Positive selection affects the human ADLH2 gene expression: genetic adaptation to alcohol consumption

Helmut Schaschl (✉ helmut.schaschl@univie.ac.at)

University of Vienna: Universität Wien <https://orcid.org/0000-0003-2826-8773>

Tobias Göllner

University of Vienna: Universität Wien

David L Morris

King's College London

Research Article

Keywords: ALDH2, alcohol metabolism, eQTLs, positive selection, genetic adaptation

Posted Date: September 13th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-891422/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Abstract

ALDH2 is a key enzyme in alcohol metabolism that protects cells from acetaldehyde toxicity. Using iHS and F_{ST} statistics, we identified regulatory acting variants affecting *ALDH2* gene expression under positive selection in populations of European ancestry. Several SNPs (rs3184504, rs4766578, rs10774625, rs597808, rs653178, rs847892, rs2013002) that function as eQTLs for *ALDH2* in various tissues showed evidence of positive selection. Very large pairwise F_{ST} values indicated high genetic differentiation at these loci between populations of European ancestry and populations of other global ancestries. Estimating the timing of positive selection on the beneficial alleles suggests that these variants were recently adapted approximately 3000 to 3700 years ago. The derived beneficial alleles are in complete linkage disequilibrium with the derived *ALDH2* promoter variant rs886205, which is associated with higher transcriptional activity. The SNPs rs4766578 and rs847892 are located in binding sequences for the transcription factor *HNF4A*, which is an important regulatory element of *ALDH2* gene expression. In contrast to the missense variant *ALDH2* rs671 (*ALDH2*2*), which is common only in East Asian populations and is associated with greatly reduced enzyme activity and alcohol intolerance, the beneficial alleles of the regulatory variants identified in this study are associated with increased expression of *ALDH2*. This suggests adaptation of Europeans to higher alcohol consumption.

1. Introduction

The Neolithic transition from a hunter-gatherer lifestyle to an agriculturist one, about 9,000–13,000 years ago, included substantial changes in food processing and dietary habits associated with plant and animal domestication (Ye and Gu 2011). One of the key questions in biological anthropology is whether these changes resulted in selective pressure, influencing the expression of genes in the human genome. Identifying such loci has the potential to identify the underlying genetic variants contributing to the risk for various human diseases such as autoimmune diseases, cancer or cardiovascular disease (Luca et al. 2010; Valente et al. 2015; Ye and Gu 2011). Alcohol consumption and culture-related drinking behavior is probably one of the major changes in human dietary habits and lifestyle over the last 10,000 years. Production of larger amounts of alcoholic beverages had probably begun by the early Neolithic. A recent study reports archaeological evidence for cereal-based beer brewing by the semi-nomadic Natufians at Raqefet Cave (Mount Carmel in the north of Israel) dating back 11,700–13,700 years ago (Liu et al. 2018). Today, large amounts of alcohol are consumed in many societies. Recent data from the World Health Organization (WHO) show that worldwide about 3 million deaths and 132.6 million disability-adjusted life years are attributable to the harmful use of alcohol (World Health Organization 2018). In particular, Europe stands out in the WHO data as the region with the highest alcohol consumption and the highest burden of alcohol-related diseases. Large amounts of episodic drinking (binge drinking) as well as chronic alcohol consumption are associated with several very harmful effects such as alcoholic liver disease, intestinal inflammation, cancer, hypertension, brain damage including adverse behavioural changes, and decreased fertility (Liu et al. 2016; Ricci et al. 2017; Rocco et al. 2014; Roerecke et al. 2019). While heavy alcohol consumption can cause complex negative physiological effects, positive effects of moderate alcohol consumption have also been reported. Light to moderate alcohol consumption has been associated with a reduced risk of some forms of cardiovascular disease and autoimmune diseases (Aguet et al. 2017; Barbhuiya and Costenbader 2016; Fernandez-Sola 2015; Lu et al. 2014; Ronksley et al. 2011). The most commonly ingested alcohol is ethanol (EtOH, $\text{CH}_3\text{CH}_2\text{OH}$) which is absorbed from the gastrointestinal tract by passive diffusion. EtOH is oxidized in the first step, mainly in the liver, to acetaldehyde by alcohol dehydrogenases (ADH). The genes *ADH4*, *ADH1A*, *ADH1B* and *ADH1C*, which are located in an array in the region of chromosome 4q23, encode closely related proteins and carry out most of the EtOH oxidation in liver. Cytochrome P450 2E1 (*CYP2E1*) and the enzyme catalase (*CAT*) also participate in this metabolic pathway, albeit to a lesser extent. In the second step of EtOH metabolism, acetaldehyde, which is a chemically reactive and toxic compound, is oxidized by aldehyde dehydrogenases (ALDHs) to acetate (Cederbaum 2012; Edenberg and McClintick 2018).

Several studies provide evidence that recent positive selection acts on the *ADH1B* locus in Asian, European and African populations (Galinsky et al. 2016; Gu et al. 2018; Han et al. 2007; Johnson and Voight 2018; Li et al. 2007; Peng et al. 2010). At this locus, two missense substitutions play a role at the SNPs rs1229984 (G > A; p.Arg48His) and rs2066702 (C > T; p.Arg370Cys). They express three different isoforms. The *ADH1B*1* isoform with arginine at both codon positions is the most common allele globally, except in populations of East Asian ancestry. In East Asia, the derived allele *ADH1B*2* (rs1229984) presents the common allele (with a frequency of about 0.70). The *ADH1B*3* allele (rs2066702) occurs only in individuals of African ancestry, with allele frequencies ranging from 0.09 to 0.28. The two derived isoforms *ADH1B*2* and *ADH1B*3* metabolise EtOH at about 11 and 3 times the rate of *ADH1B*1*, respectively, (Edenberg and McClintick 2018). Several studies report that rs1229984 in the *ADH1B* locus is associated with a reduced risk of alcoholism in different ancestries (Bierut et al. 2012; Craddock et al. 2010; Gelernter et al. 2014; Jorgenson et al. 2017; Thompson et al. 2020; Toth et al. 2010; Way et al. 2015; Xu et al. 2015). The positive selection on the derived allele was estimated to have occurred about 7,000 to 15,000 years ago (Peng et al. 2010; Peter et al. 2012; Smith et al. 2018), which overlaps with the time-frame of the origin and expansion of Neolithic agriculture in East Asia. Nonetheless, it remains unclear whether the driving selective force acting on this genetic polymorphism emanates from the protective effect against alcohol dependence or from the higher efficiency of this polymorphism in metabolizing EtOH.

The mitochondrial enzyme ALDH2 plays the key role in the second step of EtOH metabolism by converting acetaldehyde into acetate. ALDH2 is not only a major detoxification enzyme for EtOH-derived acetaldehyde, but is also involved in detoxifying reactive aldehydes derived from reactive oxygen species (ROS) (Budav et al. 2009). Aldehydes are toxic molecules that can form genotoxic DNA- and protein-adducts in cells (Rodriguez-Zavala et al. 2019). Accumulation of high levels of acetaldehyde can be mutagenic, carcinogenic (Chen et al. 2014; Lee et al. 2017; Liu et al. 2016) and may negatively affect the immune system (Ceni et al. 2014). In contrast to the *ADH* genes, *ADLH2* is expressed in most human tissues, with high levels in

the liver, heart, kidney, and muscle tissues (Stewart et al. 1998). In the coding region of *ALDH2*, the missense variant rs671 (G > A; p.Glu504Lys) expresses the isoforms *ALDH2*1* and *ALDH2*2*. The *ALDH2*2* variant is found only in individuals of East Asian ancestry, reaching frequencies of up to 40% in some East Asian populations such as Han Chinese and Japanese (Edenberg and McClintick 2018; Eng et al. 2007; Li et al. 2009; Oota et al. 2004). This allele significantly affects alcohol metabolism because it results in an inactive enzyme and thus an excess of the toxic acetaldehyde in cells, even with moderate alcohol consumption. The symptoms are severe facial flushing, nausea, headache and tachycardia (Macgregor et al. 2009). East Asians homozygous for *ALDH2*2* have a very low risk for alcohol dependency (Jorgenson et al. 2017; Macgregor et al. 2009; Quillen et al. 2014). The *ALDH2* enzyme plays a central role in protecting cells from EtOH toxicity by metabolizing acetaldehyde (and other endogenous aldehyde products), is anti-inflammatory (Pan et al. 2016), and functions in myocardial protection (Ma et al. 2011; Panisello-Rosello et al. 2018; Zhang et al. 2012). Accordingly, this gene is of great biomedical interest. *ALDH2* is located at the human chromosomal region 12q24.12. Several genome-wide association studies (GWAS) have found this genomic region to be associated with multiple human diseases such as rheumatoid arthritis (Coenen et al. 2009), systemic lupus erythematosus (Bentham et al. 2015), type 1 diabetes (Auburger et al. 2014), hypertension (Yasukochi et al. 2017) and coronary artery disease (Wild et al. 2017). This region, approximately 0.6 Mbp in size, encompasses in addition to *ALDH2* the genes *CUX2*, *FAM109A*, *SH2B2*, *ATXN2*, *BRAP*, *ACAD10* and *MAPKAPK5*, as well as the uncharacterized transcript ENST00000546840.3 (UniProt F8VP50 - Aldedh domain-containing protein), which partially overlaps with the genes *ACAD10* and *ADLH2*. High F_{ST} values at linked sites at the *ALDH2* locus point to some form of selection for this genomic region (Oota et al. 2004). A recent study analysing rare singletons in the Japanese population identified the SNP rs3782886, which is in linkage disequilibrium (LD) with the missense SNP rs671 in the 12q24.12 region, as under recent positive selection (Okada et al. 2018). We have previously found (Oberreiter et al. 2021) a signature of positive selection at the chromosomal region 12q24.12 in European populations. In this study, we applied population genetic models of natural selection and included functional genetic data to identify the targets of positive selection in this genomic region. Several lines of evidence indicate that recent positive selection is acting on regulatory variants that influence *ALDH2* gene expression in populations of European ancestry.

