-value (0.001) within this genomic region. The derived C allele reached a frequency of 1.3% 50,000 years ago, increased to 10.1% about 44,050 years ago and to 50.3% 23,250 years ago, and reached up to 72% 10,000 years ago. The allele frequency remained stably high (~80%) in East Asian populations during the past 10,000 years (Figure 4A). For SLC35F4, the lowest p-value (0.02) was observed at rs17093835. The frequency of the derived A allele reached 1.3% 28,800 years ago, increased to 10.1% 20,835 years ago and reached up to 30% 10,000 years ago. The allele frequency remained stable (~41%) in East Asian populations in the past 10,000 years (Figure 4A). For each gene, the estimated allele frequency trajectories for highly differentiated SNPs varied. However, all the results of allele frequency trajectories suggested that the putative positive selection started early, at least 10,000 years ago.

We also used startmrca to estimate the selection-onset time of a beneficial allele using all the haplotypes from the CHB population.37 The startmrca leverages the length distribution of the shared ancestral haplotype, the accumulation of derived mutations on the ancestral background, and the surrounding background haplotype diversity. The estimate was 70,680 (95% CI: 61,145-80,261) for rs66737090 at SLC19A2. The estimate was 55,544 (95% CI: 49,370-62,787) for rs61824595 at SLC35F3. The estimate was 49,309 (95% CI: 41,849-55,510) for rs17093835 at SLC35F4.

Historical fluctuations of the putative adaptive haplotype frequency were estimated using ancient genome data

We further investigated the allele frequency change over time using current available ancient DNA data from a compiled database “1240K” dataset (“Allen Ancient DNA Resource,” V44.3). We grouped the samples into different time windows. Because the currently available ancient DNA data are heavily biased toward western Eurasia and are still sparse, especially for older periods (Table S1 and Figure S7), the sample size for some particular regions and periods was still small, resulting in the frequency showing a high fluctuation. However, the number in the most recent 10,000 years in East Asia might be large enough to estimate the allele frequency in the past. Considering agriculture in East Asia emerged in the most recent 10,000 years, this still provides an opportunity to study whether this putative positive selection is associated with agriculture in human evolutionary history.

For SLC19A2, 23 of 205 extremely highly differentiated SNPs overlapped with the 1240K dataset. These 23 SNPs showed a similar pattern in allele frequency change over time in East Asia (Figure S7). We used rs36040633, which had the highest LSLB of 0.787, to tag the putative adaptive haplotype at the region encompassing SLC19A2 on chromosome 1 (Figure 4B). For two ancient samples before 30,000 years ago, the putative adaptive allele was present in AR33K (~33,000) in northern China.38 And the genotype for Tianyuan (~40,000)
in northern China was missing. Between 30,000 and 10,000 years ago, the putative adaptive haplotype had a frequency of 75%. The putative adaptive haplotype was present at a frequency of or slightly above 50% in the recent 10,000 years. This result suggested that the putative adaptive haplotype had reached a high frequency at least 10,000 years ago and has changed little in the recent 10,000 years.

For SLC35F3, 24 of 118 extremely highly differentiated SNPs overlapped with the 1240K dataset. These SNPs showed a similar pattern in allele frequency change over time in East Asia (Figure S7). We used an
SNP rs16842807, which had the highest LSBL value of 0.82, to tag the putative adaptive haplotype at SLC35F3 on chromosome 1 (Figure 4C). The putative adaptive haplotype seems to have occurred at a frequency of about 50% in East Asians before 30,000 years ago. Between 30,000 and 10,000 years ago, the putative adaptive haplotype was at a frequency in the order of 87.5%. Subsequently, the putative adaptive haplotype seems to have been present at a frequency of or slightly below 80% in the recent 10,000 years. The results showed that the putative adaptive haplotype had reached a high frequency at least 10,000 years ago and has changed little in the recent 10,000 years in East Asian populations (Figures 4C and S7), suggesting that this putative positive selection probably occurred in ancient times, at least 10,000 years ago.

For SLC35F4, 1 out of 60 highly differentiated SNPs overlapped with the 1240K dataset. We used this SNP, rs17093943, to tag this putative adaptive haplotype at SLC35F4 on chromosome 14 (Figure 4D). For this SNP, four haplotypes were successfully genotyped for samples in East Asia 30,000 years ago, and only two haplotypes were successfully genotyped for samples in East Asia between 30,000 and 10,000 years ago. We found no carrier of the putative adaptive allele among the six haplotypes available. It seems that the putative adaptive haplotype has been present at a frequency of 20% in East Asian populations and has changed little in the recent 10,000 years. In addition, the putative positive selection seemingly has occurred only in East Asian populations and not in Siberian populations. Most of the present-day samples from Siberia did not carry the putative adaptive allele (Figure S6). Accordingly, the frequency of this putative adaptive haplotype increased to about 30% when the samples from Siberia were excluded (Figure S7).

Functional consequences of these putative adaptive alleles at genes involved in thiamine metabolism

As natural selection may act directly on functionally important variants driving phenotypic variation, variants with important functions are more likely to be causal of the positive selection.40 We investigated the functional consequences of these putative adaptive alleles. This extremely highly differentiated 489 kb region contains six protein-coding genes, namely, ATP1B1, NME7, BLZF1, SLC19A2, F5, and SELP. These 205 extremely highly differentiated SNPs contain two missense variants and one synonymous variant. One SNP, rs1028180, modifies the amino acid of the BLZF1 gene, and the other SNP, rs13306334, modifies the amino acid of the F5 gene. None of these SNPs modify the amino acid sequence of the SLC19A2 gene. At the SLC35F3 region, none of these 118 extremely highly differentiated SNPs modify the amino acid sequence of any protein-coding gene. And at the SLC35F4 region, none of these 60 highly differentiated SNPs modifies the amino acid of this gene. However, the highly differentiated region at SLC19A2, SLC35F3, and SLC35F4 contains regulatory elements, including well-characterized transcription factor binding sites, CTCF binding sites, enhancers, and promoters, suggesting a putative regulatory impact of the putative adaptive alleles.