2. Materials And Methods

2.1. Genomic data

We downloaded the phased genomic datasets from the 1000 Genomes projects (phase 3; <ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/>) (Auton et al. 2015). We obtained SNP data from 12 human populations: three representative populations each from African ancestry (AFR), European ancestry (EUR), South Asian ancestry (SAS) and East Asian ancestry (EAS) (populations names in accordance to the 1000 Genomes project – see Supplementary Table 1). We excluded related individuals and did not include the admixed populations from the datasets because of the underlying statistical principle of the method used to detect positive selection. We used the software program PLINK 1.9 (Chang et al. 2015) and VCFtools v0.1.14 (Danecek et al. 2011) to process the variant call format (VCF) files. We used the following filter parameters in VCFtools: *-maf 0.05* (include only sites with a Minor Allele Frequency (MAF) greater than 0.05), *-minQ 30* (include only sites with quality value above this threshold) and *-remove-indels* (exclude sites that contain an indel). Furthermore, we excluded all SNPs that deviated from Hardy-Weinberg equilibrium (with p -value < $1e-6$) using PLINK *-hwe midp* threshold filter. We further excluded potential duplicated SNPs using bcftools version 1.10.2, (<https://github.com/samtools/bcftools/>) using the parameter *norm -Ov -check-ref w -fasta-ref human_g1k_v37.fasta* (<ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/>). SNP positions are in accordance to the human genome version GRCh37/hg19 (<https://genome-euro.ucsc.edu/>) (Kent et al. 2002).

2.2. Population genetic analyses

To detect positive selection in phased genomic population data, we used the integrated Haplotype Score (iHS) approach (Voight et al. 2006), which is implemented in the software programme selscan version 1.2.0a (Szpiech and Hernandez 2014). All scans, with default selscan model parameters, were run on phased whole chromosome data (except the Y-chromosome) with a genetic map from HapMap phase II b37 (Altshuler et al. 2010). The iHS approach compares extended haplotype homozygosity (EHH) values between alleles at a given SNP. It is based on the model of a selective sweep, in which a *de novo* adaptive mutation arises on a haplotype that is rapidly fixed in the population, thereby reducing genetic diversity around that locus (Voight et al. 2006). The unstandardized iHS scores were normalized in default frequency bins across the entire genome using the script 'norm' provided by the selscan programme. Negative iHS values (iHS score < -2.0) indicate unusually long haplotypes carrying the derived allele, and significant positive values (iHS score > 2.0) are associated with long haplotypes carrying the ancestral allele (Voight et al. 2006). We used the Ensembl Variant Effect Predictor programme package (<https://github.com/Ensembl/ensembl-vep>) (McLaren et al. 2016) to map genetic information such as gene symbol and biotype to the analysed SNPs. We calculated empirical p -values for the obtained iHS scores across all chromosomes using the R programme version 4.1.0 (R Core Team 2021). In this study we report only results for the human chromosomal region 12q24.12, the genomic location of the *ALDH2* gene. We considered statistically significant ($p < 0.01$) iHS scores > 2.4 or < -2.4; however, we also applied Bonferroni correction, which yields p -values $p < 1 \times 10^{-5}$ (= iHS scores > 4.2 or < -4.2). We used the script 'colormap.plotting.R' provided by the selscan package to display the EHH plots for the SNPs that are under positive selection. Pairwise F_{ST} were calculated using Weir & Cockerham F_{ST} calculation (Weir and Cockerham 1984) implemented in VCFtool (Danecek et al. 2011). Negative F_{ST} values were set to zero. We calculated empirical p -values for the F_{ST} values (across all chromosomes) to obtain the significant threshold ($p < 0.05$) of outlier loci. In addition, locus-specific F_{ST} values and standard deviations (sd) across all analysed populations were calculated for SNPs that were detected to be under positive selection with the Genetix programme version 4.05 (Belkhir et al. 2004) applying the jackknife resampling procedure. We used the R package ggplot2 (Wickham 2009) to plot iHS and F_{ST} values.

Allele frequency data, SNP information and ancestral/derived allele state were obtained from the Ensembl genome browser (<https://www.ensembl.org/index.html>) (Howe et al. 2021). We used LDlink, a web-based application (<https://analysistools.cancer.gov/LDlink/?tab=home>) (Machiela and Chanock 2015), to explore population-specific linkage disequilibrium (LD); we report D' and goodness-of-fit statistics (χ^2 statistics).

2.3. Estimating the timing of positive selection

We estimated the timing of selection on a beneficial allele using the R package Startmrca (Smith et al. 2018). The method applies a Markov chain Monte Carlo simulation (MCMC) that samples over the unknown ancestral haplotype to generate a sample of the posterior distribution for the time to the most recent common ancestor (TMRCA). The model takes advantage of both the length of the ancestral haplotype on each chromosome and the accumulation of derived mutations on the ancestral haplotype to generate a sample of the posterior distribution for the TMRCA. The model requires a sample (panel) containing the haplotypes with the selected allele and a reference panel of haplotypes without the selected allele. In this study, populations of European ancestry were used as samples, and populations of the other analysed genetic ancestries were used as reference panels. Because the calculated TMRCA estimates for the populations of European ancestry were very similar regardless of the reference panels used, we report in this study only the TMRCA estimates calculated for the European populations, using the European populations both as sample panel and as reference panel. We also estimated TMRCA for the East Asian-specific functional variants rs671 (*ALDH2* locus) and rs3782886 (*BRAP* locus) (Okada et al. 2018) using the East Asian population CHB as sample and reference panel (see Supplementary Table 1 for the corresponding population names). After normalising the TMRCA data we calculated 95% credible intervals (CI = 0.95) for the timing estimates using the equal-tailed interval method implemented in the R package bayestestR (Makowski et al. 2019). We used recombination rates from the sex-averaged recombination map from deCODE to model recombination rate variation across the human genome. We analysed 1 Mb regions up- and downstream of the genetic variants under selection with an assumed mutation rate of 1.6×10^{-8} . We ran three independent MCMC chains, each with 25,000 iterations. We discarded the first 9000 iterations (burn-in), retaining the remaining iterations. We assumed 25 years as generation time.

2.4. GTEx and RegulomeDB functional data

We utilized expression quantitative trait loci (eQTLs) (accessed between May and July 2021 (dbGaP Accession phs000424.v8.p2) from GTEx Portal V8 Release (<https://www.gtexportal.org/home/>) (Ardlie et al. 2015) to test whether any of the potential SNPs that are under positive selection function as eQTL. We included *cis*-eQTL variants within a 1 Mb window of analysed genes. The RegulomeDB database (<https://regulomedb.org/>) (Boyle et al. 2012) was used to obtain chromatin states; this database comprises known classes of genomic elements such as promoters, enhancers, transcription start sites, and transcription factor (TF) binding motifs. Additionally, mapped phenotype data were obtained from the NHGRI-EBI GWAS catalogue (<https://www.ebi.ac.uk/gwas/>) (Buniello et al. 2019) (accessed between May and July 2021).

3. Results

3.1 Positive selection in populations of European ancestry

The iHS analysis shows evidence that the human chromosomal region 12.q24.12 is under positive selection in populations of European ancestry. Figure 1 (a) plots the iHS scores in the European population GBR; Fig. 1 (b) shows the pairwise F_{ST} values for GBR vs. the African population LWK across 12.q24.12. The red and green lines indicate significant ($p < 0.01$ and after Bonferroni correction $p < 1 \times 10^{-5}$, respectively) iHS scores and the genome-wide threshold (95% confidence level) for F_{ST} outlier loci ($F_{ST} > 0.3$).