We next examined whether the putative adaptive alleles could affect the expression of these genes using Genotype-Tissue Expression (GTEx) Project (v8 release) Portal.41 Though these extremely highly differentiated SNPs are not gene expression quantitative trait loci (eQTLs) for the SLC19A2 and SLC35F3. Some SNPs that are in linkage disequilibrium (LD) with those extremely highly differentiated SNPs could alter the expression of SLC19A2 and SLC35F3. For example, the SNP rs10800448 was in LD (\(r^2 = 0.73\), \(D' = 1\)) with one of these extremely highly differentiated SNPs (i.e. rs11585538) at the region encompassing SLC19A2 in the CHB population (Figure S8). The frequency of the derived allele A of rs10800448 was 78.6% in CHB, 17.7% in CEU, and 36.6% in the YRI population. The C allele could increase the expression of the SLC19A2 gene in the CHB population (p = 2.4 \times 10^{-5}, normalized effect size [NES] = 0.17). One SNP rs521127 was in LD (\(r^2 = 0.97\)) with one of these extremely highly differentiated SNPs (i.e. rs11585538) at the SLC35F3 region in the CHB (Figure S8). The derived G allele of rs521127 allele was present at the frequency of 83% in CHB, 36.4% in CEU, and 95.8% in the YRI population. The G allele could significantly reduce the expression of SLC35F3 in the pituitary (p = 5 \times 10^{-6}, NES = −0.23). The SNP rs1932242 exhibited the highest LSBL of 0.41 in this SLC35F4 region. The derived C allele was 68.0% in CHB, 11.6% in CEU, and 8.3% in the YRI population. The C allele could increase the expression of the SLC35F4 gene in the thyroid (p = 8.4 \times 10^{-17}, NES = 0.59). This SNP was also in linkage disequilibrium (LD) (\(r^2 = 0.36\), \(D' = 0.97\) in CHB) with other highly differentiated SNPs with putative archaic alleles (i.e. rs17093943) (Figure S8). This might suggest that the putative adaptive haplotype could influence the expression of the SLC35F4 gene.

On the other hand, some of these extremely highly differentiated SNPs could affect the expression of other genes located in the regions. For example, one SNP rs1138486 exhibited extremely high differentiation...
with an LSBL of 0.69. The ancestral allele T was 78.2% in CHB, 4.0% in CEU, and 15.7% in YRI. The T allele could reduce the expression of ATP1B1 in Muscle-Skeletal (p = 1.2 × 10^{-4} and NES = −0.28) and increase the expression of CCDC181 in Esophagus Muscularis (p = 4.2 × 10^{-5}, NES = 0.47). This SNP was located in the enhancer region and was conversed with the Combined Annotation Dependent Depletion (CADD) score of 14.14. It was also reported to be associated with the QT interval and the T allele was the risk allele with p = 7 × 10^{-14} in Hispanics/Latino populations.42 These results suggested the putative adaptive alleles might have some functional consequences.

We also investigated the potential phenotypes that may be influenced by the putative adaptive haplotype using GWAS Catalog data. One SNP rs12049593 in the intron of SLC35F3 is reported to carry an allele from the Neanderthal population and to be associated with protein-calorie malnutrition.13 This SNP rs12049593 was located 130 kb upstream of the extremely highly differentiated region (Figures 1D and 3). The frequency of the Neanderthal allele was about 10% in East Asian populations and 5% in European populations, suggesting that this SNP had a different pattern from SNPs that may be under positive selection in East Asians. It remains unknown whether this putative adaptive haplotype was associated with malnutrition. One extremely highly differentiated SNP rs7521660 was reported to be associated with macroscopic brain infarct with p < 3 × 10^{-6} in European populations.61 The protective allele A was 89.8% in CHB, 8.6% in CEU, and 0.9% in the YRI population. In addition, one SNP rs4333882 was in complete LD (r^2 = 1 in the CHB population) with one of these extremely highly differentiated SNPs (i.e. rs11585538). This SNP rs4333882 was reported to be associated with diverticular disease (p < 4 × 10^{-22} and beta = 0.007) in the UK biobank data.44 The risk allele G of rs4333882 reached a high frequency of 82.5% in the CHB population and was 19.2% in CEU and 48.1% in the YRI population. This variant was also associated with abdominal infections with p < 2.14 × 10^{-14} in the UK biobank data.45 Previous studies also showed that SLC35F3 is associated with hypertension in both European and Han Chinese populations.12,46 Some extremely highly differentiated SNPs in this region encompassing SLC19A2 are reported to be associated with QT interval and electrocardiographic traits,42 blood protein levels,47 and immune cell traits.48 These results suggested that the putative adaptive alleles affect several phenotypes.

**DISCUSSION**

This study advances our understanding of the evolutionary history of positive selection of these genes involved in thiamine metabolism using the combination of whole-genome sequencing data from present-day people and large available ancient DNA data. For these seven protein genes involved in thiamine metabolism, we refined the evidence of positive selection and identified the putative adaptive haplotypes at SLC19A2, SLC35F3, and SLC35F4 in East Asian populations. We also found the putative adaptive haplotype at SLC35F4 was from the Neanderthal-related population, and the putative adaptive haplotypes at SLC35F3 also showed a signal of archaic introgression and the putative adaptive alleles could not match four high-coverage archaic genomes.

We further explored what might have driven this putative positive selection. Thiamine is the precursor of the essential coenzyme thiamine pyrophosphate (TPP) required for glucose metabolism.51 It helps the body convert carbohydrates into energy. The more carbohydrates are present in the diet, the more thiamine is required. One dramatic transition in the human diet occurred after the emergence of agriculture. The modern human changed from a diet based on hunting and gathering to a diet heavily based on the products of agriculture.50 Thus, the emergence of agriculture and the diet change in agricultural societies may have driven the putative positive selection. In this study, we found that the putative adaptive haplotypes had reached a high frequency similar to that of present-day populations in East Asia at least 10,000 years ago and the frequencies might have some functional consequences.

Other environmental factors, including pathogen load, may have driven the putative positive selection. Pathogens and the diseases they cause have been among the strongest selection pressures in human evolutionary history.52 Thiamine can improve immune system function. Importantly, thiamine could inhibit the production of these cytokines, increase anti-inflammatory activity,53,54 and lower the pro-inflammatory response. Upon viral infection, tissue inflammation is driven by multiple pro-inflammatory and immune-regulatory signals.55,56 Thiamine could help treat viral infections. For example, a two-center propensity score-matched study showed that the use of thiamine as adjunctive therapy may have potential survival
benefits in critical patients with COVID-19. All of these results suggest that thiamine plays an important role in the response to pathogens. When the ancestors of modern humans entered Eurasia and proceeded further into East Asia, they encountered new environmental challenges, including pathogens and different climates. The archaic hominins had lived in Eurasia for several hundred thousand years and might be well adapted to the environment. Interbreeding with archaic hominins contributed to modern humans obtaining alleles that were adaptive to the environment. Indeed, archaic introgression has been shown to affect genes involved in immunity. In addition, the Neanderthal haplotypes are protective against several viruses (e.g., SARS-CoV and SARS-CoV-2). The response to pathogens may be the selection pressure for these putative adaptive haplotypes at these genes.

Limitations of the study
Several factors may limit the power of this study. It is challenging to obtain an understanding of natural selection without a description of genetic mechanisms that linked a known variant to a plausible phenotype under the pressure of the selective force. In this study, we explicitly demonstrated evidence of positive selection and infer the evolutionary history of the putative adaptive haplotypes at these thiamine genes in East Asia. However, the causal variants and the adaptive traits remain uncertain. Specifically, SLC19A2 is a well-known thiamine transporter gene. The region encompassing SLC19A2 contains six protein-coding genes, namely, ATP1B1, NME7, BLZF1, SLC19A2, F5, and SELP. Whether SLC19A2 is the causal gene remains unknown. SLC35F3 encodes a potential thiamine transporter. Little is known about the gene SLC35F4, except that it is a paralog of SLC35F3. In addition, it remains unclear how these putative adaptive haplotypes affect the expression of the corresponding genes. We examined the effect of the putative adaptive alleles on gene expression using the GTEx dataset. Most of the samples were descendants of European ancestry. These putative adaptive alleles were under positive selection in East Asia. Whether the effect of these alleles on gene expression in European and East Asian populations was the same remains unknown. Further, it is also unknown how these putative adaptive haplotypes affect the ability to absorb thiamine. Although a large amount of evidence suggested that thiamine-related traits are putatively adaptive phenotypes, other traits except thiamine metabolism may also be the target of positive selection. Ancient DNA provided unprecedented insight into modern human history, but the small sample size of available ancient DNA data before 10,000 years ago makes it difficult to estimate the allele frequency precisely. It also prevented this study from directly estimating the time for natural selection using ancient DNA data.

STAR METHODS
Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHODS DETAILS
  - Genes involved in thiamine metabolism
  - Data
  - Detecting natural selection
  - Identifying archaic introgression
  - Neighbor-joining phylogenetic tree
  - Estimate the time to the most recent common ancestor
  - Estimation of the selection-onset time
  - Inference of the allele frequency change over time
  - The probability of archaic introgressed segment due to incomplete lineage sorting
  - Expression analysis in multiple tissues from the GTEx dataset
  - Functional annotation
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION
Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.105614.
ACKNOWLEDGMENTS

This study was supported by the Basic Science Center Program (32288101), the National Natural Science Foundation of China (NSFC) grants (32030020, 31961130380), the Strategic Priority Research Program (XDB17; XDB38000000) of the Chinese Academy of Sciences (CAS), the UK Royal Society-Newton Advanced Fellowship (NAF/R1/191094), the Shanghai Municipal Science and Technology Major Project (2017SHZDZX01). X.M. acknowledges the support of the Shanghai Post-doctoral Excellence Program (2021432). The funders had no role in the study design, data collection, analysis, decision to publish, or preparation of the article.

AUTHOR CONTRIBUTIONS

Conceptualization, S.X. and X.M.; formal analysis, X.M.; writing – original draft, X.M.; writing – review & editing, S.X.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: March 21, 2022
Revised: August 18, 2022
Accepted: November 14, 2022
Published: December 22, 2022

REFERENCES

1. Marki, P.E., Kory, P., Varon, J., Iglesias, J., and Meduri, G.U. (2021). MATH+ protocol for the treatment of SARS-CoV-2 infection: the scientific rationale. Expert Rev. Anti Infect. Ther. 19, 129–135. https://doi.org/10.1080/14787924.2020.1808462.

2. Mikkelsen, K. and Apostolopoulos, V. (2019). Vitamin B1, B2, B3, B5, and B6 and the immune system. In Nutrition and Immunity, M. Mahmoudi and N. Rezaei, eds. (Springer International Publishing), pp. 115–125. https://doi.org/10.1007/978-3-319-80339-9_7.