Table 1

SNPs under positive selection at the human chromosomal region 12q24.12 in populations with European ancestry (GBR, TSI, FIN). Given are iHS scores and the calculated (-log) p -values (in bold Bonferroni correction with $p < 1 \times 10^{-5}$), the timing (t) of positive selection on the derived beneficial allele in thousand years ago (kya) and 95% credible interval (CI) (rounded to one decimal figure), average allele frequency in % for the derived beneficial allele/ancestral allele in the different ancestries and global locus-specific F_{ST} values (sd = standard deviation) calculated across all analysed populations.

Chr	Beneficial allele/ ancestral allele	Location	iHS			iHS -logp	t (kya)	95% CI	Allele frequency in %				Locus-specific F_{ST} (sd)
			GBR	TSI	FIN				AFR	EUR	SAS	EAS	
12q24.12	rs3184504-T/C	Exon, <i>SH2B3</i>	-3.2	-3.3	-2.9	3.2; 3.3; 2.7	3.7	3.2–4.3	0.2/99.8	46/54	7/93	0.2/99.8	0.351 (0.060)
	rs4766578-T/A	Intron, <i>ATXN2</i>	-3.8	-3.1	-3.1	4.1; 3.0; 3.0	3.5	3.0–4.0	0.2/99.8	48/52	7/93	0.2/99.8	0.366 (0.063)
	rs10774625-A/G	Intron, <i>ATXN2</i>	-3.8	-3.1	-3.0	4.1; 3.0; 2.9	3.0	2.7–3.4	0.2/99.8	48/52	7/93	0.2/99.8	0.366 (0.062)
	rs597808-A/G	Intron, <i>ATXN2</i>	-3.8	-4.2	-3.1	4.1; 4.8; 3.0	3.5	3.0–4.1	0.2/99.8	47/53	7/93	0.2/99.8	0.352 (0.064)
	rs653178-C/T	Intron, <i>ATXN2</i>	-3.5	-4.4	-3.3	3.7; 5.2 ; 3.3	3.1	2.6–3.7	0.3/99.7	47/53	7/93	0/100	0.356 (0.063)
	rs847892-G/A	Intron, <i>ACAD10</i>	-2.7	-2.7	-2.7	2.5; 2.5; 2.5	6.0	5.1–7.0	0.4/99.6	69/31	30/70	7/93	0.405 (0.080)
	rs2013002-T/C	Intron, <i>ENST00000546840.3</i>	-2.7	-2.1	-2.4	2.5; 1.8; 2.1	3.1	2.8–3.6	0.3/99.7	41/59	6/94	0.3/99.7	0.315 (0.053)

3.2 Positive selection acts on regulatory variants of ALDH2

From the GTEx database, we obtained in total 1591 *cis*-QTLs that influence *ALDH2* gene expression (Supplementary Table 2); of these *cis*-QTLs, we identified 204, 217 and 53 eQTLs that had significant ($p < 0.01$) iHS scores in the European samples GBR, TSI and FIN, respectively (Supplementary Tables 3–5). We also obtained *cis*-eQTLs (1970 in total) for the other protein-coding genes located in this genomic region (*CUX2*, *FAM109A*, *SH2B2*, *ATXN2*, *BRAP*, *ACAD10*, *MAPKAPK5*) (Supplementary Table 6). In contrast to the eQTLs for *ALDH2*, we did not obtain significant iHS values for these SNP eQTLs, except for SNPs that also function as eQTLs for *ALDH2*. We further identified seven SNPs (rs3184504, rs4766578, rs10774625, rs597808, rs653178, rs847892, rs2013002) that are under positive selection in European populations that have very large global locus-specific F_{ST} values > 0.3 , i.e. are outlier loci (Table 1). The corresponding EHH plots and pairwise F_{ST} values of these SNPs can be found in Supplementary Fig. 1 and Supplementary Table 7, respectively. The pairwise F_{ST} values for these SNPs comparing populations of European ancestries vs. African, East Asian and South Asian ancestries ranged from 0.253 to 0.691. These SNPs function as eQTLs for *ALDH2*, and the beneficial alleles are associated with increased *ALDH2* gene expression in various human tissues (in accordance to the set of tissues represented in GTEx) such as esophagus – mucosa, skin, muscle – skeletal, brain – nucleus accumbens, artery – tibial, artery – aorta, and thyroid. The average allele frequencies for these SNPs are given in Table 1; for most of these SNPs the frequency of the derived beneficial alleles reaches almost 50% in the European populations. In contrast, the derived alleles are very rare ($< 0.3\%$) in African and East Asian ancestries and at low frequency in populations of South Asian ancestry ($< 7\%$ with the exception of rs847892). We also compared the allele frequency at these loci with ancient Eurasians, including ancient hunter-gatherers (8.2–7.5 kya) from the study of Mathieson et al. (2015). The allele frequency data from the latter study show for the SNPs rs3184504, rs4766578, rs10774625, and rs653178 that the ancestral alleles were fixed in ancient European hunter-gatherers. As expected, the Neanderthal and Denisovan data on the UCSC Genome Browser also show only ancestral alleles at these loci. In contrast, early European farmers (8.4–4.2 kya) and individuals with steppe ancestry (5.4–3.6 kya) had frequencies between 8% and 25% of the derived alleles at these loci.

In the European sample GBR we estimated the timing of positive selection of the derived beneficial alleles to be from about 3.0 to 3.7 kya with the exception of SNP rs847892, which we date at 6.0 kya (Table 1). This range of estimates are very similar to the TMRCA estimates calculated for the other European samples (TSI and FIN) (Supplementary Table 8). We also calculated the TMRCA for the derived allele of the East Asian-specific polymorphism (missense variants) at rs671-A/G and rs3782886-C/T in the East Asian population CHB, which yielded an estimation of 5.8 kya (CI: 4.8–6.7) and 5.4 kya (CI: 3.3–6.5), respectively.

We then further included in the analysis the *ALDH2* promoter variant rs886205-A/G, which is located – 360 bp from the ATG start codon of the *ALDH2* gene (Chou et al. 1999). This promoter variant shows very large genetic differentiation with global locus-specific $F_{ST} = 0.378$ (s.d. = 0.055). In the 1000 Genome data the derived allele A is the common allele in European and South Asian populations with average frequencies of about 83% and

71%, respectively. In contrast, in populations of African and East Asian ancestry the common allele is the ancestral allele G with frequencies of about 78% and 84%, respectively. For the *ALDH2* promoter variant, a study showed (in vivo and in vitro experiments) that the - 360G (ancestral) allele has a significantly lower basal transcriptional activity than the - 360A (derived) allele (Kimura et al. 2009). Our LD analysis revealed that the positively selected SNPs are in complete LD ($D' = 1$) with the *ALDH2* promoter variant rs886205 (Table 2). The chromatin state data from RegulomeDB showed that the identified SNPs are associated with active transcription start site (TSS), enhancers and strong transcription in different tissues (Table 3). Importantly, the positively selected SNPs rs4766578 and rs847892 are located in the binding motif for the transcription factor *hepatocyte nuclear factor 4 alpha* (*HNF4A*). This transcription factor is an important regulatory element of *ALDH2*. The mapped phenotypes (Table 3) show that the positively selected SNPs are associated with various traits and diseases, in particular with blood pressure, cardiovascular disease, cholesterol level and autoimmune diseases. The variants rs597808 and rs2013002 are also associated with alcohol drinking and physiological traits such as blood pressure. We pooled related traits (Supplementary Table 9) into four main trait category namely autoimmune diseases (AIS), blood pressure (BP), cardiovascular disease (CDS) and cancer to test the null hypothesis that the traits and the allele state are independent. We found a significant ($\chi^2 = 28.828$, $df = 3$, $p\text{-value} = 2.4\text{e-}06$) relationship between the allele state and trait; the derived beneficial alleles are positively associated with AIS, BP and CDS whereas the ancestral alleles with cancer.

Table 2
Pairwise LD (D') of SNPs under positive selection in populations of European ancestry (GBR, TSI, FIN) and the *ALDH2* promoter (*) variant rs886205; all calculated D' values with $p\text{-value} < 0.0001$ (χ^2 statistics).