3. Al Sulaiman, K., Aljuhani, O., Al Dossari, M., Alshahrani, A., Alharbi, A., Algarni, R., Al Jeraisy, M., Al Harbi, S., Al Katheri, A., Al Erani, F., et al. (2021). Evaluation of thiamine as adjunctive therapy in COVID-19 critically ill patients: a two-center propensity score matched study. Crit. Care 25, 223. https://doi.org/10.1186/s13054-021-03648-9.

4. Whitfield, K.C., Bourassa, M.W., Adamolekun, B., Bergeron, G., Bettendorff, L., Brown, K.H., Cox, L., Fatali-Valevski, A., Fischer, P.R., Frank, E.L., et al. (2018). Thiamine deficiency disorders: diagnosis, prevalence, and a roadmap for global control programs. Ann. N. Y. Acad. Sci. 1430, 3–43. https://doi.org/10.1111/nyas.13919.

5. Tallaksen, C.M., Sande, A., Behmner, T., Bell, H., and Karlsen, J. (1993). Kinetics of thiamin and thiamin phosphate esters in human blood, plasma and urine after 50 mg intravenously or orally. Eur. J. Clin. Pharmacol. 44, 73–78. https://doi.org/10.1007/BF00315284.

6. Bettendorff, L., Wins, P., and Lesourd, M. (1994). Subcellular localization and compartmentation of thiamine derivatives in rat brain. Biochim. Biophys. Acta 1222, 1–6. https://doi.org/10.1016/0167-4889(94)90018-3.

7. Smith, T.J., and Hess, S.Y. (2021). Infantile thiamine deficiency in South and Southeast Asia: an age-old problem needing new solutions. Nutr. Bull. 46, 12–25. https://doi.org/10.1111/nbu.12581.

8. Lonsdale, D., and Mars, C. (2017). Chapter 6 - high-calorie malnutrition and its impact on health. In Thiamine Deficiency Disease, Dysautonomia, and High Calorie Malnutrition, D. Lonsdale and C. Mars, eds. (Academic Press), pp. 213–261. https://doi.org/10.1016/B978-0-12-810387-6.00006-X.

9. Kerns, J.C., Arundel, C., and Chawla, L.S. (2015). Thiamin deficiency in people with obesity. Adv. Nutr. 6, 147–153. https://doi.org/10.3945/an.114.007526.

10. Thormar, P.J., Babaei-Jadidi, R., Al Ali, H., Rabbani, N., Antonsunil, A., Larkin, J., Ahmed, A., Rayman, G., and Bodmer, C.W. (2007). High prevalence of low plasma thiamine concentration in diabetes linked to a marker of vascular disease. Diabetologia 50, 2164–2170. https://doi.org/10.1007/s00125-007-0771-4.

11. Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Coftasapás, C., Xie, X., Byrne, E.H., McCarroll, S.A., Gaudet, R., et al. (2007). Genome-wide detection and characterization of positive selection in human populations. Nature 449, 913–918. https://doi.org/10.1038/nature06200.

12. Zhang, K., Huentelman, M.J., Rao, F., Sun, E.I., Cornevaux, J.J., Schork, A.J., Wei, Z., Wisalen, J., Miramontes-Gonzalez, J.P., Hightower, C.M., et al. (2014). Genetic implication of a novel thiamine transporter in human hypertension. J. Am. Coll. Cardiol. 63, 1542–1555. https://doi.org/10.1016/j.jacc.2014.01.007.

13. Cheng, X., Xu, C., and DeGiorgio, M. (2017). Fast and robust detection of ancestral selective sweeps. Mol. Ecol. 26, 6871–6891. https://doi.org/10.1111/mec.14416.

14. Racimo, F. (2016). Testing for ancient selection using cross-population allele frequency differentiation. Genetics 202, 733–750. https://doi.org/10.1534/genetics.115.178995.

15. Simonti, C.N., Vernot, B., Bastarache, L., Bottinger, E., Carrell, D.S., Chisholm, R.L., Crosslin, D.R., Hebbring, S.J., Janik, G.P., Kullo, I.J., et al. (2016). The phenotypic legacy of admixture between modern humans and Neandertals. Science 351, 737–741. https://doi.org/10.1126/science.aad2149.

16. Gittelman, R.M., Schraiber, J.G., Vernot, B., Mikacenic, C., Wurfel, M.M., and Akery, J.M. (2016). Archaic hominin admixture facilitated adaptation to out-of-Africa environments. Curr. Biol. 26, 3375–3382. https://doi.org/10.1016/j.cub.2016.10.041.

17. Voight, B.F., Kudaravalli, S., Wen, X., and Pritchard, J.K. (2006). A map of recent positive selection in the human genome. PLoS Biol. 4, e72. https://doi.org/10.1371/journal.pbio.0040072.

18. Shriver, M.D., Kennedy, G.C., Parra, E.J., Lawson, H.A., Sonpar, V., Huang, J., Akery, J.M., and Jones, K.W. (2004). The genomic distribution of population substructure in four populations using 8, 525 autosomal SNPs. Hum. Genomics 1, 274–286.

19. 1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Macht, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. Nature 526, 68–74. https://doi.org/10.1038/nature15393.