Chr:pos	SNP	LD (<i>D'</i>)						
		rs3184504	rs4766578	rs10774625	rs597808	rs653178	rs847892	rs2013002
chr12:111884608	rs3184504	-						
chr12:111904371	rs4766578	1.0	-					
chr12:111910219	rs10774625	1.0	1.0	-				
chr12:111973358	rs597808	0.98	0.986	0.986	-			
chr12:112007756	rs653178	0.98	0.979	0.979	0.859			
chr12:112141570	rs847892	0.846	0.851	0.851	0.846	1.0	-	
chr12:112200150	rs2013002	0.956	0.970	0.97	1.0	1.0	1.0	-
chr12:112204427	*rs886205	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Table 3

GTEEx and RegulomeDB data on SNPs under positive selection in European populations (GBR, TSI, FIN). Given is also a summary of reported traits from the NHGRI-EBI GWAS catalogue. GTEEx eQTLs–eGene interaction with $p < 0.0001$. RegulomeDB rank: 2b: TF binding + any motif + DNase Footprint + DNase peak; 3a: TF binding + any motif + DNase peak; 4–5: TF binding + DNase peak; 6: motif hit. The RegulomeDB probability score ranges from 0 to 1, with 1 being most likely to be a regulatory variant (for further details see {Boyle, 2012} and {Dong, 2019}). Transcription factor *HNF4A*, an important regulatory element of the *ALDH2* gene expression, is given in bold.

GTEEx		RegulomeDB				GWAS
eQTL	eGene	rank	score	chromatin state	motif	reported traits
rs3184504	<i>ALDH2</i> , <i>LINC01405</i> , <i>TMEM116</i>	3a	0.67022	strong transcription; enhancers	<i>MTF1</i>	Cardiovascular disease, blood pressure, ischemic stroke, glaucoma, rheumatoid arthritis, cancer, celiac disease, type I diabetes mellitus, parental longevity, inflammatory bowel disease, multiple sclerosis, blood cell count, hypothyroidism, haemoglobin measurement.
rs4766578	<i>ALDH2</i> , <i>TMEM116</i>	2b	0.63936	strong transcription; enhancers	<i>ESRRA</i> , <i>ESRRB</i> , <i>HNF4A</i> , <i>NR6A1</i>	Sjögren's syndrome, reticulocyte fraction of red cells, arthritis, vitiligo, HDL cholesterol, smoking status, coronary artery disease.
rs10774625	<i>ALDH2</i> , <i>ADAM1B</i> , <i>TMEM116</i>	5	0	strong transcription; enhancers	<i>FOXJ2</i> , <i>FOXQ1</i>	Hypertension, myocardial infarction, coronary artery disease, asthma, cholesterol levels, systemic lupus erythematosus, urate measurement, life span, systolic blood pressure, hypothyroidism, glomerular filtration rate.
rs597808	<i>ALDH2</i> , <i>LINC01405</i> , <i>ADAM1B</i>	5	0.13454	strong transcription	-	Systolic blood pressure, alcohol drinking, diastolic blood pressure, cholesterol levels, apolipoprotein B levels, colorectal cancer, allergic diseases, haematocrit, systemic lupus erythematosus, allergy.
rs653178	<i>ALDH2</i> , <i>LINC01405</i>	4	0.60906	active TSS; strong transcription; enhancers	-	Allergic disease, asthma, celiac disease, cholesterol level, eczema, Crohn's disease, chronic kidney disease, blood pressure, eosinophil counts, inflammatory bowel disease, type 1 diabetes, urate level.
rs847892	<i>ALDH2</i> , <i>TMEM116</i> , <i>NAA25</i>	6	0.20016	active TSS; strong transcription; enhancers	<i>HNF4A</i>	No data.
rs2013002	<i>ALDH2</i> , <i>ADAM1B</i>	6	0.55195	active TSS; enhancers	<i>MAFB</i> , <i>MAFK</i>	Alcohol drinking and blood pressure.

4. Discussion

4.1 Evidence of position selection acting on regulatory variants of *ALDH2*

This study provides evidence of positive selection across the human chromosomal region 12q24.12. This finding is in line with two previous studies (Akbari et al. 2018; Barreiro and Quintana-Murci 2010). We further found that this genomic region is enriched in eQTLs that influence *ALDH2* gene expression. A high number of these SNP eQTLs had significant iHS scores in the populations of European ancestry. In contrast, *cis*-eQTLs of the other genes located at chr12q24.12 showed no significant iHS values. This indicates that the target of positive selection are regulatory acting variants that influence *ALDH2* gene expression. We identified seven SNPs that are under positive selection and show very large global locus-specific F_{ST} values (> 0.3), indicating high genetic differentiation between populations of European ancestry and populations from other global ancestries. The GTEEx data show that these SNPs function primarily as eQTLs for the *ALDH2* gene. The derived beneficial alleles at these SNP eQTLs are associated with increased expression of *ALDH2* in multiple human tissues. Importantly, the two variants rs4766578 and rs847892 are located in binding sequences for transcription factor *HNF4A*. That transcription factor is considered to be a master regulator of liver-specific gene expression (Bolotin et al. 2010) and is an important regulatory element of *ALDH2* gene expression (Stewart et al. 1998; You et al. 2002). Furthermore, these SNP eQTLs are in complete LD with the *ALDH2* promoter variant rs847892. This promoter polymorphism influences individual differences in acetaldehyde elimination. The ancestral allele G, the common allele in populations of African and East Asian ancestry, has a lower basal transcriptional activity than the derived allele A, the common allele in populations of European and South Asian ancestry (Kimura et al. 2009). These results suggest that higher transcriptional activity and increased *ALDH2* expression in individuals of European ancestry represent a form of genetic adaptation to increased alcohol consumption, possibly enabling faster detoxification of acetaldehyde.

The derived beneficial alleles of these loci reach almost 50% in the European population, whereas in African and East Asian populations the frequencies are very low (< 0.003). The ancestral alleles at these positively selected loci appear to be fixed in ancient European hunter-gatherers, but in early farmers and individuals with steppe ancestry the frequencies of the derived alleles already range between 8% and 25% (Mathieson et al. 2015). The estimated timing of positive selection on the beneficial alleles in the European population GBR ranges from about 3.0 kya to 3.7 kya

(except for rs847892 for which TMRCA was estimated to 6.0 kya). We also calculated the TMRCA for the East Asian-specific derived alleles rs671-A and rs3782886-C, yielding an estimation of 5.8 kya (CI: 4.8–6.7) and 5.4 kya (CI: 4.3–6.5), respectively. Rs3782886, which is in LD with rs671, shows signals of very recent selection for the past 2000–3000 years in the Japanese population as reported in a recent study (Okada et al. 2018). Noteworthy, rs671-A and rs1229984-A (*ADH1B* locus) were found in a subsequent study to be significantly associated with better survival in the Japanese population (Sakaue et al. 2020). The estimated TMRCA for the derived alleles in our study suggests that these alleles spread in East Asia at a much earlier time than the beneficial alleles in populations of European ancestry. Archaeological evidence indicates early production of fermented alcohol in China (McGovern et al. 2004). Analysis of starch granules, phytoliths and fungi in food residues adhering to 8000–7000 year-old alcohol-making pottery vessels suggests that, in East Asia in the early Neolithic, alcoholic beverages were already being produced (Liu et al. 2019). For Europe, archaeologically recognizable brewing material in Central European lakeside settlements show that alcoholic beverages were being produced in this region in the late Neolithic period about 6000 years ago (Heiss et al. 2020). Later, in Greek-Roman antiquity, a richly developed viticulture with high wine production was achieved and, in this period, wine became part of the daily diet of many people (Retief and Cilliers 2015). Alcohol consumption has apparently increased steadily since then in Europe, especially in the 19th century. In Germany, for example, the high level of consumption, in particular of strong spirits, in the early 19th century was – in analogy to the plague – referred to as Branntweinpest (brandy plague). Since the rs671-A allele leads to an inactive enzyme and thus to an excess of toxic acetaldehyde in cells with its negative physiological effects, we suggest that this allele may explain the differences in the signature of positive selection between populations of European and East Asian ancestry.

4.2 Wider biomedical context

The ALDH2 enzyme plays a critical role both in the detoxification of both acetaldehyde and ROS-generated aldehyde adducts such as 4-hydroxy-2-nonenal and malondialdehyde. This enzyme thus has cytoprotective effects reducing oxidative stress (Budasz et al. 2009; Guo et al. 2013). In particular, the *ALDH2*2* variant (rs671), which is common only in individuals of East Asian ancestry, has been intensively studied in East Asians. While individuals with the *ALDH2*2* allele have a reduced risk of developing alcoholism, it increases their cancer risk (Crabb et al. 2004; Zhang and Fu 2021). Nevertheless, this allele was found to be associated in the Japanese population with better survival (Sakaue et al. 2020). In European populations this allele is virtually absent. In our study, however, the identified variants that are under recent positive selection in European populations act as regulatory variants and are associated with increased *ADLH2* gene expression in various human tissues. This suggests that individuals carrying these beneficial alleles should be more quickly able to detoxify the body from higher amounts of acetaldehyde and ROS-generated aldehyde adducts. However, a recent study reports higher methylation in alcohol-dependent patients compared to controls in the *ALDH2* promoter region (Pathak et al. 2017). Furthermore, that study suggests that positive and negative regulatory elements interact at the *ALDH2* promoter to induce genotype-mediated epigenetic changes, leading to differential transcriptional activity of this gene. We therefore suggest that individuals carrying the beneficial alleles may be able to consume more alcohol (over longer time periods), but may also have a higher likelihood of becoming heavy drinkers and alcohol dependent. This, then, could lead to increased methylation of the *ALDH2* promoter, resulting in decreased *ALDH2* gene expression. Accordingly, the protective effects of *ALDH2* against oxidative damage through acetaldehyde would be lost, resulting in increased risk of numerous oxidative stress-related diseases such as cancer, diabetes, inflammatory disorders and cardiovascular conditions such as hypertension and stroke. Indeed, we found in this study that the derived beneficial alleles are positively associated with AIS, BP and CDS whereas the ancestral alleles with cancer.

5. Conclusions

We found that positive selection acts on regulatory variants affecting *ALDH2* gene expression in populations of European ancestry. In contrast to the known functional consequence of the *ALDH2*2* variant (rs671) in East Asians, which is associated with alcohol intolerance, in Europeans the beneficial derived alleles are associated with increased *ALDH2* gene expression. This suggests local adaptation to higher alcohol consumption in Europeans. Estimation of the timing of positive selection on the beneficial alleles suggests that these variants were recently adapted, approximately 3000 to 3700 years ago. We hypothesize that the beneficial effects of higher *ALDH2* expression leads to an increased detoxification capacity for acetaldehyde, but possibly also to increased likelihood of chronic alcohol abuse, leading to decreased *ALDH2* expression and thus increased cell toxicity from EtOH-derived acetaldehyde as well as from ROS-generated aldehydes.

Declarations

Funding: Not applicable.

Conflicts of interest: All authors declare that the research was conducted without commercial or financial relationships and therefore no potential conflicts of interest occurred.

Availability of data and material: The genomic data can be obtained from 1000 Genomes database. The generated iHS and Fst dataset are available from the corresponding author on reasonable request.

Code availability: Not applicable.

Data Availability Statement: Not applicable.

Authors' contributions:

HS conceived and designed this study; HS, TG and DM performed statistical analyses. HS wrote the draft manuscript; all authors contributed to the results, edited, read and approved the final manuscript.

Ethics approval: Not applicable.

Consent to participate: Not applicable. The 1000 Genomes data are publicly available.

Consent for publication: Not applicable.

Acknowledgments:

We thank Michael Stachowitsch from the Department of Evolutionary Anthropology, University of Vienna and Franz Suchentrunk from the University of Veterinary Medicine Vienna for valuable comments on the manuscript.

References

- Aguet F, Brown AA, Castel SE, Davis JR, He Y, Jo B, Mohammadi P, Park Y, Parsana P, Segre AV, Strober BJ, Zappala Z, Cummings BB, Gelfand ET, Hadley K, Huang KH, Lek M, Li X, Nedzel JL, Nguyen DY, Noble MS, Sullivan TJ, Tukiainen T, MacArthur DG, Getz G, Management NP, Addington A, Guan P, Koester S, Little AR, Lockhart NC, Moore HM, Rao A, Struwing JP, Volpi S, Collection B, Brigham LE, Hasz R, Hunter M, Johns C, Johnson M, Kopen G, Leinweber WF, Lonsdale JT, McDonald A, Mestichelli B, Myer K, Roe B, Salvatore M, Shad S, Thomas JA, Walters G, Washington M, Wheeler J, Bridge J, Foster BA, Gillard BM, Karasik E, Kumar R, Miklos M, Moser MT, Jewell SD, Montroy RG, Rohrer DC, Valley D, Mash DC, Davis DA, Sobin L, Barcus ME, Branton PA, Grp EMW, Abell NS, Balliu B, Delaneau O, Fresard L, Gamazon ER, Garrido-Martin D, Gewirtz ADH, Gliner G, Gloudemans MJ, Han B, He AZ, Hormozdiari F, Liu B, Kang EY, McDowell IC, Ongen H, Palowitch JJ, Peterson CB, Quon G, Ripke S, Saha A, Shabalin AA, Shimko TC, Sul JH, Teran NA, Tsang EK, Zhang H, Zhou YH, Bustamante CD, et al. (2017) Genetic effects on gene expression across human tissues. *Nature* 550: 204-+.
- Akbari A, Vitti JJ, Iranmehr A, Bakhtiari M, Sabeti PC, Mirarab S, Bafna V (2018) Identifying the favored mutation in a positive selective sweep. *Nature Methods* 15: 279-+.
- Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu FL, Bonnen PE, de Bakker PIW, Deloukas P, Gabriel SB, Gwilliam R, Hunt S, Inouye M, Jia XM, Palotie A, Parkin M, Whittaker P, Chang K, Hawes A, Lewis LR, Ren YR, Wheeler D, Muzny DM, Barnes C, Darvishi K, Hurles M, Korn JM, Kristiansson K, Lee C, McCarroll SA, Nemesh J, Keinan A, Montgomery SB, Pollack S, Price AL, Soranzo N, Gonzaga-Jauregui C, Anttila V, Brodeur W, Daly MJ, Leslie S, McVean G, Moutsianas L, Nguyen H, Zhang QR, Ghorri MJR, McGinnis R, McLaren W, Takeuchi F, Grossman SR, Shlyakhter I, Hostetter EB, Sabeti PC, Adebamowo CA, Foster MW, Gordon DR, Licinio J, Manca MC, Marshall PA, Matsuda I, Ngare D, Wang VO, Reddy D, Rotimi CN, Royal CD, Sharp RR, Zeng CQ, Brooks LD, McEwen JE, Int HapMap C (2010) Integrating common and rare genetic variation in diverse human populations. *Nature* 467: 52-58.
- Ardlie KG, DeLuca DS, Segre AV, Sullivan TJ, Young TR, Gelfand ET, Trowbridge CA, Maller JB, Tukiainen T, Lek M, Ward LD, Kheradpour P, Iriarte B, Meng Y, Palmer CD, Esko T, Winckler W, Hirschhorn JN, Kellis M, MacArthur DG, Getz G, Shabalin AA, Li G, Zhou YH, Nobel AB, Rusyn I, Wright FA, Lappalainen T, Ferreira PG, Ongen H, Rivas MA, Battle A, Mostafavi S, Monlong J, Sammeth M, Mele M, Reverter F, Goldmann JM, Koller D, Guigo R, McCarthy MI, Dermitzakis ET, Gamazon ER, Im HK, Konkashbaev A, Nicolae DL, Cox NJ, Flutre T, Wen XQ, Stephens M, Pritchard JK, Tu ZD, Zhang B, Huang T, Long Q, Lin L, Yang JL, Zhu J, Liu J, Brown A, Mestichelli B, Tidwell D, Lo E, Salvatore M, Shad S, Thomas JA, Lonsdale JT, Moser MT, Gillard BM, Karasik E, Ramsey K, Choi C, Foster BA, Syron J, Fleming J, Magazine H, Hasz R, Walters GD, Bridge JP, Miklos M, Sullivan S, Barker LK, Traino HM, Mosavel M, Siminoff LA, Valley DR, Rohrer DC, Jewell SD, Branton PA, Sobin LH, Barcus M, Qi LQ, McLean J, Hariharan P, Um KS, Wu SP, Tabor D, Shive C, Smith AM, Buia SA, et al. (2015) The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* 348: 648-660.
- Auburger G, Gispert S, Lahut S, Omur O, Damrath E, Heck M, Basak N (2014) 12q24 locus association with type 1 diabetes: SH2B3 or ATXN2 ? *World Journal of Diabetes* 5: 316-327.
- Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, Flicek P, Gabriel SB, Gibbs RA, Green ED, Hurles ME, Knoppers BM, Korbel JO, Lander ES, Lee C, Lehrach H, Mardis ER, Marth GT, McVean GA, Nickerson DA, Schmidt JP, Sherry ST, Wang J, Wilson RK, Boerwinkle E, Doddapaneni H, Han Y, Korchina V, Kovar C, Lee S, Muzny D, Reid JG, Zhu Y, Chang Y, Feng Q, Fang X, Guo X, Jian M, Jiang H, Jin X, Lan T, Li G, Li J, Li Y, Liu S, Liu X, Lu Y, Ma X, Tang M, Wang B, Wang G, Wu H, Wu R, Xu X, Yin Y, Zhang D, Zhang W, Zhao J, Zhao M, Zheng X, Gupta N, Gharani N, Toji LH, Gerry NP, Resch AM, Barker J, Clarke L, Gil L, Hunt SE, Kelman G, Kulesha E, Leinonen R, McLaren WM, Radhakrishnan R, Roa A, Smirnov D, Smith RE, Streeter I, Thormann A, Toneva I, Vaughan B, Zheng-Bradley X, Grocock R, Humphray S, James T, Kingsbury Z, Sudbrak R, Albrecht MW, Amstislavskiy VS, Borodina TA, Lienhard M, Mertes F, Sultan M, Timmermann B, Yaspo M-L, Fulton L, Fulton R, et al. (2015) A global reference for human genetic variation. *Nature* 526: 68-74.