20. Bergström, A., McCarthy, S.A., Hui, R., Almarri, M.A., Ayub, Q., Danecek, P., Chen, Y., Feikel, S., Hallast, P., Kamph, J., et al. (2020). Insights into human genetic variation and population history from 929 diverse genomes. Science 367.
54. Menezes, R.R., Godin, A.M., Rodrigues, F.F., Bertolla, C.M., Paulino, T.P., Rachid, M.A., Machado, R.R., and Coelho, M.M. (2017). Thiamine and riboflavin inhibit production of cytokines and increase the anti-inflammatory activity of a corticosteroid in a chronic model of infection induced by complete Freund’s adjuvant. Pharmacol. Rep. 69, 1036–1043. https://doi.org/10.1016/j.pharep.2017.04.011.

55. Vatsalya, V., Li, F., Frimodig, J., Gala, K.S., Srivastava, S., Kong, M., Ramchandani, V.A., Feng, W., Zhang, X., and McClain, C.J. (2020). Repurposing treatment of Wernicke-Korsakoff syndrome for Th-17 cell immune storm and neurological symptoms in COVID-19: thiamine efficacy and safety, in vitro evidence and pharmacokinetic profile. Front. Pharmacol. 11, 598128.

56. Glass, W.G., Rosenberg, H.F., and Murphy, P.M. (2003). Chemokine regulation of inflammation during acute viral infection. Curr. Opin. Allergy Clin. Immunol. 3, 467–473.

57. Dannemann, M., and Racimo, F. (2018). Something old, something borrowed: admixture and adaptation in human evolution. Curr. Opin. Genet. Dev. 53, 1–8. https://doi.org/10.1016/j.gde.2018.05.009.

58. Abi-Rached, L., Jobin, M.J., Kulkarni, S., Quach, H., Rotival, M., Pothlichet, J., Loh, P.M., et al. (2011). The shaping of human populations. Cell 145, 1272–1283. https://doi.org/10.1016/j.cell.2011.06.016.

59. Deschamps, M., Laval, G., Fagny, M., Itan, Y., Abel, L., Casanova, J.-L., Patin, E., and Quintana-Murci, L. (2016). Genomic signatures of selective pressures and introgression from archaic hominins at human innate immunity genes. Am. J. Hum. Genet. 98, 5–21. https://doi.org/10.1016/j.ajhg.2015.11.014.

60. Lu, D., Lou, H., Yuan, K., Wang, X., Wang, Y., Zhang, C., Lu, Y., Yang, X., Deng, L., Zhou, Y., et al. (2016). Ancestral origins and genetic history of Tibetan highlanders. Am. J. Hum. Genet. 99, 580–594. https://doi.org/10.1016/j.ajhg.2016.07.002.

61. MacArthur, J., Bowling, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., et al. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 45, D896–D901. https://doi.org/10.1093/nar/gkw1133.

62. Flicek, P., Wilder, S.P., Johnson, N., Juettemann, T., and Flicek, P. (2015). The archaic hominins at human innate immunity genes. Am. J. Hum. Genet. 98, 5–21. https://doi.org/10.1016/j.ajhg.2015.11.014.

63. Sidiropoulos, K., Viteri, G., De la Higuera, C., Webber, M., Orlic-Milacic, M., Jassal, B., Shamovsky, V., Duenas, C., et al. (2017). PathCards: multi-source consolidation of human biological pathways. Database 2015, bav006. https://doi.org/10.1093/database/bav006.

64. Lane, J., Keller, L., and Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28, 28–31. https://doi.org/10.1093/nar/28.1.28.

65. Saitou, N., and Nei, M. (1987). The neighbor-joining method - a new method for determining the group of evolutionary relationships. J. Mol. Biol. 165, 107–119. https://doi.org/10.1016/S0022-2836(87)80157-2.

66. Zhang, R., Yuan, K., and Xu, S. (2022). Detecting archaic introgression and modeling multiple-wave admixture with ArcaicSeeker 2.0. STAR Protoc. 3, 101314. https://doi.org/10.1016/j.xpro.2022.101314.

67. Kircher, M., Witten, D.M., Jain, P., O’Roak, B.J., Cooper, G., Fitzgerald, S., et al. (2014). EndSmiler 2014. Nucleic Acids Res. 42, D749–D755. https://doi.org/10.1093/nar/gkt1196.

68. Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. 46, 310–315. https://doi.org/10.1038/ng.2892.
**STAR METHODS**

**KEY RESOURCES TABLE**

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---------------------|--------|------------|
| Deposited data      |        |            |
| 90 Han Chinese      | Lu et al. [63, 64] | https://ngdc.cnbc.ac.cn/bioproject/browse/PRJCA000246 |
| 1000 Genomes Project – Phase 3 | 1000 Genomes Project Consortium et al. [19] | ftp://ftp.1000genomes.ebi.ac.uk/ vol1/ftp/release/20130502/ |
| Human Genome Diversity Project (HGDP) | Bergström et al. [20] | ftp://ngs.sanger.ac.uk/production/hgdp |
| Simons Genome Diversity Project data (SGDP) | Mallick et al. [21] | https://sharehost.hms.harvard.edu/genetics/reich_lab/sgdp/phased_data/PS2_multisample_public/ |
| Sprime scores for 1000 Genomes populations | Browning et al. [21] | https://doi.org/10.17632/y7hry83xvr.1 |
| Ancient DNA         | Reich et al. | https://reich.hms.harvard.edu/Allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data |
| Altai Neanderthal   | Prufer et al. [24] | http://cdna.eva.mpg.de/Neandertal/altai/AltaiNeandertal/ |
| Altai Denisovan     | Meyer et al. [27] | http://cdna.eva.mpg.de/Neandertal/AltaiDenisovan/ |
| Vindija 33.19       | Prufer et al. [25] | http://cdna.eva.mpg.de/Neandertal/Vindija/VCF/Vindija33.19/ |
| Chagyrskaya Neanderthal | Mafessoni et al. [26] | http://cdna.eva.mpg.de/Neandertal/Chagyrskaya/ |
| Ancestral sequences for Homo sapiens | Paten et al. [65] | http://ftp.ensembl.org/pub/release-71/fasta/ancestral_alleles/ |
| HapMap genetic map  | International HapMap Consortium et al. [29] | https://ftp.ncbi.nlm.nih.gov/hapmap/ |
| Expression data     | The GTEx Consortium et al. [41] | https://www.gtexportal.org/home/ |
| GWAS                | MacArthur et al. [66] | https://www.ebi.ac.uk/gwas/ |
| The Ensembl regulatory build | Zerbino et al. [67] | https://ftp.ensembl.org/pub/annotation/homo_sapiens/ |