- Barbhaiya M, Costenbader KH (2016) Environmental exposures and the development of systemic lupus erythematosus. *Current Opinion in Rheumatology* 28: 497-505.
- Barreiro LB, Quintana-Murci L (2010) From evolutionary genetics to human immunology: how selection shapes host defence genes. *Nature Reviews Genetics* 11: 17-30.
- Belkhir K, Borsa P, Chikhi L, Raufaste N, Bonhomme F (2004) GENETIX4. 05, logiciel sous Windows TM pour la génétique des populations. Laboratoire génome, populations, interactions, CNRS UMR 5000: 1996-2004.
- Bentham J, Morris DL, Graham DSC, Pinder CL, Tombleson P, Behrens TW, Martin J, Fairfax BP, Knight JC, Chen L, Replogle J, Syvanen A-C, Ronnblom L, Graham RR, Wither JE, Rioux JD, Alarcon-Riquelme ME, Vyse TJ (2015) Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nature Genetics* 47: 1457-+.
- Bierut LJ, Goate AM, Breslau N, Johnson EO, Bertelsen S, Fox L, Agrawal A, Bucholz KK, Gruzza R, Hesselbrock V, Kramer J, Kuperman S, Nurnberger J, Porjesz B, Saccone NL, Schuckit M, Tischfield J, Wang JC, Foroud T, Rice JP, Edenberg HJ (2012) ADH1B is associated with alcohol dependence and alcohol consumption in populations of European and African ancestry. *Molecular Psychiatry* 17: 445-450.
- Bolotin E, Liao HL, Ta TC, Yang CH, Hwang-Verslues W, Evans JR, Jiang T, Sladek FM (2010) Integrated Approach for the Identification of Human Hepatocyte Nuclear Factor 4 alpha Target Genes Using Protein Binding Microarrays. *Hepatology* 51: 642-653.
- Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, Karczewski KJ, Park J, Hitz BC, Weng S, Cherry JM, Snyder M (2012) Annotation of functional variation in personal genomes using RegulomeDB. *Genome Research* 22: 1790-1797.
- Budas GR, Disatnik MH, Mochly-Rosen D (2009) Aldehyde Dehydrogenase 2 in Cardiac Protection: A New Therapeutic Target? *Trends in Cardiovascular Medicine* 19: 158-164.
- Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E, Suveges D, Vrousitou O, Whetzel PL, Amode R, Guillen JA, Riat HS, Trevanion SJ, Hall P, Junkins H, Flicek P, Burdett T, Hindorf LA, Cunningham F, Parkinson H (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research* 47: D1005-D1012.
- Cederbaum AI (2012) Alcohol Metabolism. *Clinics in Liver Disease* 16: 667-+.
- Ceni E, Mello T, Galli A (2014) Pathogenesis of alcoholic liver disease: Role of oxidative metabolism. *World Journal of Gastroenterology* 20: 17756-17772.
- Chang CC, Chow CC, Tellier L, Vattikuti S, Purcell SM, Lee JJ (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4.
- Chen CH, Ferreira JCB, Gross ER, Mochly-Rosen D (2014) TARGETING ALDEHYDE DEHYDROGENASE 2: NEW THERAPEUTIC OPPORTUNITIES. *Physiological Reviews* 94: 1-34.
- Chou WY, Stewart MJ, Carr LG, Zheng D, Stewart TR, Williams A, Pinaire J, Crabb DW (1999) An A/G polymorphism in the promoter of mitochondrial aldehyde dehydrogenase (ALDH2): Effects of the sequence variant on transcription factor binding and promoter strength. *Alcoholism-Clinical and Experimental Research* 23: 963-968.
- Coenen MJH, Trynka G, Heskamp S, Franke B, van Diemen CC, Smolonska J, van Leeuwen M, Brouwer E, Boezen MH, Postma DS, Platteel M, Zanen P, Lammers J, Groen HJM, Mali W, Mulder CJ, Tack GJ, Verbeek WHM, Wolters VM, Houwen RHJ, Mearin ML, van Heel DA, Radstake T, van Riel P, Wijmenga C, Barrera P, Zhernakova A (2009) Common and different genetic background for rheumatoid arthritis and coeliac disease. *Human Molecular Genetics* 18: 4195-4203.
- Crabb DW, Matsumoto M, Chang D, You M (2004) Overview of the role of alcohol dehydrogenase and aldehyde dehydrogenase and their variants in the genesis of alcohol-related pathology. *Proceedings of the Nutrition Society* 63: 49-63.
- Craddock N, Hurles ME, Cardin N, Pearson RD, Plagnol V, Robson S, Vukcevic D, Barnes C, Conrad DF, Giannoulatou E, Holmes C, Marchini JL, Stirrups K, Tobin MD, Wain LV, Yau C, Aerts J, Ahmad T, Andrews TD, Arbury H, Attwood A, Auton A, Ball SG, Balmforth AJ, Barrett JC, Barroso I, Barton A, Bennett AJ, Bhaskar S, Blaszczyk K, Bowes J, Brand OJ, Braund PS, Bredin F, Breen G, Brown MJ, Bruce IN, Bull J, Burren OS, Burton J, Byrnes J, Caesar S, Clee CM, Coffey AJ, Connell JMC, Cooper JD, Dominiczak AF, Downes K, Drummond HE, Dudakia D, Dunham A, Ebbs B, Eccles D, Edkins S, Edwards C, Elliot A, Emery P, Evans DM, Evans G, Eyre S, Farmer A, Ferrier IN, Feuk L, Fitzgerald T, Flynn E, Forbes A, Forty L, Franklyn JA, Freathy RM, Gibbs P, Gilbert P, Gokumen O, Gordon-Smith K, Gray E, Green E, Groves CJ, Grozeva D, Gwilliam R, Hall A, Hammond N, Hardy M, Harrison P, Hassanali N, Hebaishi H, Hines S, Hinks A, Hitman GA, Hocking L, Howard E, Howard P, Howson JMM, Hughes D, Hunt S, Isaacs JD, Jain M, Jewell DP, Johnson