**Software and algorithms**

| Software and algorithms | Source | Identifier |
|-------------------------|--------|------------|
| Shapeit2                | Delaneau et al. [68] | https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html |
| ArchaicSeeker 2.0       | Yuan et al. [23] | https://github.com/Shuhua-Group/ArchaicSeeker2.0 |
| FASTME                  | Lefort et al. [69] | http://www.atgc-montpellier.fr/fastme/ |
| Startmrca               | Smith et al. [37] | https://github.com/jhavsmith/startmrca |
| RELATE                  | Speidel et al. [35] | https://myersgroup.github.io/relate/ |
| CLUES                   | Stern et al. [36] | https://github.com/35ajstern/clues |

**RESOURCE AVAILABILITY**

**Lead contact**

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Shuhua Xu (xushua@fudan.edu.cn).
Materials availability
This study did not generate new unique reagents.

Data and code availability
This paper analyzes existing, publicly available data. These accession numbers for the datasets are listed in the key resources table.

This paper does not report any original code.

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

EXPERIMENTAL MODEL AND SUBJECT DETAILS
All the source of bioinformatics data used in the analyses is provided in the key resources table.

METHODS DETAILS
Genes involved in thiamine metabolism
We queried public databases including REACTOM,70 GeneCards,71 and KEGG72 to obtain the list of genes involved in thiamine metabolism. We also performed a comprehensive search on recent publications to get genes potentially involved in this pathway.17 Briefly, this pathway contained seven protein-coding genes (Figure S1): SLC19A2, SLC19A3, TPK1, THTPA, SLC25A19, SLC35F3, and SLC35F4. In particular, SLC19A2, SLC19A3, SLC35F3, and SLC35F4 are involved in the transport of thiamine from the extracellular space to the cytosol. In this study, we investigated the footprints of positive natural selection on these seven protein-coding genes in East Asian populations.

Data
We used 90 high coverage (30x) whole-genome sequencing Han Chinese samples to perform a natural selection analysis to detect genes with signals of positive selection in East Asia.63,64 These 90 Han Chinese samples were referred to as HAN in this study. The data of Han genomes were phased using shapeit268 with the 1000 Genomes Project (1KG) phase 3 data as a reference panel with default parameters. The genetic map from HapMap II was used to calculate the genetic distance between pairwise SNPs.29

The 1000 Genomes Project phase 3 dataset19 was also included in the analysis to represent the worldwide reference populations including 2504 individuals from 26 worldwide populations and was used in most of the other analyses, including constructing the haplotype network, constructing the phylogenetic tree, and estimating the time to the most recent common ancestor (TMRCA).

To investigate the frequency distribution of these putative adaptive alleles in more diverse populations, we also obtained 929 high coverage (30x) whole-genome sequencing data from 54 populations from the Human Genome Diversity Project (HGDP).20 The GRCh38 coordinate was converted to GRCh37 using liftover from the UCSC genome browser.73 Only single nucleotide variants (SNVs) located on the same chromosome in different coordinates were used for further analyses. In addition, the high-quality genomes from 300 individuals from 142 diverse populations from the Simons Genome Diversity Project (SGDP) dataset31 were also included in the analysis.

To investigate how the frequencies of the putative adaptive alleles changed over time, we employed available ancient DNA data from a compiled database ‘1240K’ dataset (“Allen Ancient DNA Resource,” https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data, version 44.3). We clustered these samples into different groups by time and geography (Table S1).

We also obtained the high-coverage reference genomes for Altai Neanderthal,24 Vindija 33.19 Neanderthal,25 Chagyrskaya Neanderthal,26 and Altai Denisovan.27 When comparing putative adaptive alleles to an archaic genome, we applied the masks provided with data (http://cdna.eva.mpg.de/neandertal/Vindija/FilterBed), which filters for coverage (stratified by GC content) and filter sites with coverage depth below 10, or mapping quality below 25 or that have poor mappability or that are within in tandem repeats.
or indels. For sites that pass those filters, we calculate a match if the archaic genotype includes the putative archaic-specific allele and a mismatch otherwise. The match rate is calculated as the number of matches divided by the number of compared sites (matches and mismatches). Sites that do not pass the filters do not contribute to the match rate.

**Detecting natural selection**

We mainly used two different strategies to investigate the signal of positive selection in East Asian populations. One was based on allele frequency differentiation and the other was based on extended haplotype homozygosity.

We first merged HAN data with 1000 Genomes Project data. We removed SNPs with minor allele frequency < 0.05 in all the populations. This resulted in 10,409,928 SNPs for the natural selection analysis.

We then calculated the locus-specific branch length (LSBL)\(^{18}\) for the HAN population with the European population (i.e. CEU) and the African population (i.e. YRI) as reference populations. We used LSBL instead of population branch statistic (PBS)\(^{74}\) as the value of LSBL shows the genetic differentiation more intuitively. For each locus, the \(F_{ST}\) value was estimated between pairs of these three populations (HAN, CEU, and YRI) using Weir and Cockham \(F_{ST}\).\(^{75}\) Then the LSBL was calculated as follows: \(LSBL = (F_{ST \text{HAN,CEU}} + F_{ST \text{HAN,YRI}} - F_{ST \text{CEU,YRI}})/2\). The 1% threshold of the whole genome was 0.36 and the 0.1% threshold was 0.55. SNPs within the top 1% were considered highly differentiated. SNPs within the top 0.1% were considered extremely highly differentiated.

We further implemented the integrated Haplotype Score (iHS)\(^{17}\) and the Cross Population Extended Haplotype Homozygosity (XP-EHH) test\(^{11}\) using R package rehh 2.0.\(^{76}\) The inferred ancestral state for *Homo sapiens* of each SNP was retrieved from Ensembl database release 71.\(^{65,77}\) A genetic map from HapMap II was used to calculate the distance between markers.\(^{29}\) This map was average across all three HapMap populations, and it was unlikely to be biased by selection in any individual population. We calculated the iHS value for each locus for the HAN population. SNPs with |iHS| > 2 were considered to exhibit a signal of positive selection.\(^{17}\) The XP-EHH score for each locus for the HAN population was calculated with CEU as the reference population. SNPs with XP-EHH > 2 were thought to exhibit a signal of positive selection.\(^{78}\)

We partitioned the whole genome into fixed non-overlapping 100-kb windows. Windows with less than 10 SNPs were removed. Inspired by previous study,\(^{78}\) we employed different strategies to calculate the score of each window for different tests. For LSBL, the mean of the top 10 LSBL within each window was used as the score of that window. For iHS, the fraction of SNPs with |iHS| > 2 within each window was used as the score. For XP-EHH, the maximum XP-EHH within each window was used as the score of that window. For each test, the windows were ranked from the highest to lowest, respectively. The empirical p value was used to test the significance.

**Identifying archaic introgression**

*Identifying archaic introgressed regions and variants using Sprime*

To define genomic regions with the signal of archaic introgression, we got segments identified by Browning et al. using Sprime, which uses a heuristic scoring strategy that compares high-LD regions in a target population (e.g. East Asians) with an unadmixed outgroup (i.e. Africans) to identify putatively introgressed regions.\(^{21}\) They used each non-African population as the target population and YRI from the 1000 Genomes Project as an outgroup. In this study, we only considered the segments identified independently in five East Asian subpopulations (i.e. CHB, CHS, JPT, KHV, and CDX). For each East Asian subpopulation, we removed segments with less than 30 putatively introgressed alleles as the authors suggested. We then merged these segments with a strategy to get the finest coordination. For example, one segment (a1-b1) overlaps with another segment (a2-b2), and the relationship of these four positions was a1<a2<b1<b2. The new merged segments would be three segments (a1-a2, a2-b1, b1-b2). This resulted in 5,397 segments covering 720,406,243 bp. We kept segments that showed signals of archaic introgression in at least four subpopulations. This finally resulted in 1,610 segments that covered 304,398,043 bp.

Because of the long divergence between modern humans and archaic hominin groups, archaic hominin groups carry many alleles that are specific to their lineage. Such alleles are presented on introgressed
haplotypes but are absent or very rare in African populations. Variants carrying such putatively introgressed alleles were considered archaic informative variants.\textsuperscript{21,22} To define archaic informative variants, we merged the archaic informative variants independently identified in five East Asian subpopulations. This resulted in 370,123 archaic informative variants. We only keep variants that were identified in at least four subpopulations and that are located in the genomic regions identified in the above step. This resulted in 138,323. And 81,085 (58.6%) passed the mask filter. When comparing the putatively introgressed alleles with archaic genomes, variants that pass the mask filter were used.

**Identifying archaic introgressed regions using ArchaicSeeker 2.0**

We also used ArchaicSeeker 2.0\textsuperscript{23} to detect the archaic segments in East Asian populations from the 1000 Genomes Project phase 3 data. We used five East Asian populations as target populations. We used Altai Neanderthal and Altai Denisovan as archaic references and YRI from the 1000 Genomes Project as an African reference. The default was used for all parameters as suggested in protocol paper.\textsuperscript{79} ArchaicSeeker 2.0 used a hidden Markov Model to detect archaic segments on each haplotype. We merged these segments and keep segments detected in all the East Asian subpopulations. This resulted in 134,961 segments that covered 651,657,769 bp.

For introgressed haplotypes identified by Sprime in five East Asian subpopulations (1610), 1319 (82%) are more than 50% covered by segments identified by ArchaicSeeker 2.0 and 94 (6%) have 0% coverage. Conversely, for introgressed segments identified by ArchaicSeeker 2.0 in East Asians (134,961), 60570 (44.9%) are more than 50% coverage by segments identified by Sprime, and 73829 (55%) have 0% coverage.

**Neighbor-joining phylogenetic tree**

We constructed a neighbor-joining phylogenetic tree\textsuperscript{80} for the haplotypes using all of the SNPs in the entire region based on nucleotide distance. We also obtained the sequence from the Chimpanzee genome.\textsuperscript{81} The inferred ancestral sequences for Homo sapiens were obtained from Ensembl release 71\textsuperscript{65,77}. Before constructing the haplotype network, we first merged the resulting haplotypes into core haplotypes that differed by < 10% of the total number of nucleotides (i.e., ~1/1000 base pairs in the region were allowed to differ between two haplotypes). This results in a set of modern human core haplotypes. For each core cluster, the haplotype with the highest frequency was used to represent this core cluster. We constructed a phylogenetic tree for the modern human core haplotypes together with Neanderthal, Denisovan, Chimpanzee, and ancestral sequences for Homo sapiens using FastME\textsuperscript{69} and the substitution model of Tamura-Nei\textsuperscript{82} with 1000 bootstrap.