- T, Jolley JD, Jones IR, Jones LA, et al. (2010) Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* 464: 713-U86.
- Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, Genomes Project Anal G (2011) The variant call format and VCFtools. *Bioinformatics* 27: 2156-2158.
- Edenberg HJ, McClintick JN (2018) Alcohol Dehydrogenases, Aldehyde Dehydrogenases, and Alcohol Use Disorders: A Critical Review. *Alcoholism-Clinical and Experimental Research* 42: 2281-2297.
- Eng MY, Luczak SE, Wall TL (2007) ALDH2, ADH1B, and ADH1C genotypes in Asians: A literature review. *Alcohol Research & Health* 30: 22-27.
- Fernandez-Sola J (2015) Cardiovascular risks and benefits of moderate and heavy alcohol consumption. *Nature Reviews Cardiology* 12: 576-587.
- Galinsky KJ, Bhatia G, Loh PR, Georgiev S, Mukherjee S, Patterson NJ, Price AL (2016) Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1B in Europe and East Asia. *American Journal of Human Genetics* 98: 456-472.
- Gelernter J, Kranzler HR, Sherva R, Almasy L, Koesterer R, Smith AH, Anton R, Preuss UW, Ridinger M, Rujescu D, Wodarz N, Zill P, Zhao H, Farrer LA (2014) Genome-wide association study of alcohol dependence: significant findings in African-and European-Americans including novel risk loci. *Molecular Psychiatry* 19: 41-49.
- Guo JM, Liu AJ, Zang P, Dong WZ, Ying L, Wang W, Xu P, Song XR, Cai J, Zhang SQ, Duan JL, Mehta JL, Su DF (2013) ALDH2 protects against stroke by clearing 4-HNE. *Cell Research* 23: 915-930.
- Gu S, Li H, Pakstis AJ, Speed WC, Gurwitz D, Kidd JR, Kidd KK (2018) Recent Selection on a Class I ADH Locus Distinguishes Southwest Asian Populations Including Ashkenazi Jews. *Genes* 9.
- Han Y, Gu S, Oota H, Osier MV, Pakstis AJ, Speed WC, Kidd JR, Kidd KK (2007) Evidence of positive selection on a class I ADH locus. *American Journal of Human Genetics* 80: 441-456.
- Heiss AG, Azorin MB, Antolin F, Kubiak-Martens L, Marinova E, Arendt EK, Biliaderis CG, Kretschmer H, Lazaridou A, Stika HP, Zarnkow M, Baba M, Bleicher N, Cialowicz KM, Chlodnicki M, Matuschik I, Schlichtherle H, Valamoti SM (2020) Mashies to Mashies, Crust to Crust. Presenting a novel microstructural marker for malting in the archaeological record. *Plos One* 15.
- Howe KL, Achuthan P, Allen J, Alvarez-Jarreta J, Amodio MR, Armean IM, Azov AG, Bennett R, Bhai J, Billis K, Boddu S, Charkhchi M, Cummins C, Fioretto LR, Davidson C, Dodiya K, El Houdaigui B, Fatima R, Gall A, Giron CG, Grego T, Guijarro-Clarke C, Haggerty L, Hemrom A, Hourlier T, Izuogu OG, Juettemann T, Kaikala V, Kay M, Lavidas I, Le T, Lemos D, Martinez JG, Marugan JC, Maurel T, McMahon AC, Mohanan S, Moore B, Muffato M, Oheh DN, Paraschas D, Parker A, Parton A, Prosovetskaia I, Sakthivel MP, Salam AIA, Schmitt BM, Schuilenburg H, Sheppard N, Steed E, Szpak M, Szuba M, Taylor K, Thormann A, Threadgold G, Walts B, Winterbottom A, Chakiachvili M, Chaubal A, De Silva N, Flint B, Frankish A, Hunt SE, Ilesley GR, Langridge N, Loveland JE, Martin FJ, Mudge JM, Morales J, Perry E, Ruffier M, Tate J, Thybert D, Trevanion SJ, Cunningham F, Yates AD, Zerbino DR, Flicek P (2021) Ensembl 2021. *Nucleic Acids Research* 49: D884-D891.
- Johnson KE, Voight BF (2018) Patterns of shared signatures of recent positive selection across human populations. *Nature Ecology & Evolution* 2: 713-720.
- Jorgenson E, Thai KK, Hoffmann TJ, Sakoda LC, Kvale MN, Banda Y, Schaefer C, Risch N, Mertens J, Weisner C, Choquet H (2017) Genetic contributors to variation in alcohol consumption vary by race/ethnicity in a large multi-ethnic genome-wide association study. *Molecular Psychiatry* 22: 1359-1367.
- Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D (2002) The human genome browser at UCSC. *Genome Research* 12: 996-1006.
- Kimura Y, Nishimura FT, Abe S, Fukunaga T, Tanii H, Saijoh K (2009) A Promoter Polymorphism in the ALDH2 Gene Affects Its Basal and Acetaldehyde/Ethanol-Induced Gene Expression in Human Peripheral Blood Leukocytes and HepG2 Cells. *Alcohol and Alcoholism* 44: 261-266.
- Lee DJ, Lee HM, Kim JH, Park IS, Rho YS (2017) Heavy alcohol drinking downregulates ALDH2 gene expression but heavy smoking up-regulates SOD2 gene expression in head and neck squamous cell carcinoma. *World Journal of Surgical Oncology* 15.
- Li H, Borinskaya S, Yoshimura K, Kal'ina N, Marusin A, Stepanov VA, Qin ZD, Khaliq S, Lee MY, Yang YJ, Mohyuddin A, Gurwitz D, Mehdi SQ, Rogaev E, Jin L, Yankovsky NK, Kidd JR, Kidd KK (2009) Refined Geographic Distribution of the Oriental ALDH2*504Lys (nee 487Lys) Variant. *Annals of Human Genetics* 73: 335-345.
- Li H, Mukherjee N, Soundararajan U, Tarnok Z, Barta C, Khaliq S, Mohyuddin A, Kajuna SLB, Mehdi SQ, Kidd JR, Kidd KK (2007) Geographically separate increases in the frequency of the derived ADH1B*47His allele in eastern and western Asia. *American Journal of Human Genetics* 81: 842-

- Liu L, Wang JJ, Levin MJ, Sinnott-Armstrong N, Zhao H, Zhao YA, Shao J, Di N, Zhang TE (2019) The origins of specialized pottery and diverse alcohol fermentation techniques in Early Neolithic China. *Proceedings of the National Academy of Sciences of the United States of America* 116: 12767-12774.
- Liu L, Wang JJ, Rosenberg D, Zhao H, Lengyel G, Nadel D (2018) Fermented beverage and food storage in 13,000 y-old stone mortars at Raqefet Cave, Israel: Investigating Natufian ritual feasting. *Journal of Archaeological Science-Reports* 21: 783-793.
- Liu X, Hu C, Bao MH, Li J, Liu XY, Tan XR, Zhou Y, Chen YQ, Wu SL, Chen SH, Zhang R, Jiang F, Jia WP, Wang XY, Yang XC, Cai J (2016) Genome Wide Association Study Identifies L3MBTL4 as a Novel Susceptibility Gene for Hypertension. *Scientific Reports* 6.
- Lu B, Solomon DH, Costenbader KH, Karlson EW (2014) Alcohol Consumption and Risk of Incident Rheumatoid Arthritis in Women A Prospective Study. *Arthritis & Rheumatology* 66: 1998-2005.
- Luca F, Perry GH, Di Rienzo A (2010) Evolutionary Adaptations to Dietary Changes. *Annual Review of Nutrition*, Vol 30 30: 291-314.
- Macgregor S, Lind PA, Bucholz KK, Hansell NK, Madden PAF, Richter MM, Montgomery GW, Martin NG, Heath AC, Whitfield JB (2009) Associations of ADH and ALDH2 gene variation with self report alcohol reactions, consumption and dependence: an integrated analysis. *Human Molecular Genetics* 18: 580-593.
- 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: 3555-3557.
- Ma H, Guo R, Yu L, Zhang YM, Ren J (2011) Aldehyde dehydrogenase 2 (ALDH2) rescues myocardial ischaemia/reperfusion injury: role of autophagy paradox and toxic aldehyde. *European Heart Journal* 32: 1025-1038.
- Makowski D, Ben-Shachar MS, Chen SHA, Ludecke D (2019) Indices of Effect Existence and Significance in the Bayesian Framework. *Frontiers in Psychology* 10.
- Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, Harney E, Stewardson K, Fernandes D, Novak M, Sirak K, Gamba C, Jones ER, Llamas B, Dryomov S, Pickrell J, Arsuaga JL, de Castro JMB, Carbonell E, Gerritsen F, Khokhlov A, Kuznetsov P, Lozano M, Meller H, Mochalov O, Moiseyev V, Guerra MAR, Roodenberg J, Verges JM, Krause J, Cooper A, Alt KW, Brown D, Anthony D, Lalueza-Fox C, Haak W, Pinhasi R, Reich D (2015) Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* 528: 499+.
- McGovern PE, Zhang JH, Tang JG, Zhang ZQ, Hall GR, Moreau RA, Nunez A, Butrym ED, Richards MP, Wang CS, Cheng GS, Zhao ZJ (2004) Fermented beverages of pre- and proto-historic China. *Proceedings of the National Academy of Sciences of the United States of America* 101: 17593-17598.
- McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F (2016) The Ensembl Variant Effect Predictor. *Genome Biology* 17.
- Oberreiter V, Goellner T, Morris LD, Schaschl H (2021) Positive Selection Affects the Expression of Systemic Lupus Erythematosus Associated Loci in Human Populations. *BMC Genomics*.
- Okada Y, Momozawa Y, Sakaue S, Kanai M, Ishigaki K, Akiyama M, Kishikawa T, Arai Y, Sasaki T, Kosaki K, Suematsu M, Matsuda K, Yamamoto K, Kubo M, Hirose N, Kamatani Y (2018) Deep whole-genome sequencing reveals recent selection signatures linked to evolution and disease risk of Japanese. *Nature Communications* 9.
- Oota H, Pakstis AJ, Bonne-Tamir B, Goldman D, Grigorenko E, Kajuna SLB, Karoma NJ, Kungulilo S, Lu RB, Odunsi K, Okonofua F, Zhukova OV, Kidd JR, Kidd KK (2004) The evolution and population genetics of the ALDH2 locus: random genetic drift, selection, and low levels of recombination. *Annals of Human Genetics* 68: 93-109.
- World Health Organization (2018) Global status report on alcohol and health 2018. Geneva
- Pan C, Xing JH, Zhang C, Zhang YM, Zhang LT, Wei SJ, Zhang MX, Wang XP, Yuan QH, Xue L, Wang JL, Cui ZQ, Zhang Y, Xu F, Chen YG (2016) Aldehyde dehydrogenase 2 inhibits inflammatory response and regulates atherosclerotic plaque. *Oncotarget* 7: 35562-35576.
- Panisello-Rosello A, Lopez A, Folch-Puy E, Carbonell T, Rolo A, Palmeira C, Adam R, Net M, Rosello-Catafau J (2018) Role of aldehyde dehydrogenase 2 in ischemia reperfusion injury: An update. *World Journal of Gastroenterology* 24: 2984-2994.
- Pathak H, Frieling H, Bleich S, Glahn A, Heberlein A, Nassab MH, Hillemecher T, Burkert A, Rhein M (2017) Promoter Polymorphism rs886205 Genotype Interacts With DNA Methylation of the ALDH2 Regulatory Region in Alcohol Dependence. *Alcohol and Alcoholism* 52: 269-276.