**Estimate the time to the most recent common ancestor**

To estimate the TMRCA of a group of sequences (or haplotypes) in a certain region, we first computed the mean pairwise nucleotide differences of these sequences ($\pi$) as follows:

$$\pi = \frac{\sum_{i<j} \pi_{ij}}{\binom{n}{2}},$$

where $n$ is the number of sequences analyzed in a given region ab and $\pi_{ij}$ is the total nucleotide difference between haplotypes $i$ and $j$ ($i \neq j$).

Then the TMRCA can be estimated with the following formula:

$$\text{TMRCA} = \frac{\pi}{2 \times \mu_{ab} \times l_{ab}},$$

where $\mu_{ab}$ is the local mutation rate of a genomic region with length $l_{ab}$ from position a to b. The mutation rate can be estimated as follows:

$$\mu_{ab} = \frac{d_{\text{Hum}} - \text{AncHumChimp}}{l_{ab} \times T_{\text{Human}} - \text{Chimp}},$$

where $d_{\text{Hum}} - \text{AncHumChimp}$ is the nucleotide difference between the human reference genome and that of the most recent common ancestor of humans and chimpanzees of the region. $T_{\text{Human}} - \text{Chimp}$ is the divergence time of humans and chimpanzees; a value of 13 million years was used in this study.\textsuperscript{83}
Estimation of the selection-onset time
We used startmrca to estimate the selection-onset time of these beneficial alleles\textsuperscript{37} based on all the haplotypes of CHB. startmrca uses a hidden Markov model which takes into account the length of distribution of the shared ancestral haplotype, the accumulation of derived mutations on the ancestral background, and the surrounding background haplotype diversity. For each locus, we used a 1-Mb region surrounding it and the HapMap combined recombination map in this region. The local mutation rate was estimated following a previous study. We set the maximum number of haplotypes to include in the selected and reference panels to be 100 and 20, respectively. We ran this with 15,000 iterations. The first 9,000 iterations were removed as a burn-in process. We retained the remaining successful iterations as the final result. We took the 2.5 and 97.5 quantiles to obtain the 95% CIs.

Inference of the allele frequency change over time
We inferred the frequency trajectories of these putative adaptive alleles of these highly differentiated SNPs at these thiamine-related genes using RELATE\textsuperscript{35} and CLUES.\textsuperscript{36} We first estimated the population history of the CHB population for each chromosome using RELATE. This step estimated the coalescence rate that was used by CLUES. We then constructed the local genealogical tree using all the haplotypes from the CHB population. We next calculated the evidence of positive selection of each locus. SNPs with signals of positive selection were used for further analyses. We further resampled the branch length 100 times for each locus with evidence of positive selection using the coalescence rate of the corresponding chromosome. We used CLUES to model the allele frequencies in the last 4000 generations (100,000 years, assuming 25 years per generation).

The probability of archaic introgressed segment due to incomplete lineage sorting
We calculated the probability of observing a haplotype, of a certain length or longer, shared by modern humans and archaic hominins due to incomplete lineage sorting (ILS) as described previously.\textsuperscript{30} Briefly, the expected length of segments of ILS was modeled given the local recombination rate, time since the divergence from the common ancestor, and generation time. We compared this expected length to the length of the observed introgressed haplotype. We let $r$ be the recombination rate per generation per bp, $t_1$ the length of the human branch since divergence, and $t_2$ the archaic branch length since divergence (corrected for branch shortening). We assume a generation time ($g$) of 25 years. We get $L = 1/r \times (t_1 + t_2)/g$. The probability of a length of at least $L_0$ bp is $1 - \text{GammaCDF}(L_0, \text{shape} = 2, \text{rate} = 1/L)$.

The regional recombination rate was estimated according to the HapMap genetic map. In this study, we assumed a split time between Neanderthals and modern humans of 550,000 years ago followed by interbreeding ~50,000 years ago. We used the split time between Neanderthals and modern humans and this yielded longer haplotypes. If some archaic hominins split with modern humans earlier, the expected length would be shorter. This strategy was thus the most conservative.

Expression analysis in multiple tissues from the GTEx dataset
To determine whether the putative adaptive alleles have any effect on gene expression, we directly used the rs ID and queried the Genotype-Tissue Expression (GTEx) Project (v8 release) Portal.\textsuperscript{41} This dataset contains 52 tissues and 2 cell lines. Between 4 and 803 individuals were available for each tissue. Forty-eight tissues have samples larger than 100, suggesting adequate power to detect the expression quantitative trait loci (eQTLs).

Functional annotation
The variants were annotated using the Ensembl Variant Effect Predictor (VEP)\textsuperscript{84} with the corresponding VEP-compiled annotation database (v86_GRCh37). We retrieved the Combined Annotation Dependent Depletion (CADD) scores of these SNPs.\textsuperscript{85} We also analyzed whether the putative adaptive sequences have a regulatory effect on genes using the Ensembl regulatory annotation database.\textsuperscript{67}

We further investigated the phenotypic legacy of the putative adaptive alleles using the association results from GWAS Catalog\textsuperscript{66} (accessed on 15 December, 2020). We also did a comprehensive search of publications that studied traits affected by loci in these genes involved in thiamine metabolism. We
investigated these traits that may be affected by these SNPs with the signal of positive selection or SNPs in LD with SNPs potentially under positive selection.

**QUANTIFICATION AND STATISTICAL ANALYSIS**

The Method details provide in-depth descriptions of the quantifications and statistical analyses used in this manuscript.