- Peng Y, Shi H, Qi XB, Xiao CJ, Zhong H, Ma RLZ, Su B (2010) The ADH1B Arg47His polymorphism in East Asian populations and expansion of rice domestication in history. *Bmc Evolutionary Biology* 10.
- Peter BM, Huerta-Sanchez E, Nielsen R (2012) Distinguishing between Selective Sweeps from Standing Variation and from a De Novo Mutation. *Plos Genetics* 8.
- Quillen EE, Chen XD, Almasy L, Yang F, He H, Li X, Wang XY, Liu TQ, Hao W, Deng HW, Kranzler HR, Gelernter J (2014) ALDH2 Is Associated to Alcohol Dependence and Is the Major Genetic Determinant of "Daily Maximum Drinks" in a GWAS Study of an Isolated Rural Chinese Sample. *American Journal of Medical Genetics Part B-Neuropsychiatric Genetics* 165: 103-110.
- R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Retief F, Cilliers L Wine in Graeco-Roman Antiquity with Emphasis on Its Effect on Health 2015
- Ricci E, Al Beitawi S, Cipriani S, Candiani M, Chiaffarino F, Vigano P, Noli S, Parazzini F (2017) Semen quality and alcohol intake: a systematic review and meta-analysis. *Reproductive Biomedicine Online* 34: 38-47.
- Rocco A, Compare D, Angrisani D, Zamparelli MS, Nardone G (2014) Alcoholic disease: Liver and beyond. *World Journal of Gastroenterology* 20: 14652-14659.
- Rodriguez-Zavala JS, Calleja LF, Moreno-Sanchez R, Yoval-Sanchez B (2019) Role of Aldehyde Dehydrogenases in Physiopathological Processes. *Chemical Research in Toxicology* 32: 405-420.
- Roerecke M, Vafaei A, Hasan OSM, Chrystoja BR, Cruz M, Lee R, Neuman MG, Rehm J (2019) Alcohol Consumption and Risk of Liver Cirrhosis: A Systematic Review and Meta-Analysis. *American Journal of Gastroenterology* 114: 1574-1586.
- Ronksley PE, Brien SE, Turner BJ, Mukamal KJ, Ghali WA (2011) Association of alcohol consumption with selected cardiovascular disease outcomes: a systematic review and meta-analysis. *Bmj-British Medical Journal* 342.
- Sakaue S, Akiyama M, Hirata M, Matsuda K, Murakami Y, Kubo M, Kamatani Y, Okada Y (2020) Functional variants in ADH1B and ALDH2 are non-additively associated with all-cause mortality in Japanese population. *European Journal of Human Genetics* 28: 378-382.
- Smith J, Coop G, Stephens M, Novembre J (2018) Estimating Time to the Common Ancestor for a Beneficial Allele. *Molecular Biology and Evolution* 35: 1003-1017.
- Stewart MJ, Dipple KM, Estonius M, Nakshatri H, Everett LM, Crabb DW (1998) Binding and activation of the human aldehyde dehydrogenase 2 promoter by hepatocyte nuclear factor 4. *Biochimica Et Biophysica Acta-Gene Structure and Expression* 1399: 181-186.
- Szpiech ZA, Hernandez RD (2014) selscan: An Efficient Multithreaded Program to Perform EHH-Based Scans for Positive Selection. *Molecular Biology and Evolution* 31: 2824-2827.
- Team RC (2021) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>. Vienna, Austria
- Thompson A, Cook J, Choquet H, Jorgenson E, Yin J, Kinnunen T, Barclay J, Morris AP, Pirmohamed M (2020) Functional validity, role, and implications of heavy alcohol consumption genetic loci. *Science Advances* 6.
- Toth R, Pocsai Z, Fiatal S, Szeles G, Kardos L, Petrovski B, McKee M, Adany R (2010) ADH1B*2 allele is protective against alcoholism but not chronic liver disease in the Hungarian population. *Addiction* 105: 891-896.
- Valente C, Alvarez L, Marks SJ, Lopez-Parra AM, Parson W, Oosthuizen O, Oosthuizen E, Amorim A, Capelli C, Arroyo-Pardo E, Gusmao L, Prata MJ (2015) Exploring the relationship between lifestyles, diets and genetic adaptations in humans. *Bmc Genetics* 16.
- Voight BF, Kudaravalli S, Wen XQ, Pritchard JK (2006) A map of recent positive selection in the human genome (vol 4, pg 154, 2006). *Plos Biology* 4: 659-659.
- Way M, McQuillin A, Saini J, Ruparelia K, Lydall GJ, Guerrini I, Ball D, Smith I, Quadri G, Thomson AD, Kasiakogia-Worley K, Cherian R, Gunwardena P, Rao H, Kottalgi G, Patel S, Hillman A, Douglas E, Qureshi SY, Reynolds G, Jauhar S, O'Kane A, Dedman A, Sharp S, Kandaswamy R, Dar K, Curtis D, Morgan MY, Gurling HMD (2015) Genetic variants in or near ADH1B and ADH1C affect susceptibility to alcohol dependence in a British and Irish population. *Addiction Biology* 20: 594-604.
- Weir BS, Cockerham CC (1984) ESTIMATING F-STATISTICS FOR THE ANALYSIS OF POPULATION-STRUCTURE. *Evolution* 38: 1358-1370.

Wickham H (2009) ggplot2: Elegant Graphics for Data Analysis. Ggplot2: Elegant Graphics for Data Analysis: 1-212.

Wild PS, Felix JF, Schillert A, Teumer A, Chen MH, Leening MJG, Volker U, Grossmann V, Brody JA, Irvin MR, Shah SJ, Pramana S, Lieb W, Schmidt R, Stanton AV, Malzahn D, Smith AV, Sundstrom J, Minelli C, Ruggiero D, Lyytikainen LP, Tiller D, Smith JG, Monnereau C, Di Tullio MR, Musani SK, Morrison AC, Pers TH, Morley M, Kleber ME, Aragam J, Benjamin EJ, Bis JC, Bisping E, Broeckel U, Cheng S, Deckers JW, Del Greco MF, Edelmann F, Fornage M, Franke L, Friedrich N, Harris TB, Hofer E, Hofman A, Huang J, Hughes AD, Kahonen M, Kruppa J, Lackner KJ, Lannfelt L, Laskowski R, Launer LJ, Leosdottir M, Lin HH, Lindgren CM, Loley C, MacRae CA, Mascalzoni D, Mayet J, Medenwald D, Morris AP, Muller C, Muller-Nurasyid M, Nappo S, Nilsson PM, Nuding S, Nutile T, Peters A, Pfeufer A, Pietzner D, Pramstaller PP, Raitakari OT, Rice KM, Rivadeneira F, Rotter JI, Ruohonen ST, Sacco RL, Samdarshi TE, Schmidt H, Sharp ASP, Shields DC, Sorice R, Sotoodehnia N, Stricker BH, Surendran P, Thom S, Toglhofer AM, Uitterlinden AG, Wachter R, Volzke H, Ziegler A, Munzel T, Marz W, Cappola TP, Hirschhorn JN, Mitchell GF, Smith NL, Fox ER, Dueker ND, et al. (2017) Large-scale genome-wide analysis identifies genetic variants associated with cardiac structure and function. *Journal of Clinical Investigation* 127: 1798-1812.

Xu K, Kranzler HR, Sherva R, Sartor CE, Almasy L, Koesterer R, Zhao HY, Farrer LA, Gelemtier J (2015) Genomewide Association Study for Maximum Number of Alcoholic Drinks in European Americans and African Americans. *Alcoholism-Clinical and Experimental Research* 39: 1137-1147.

Yasukochi Y, Sakuma J, Takeuchi I, Kato K, Oguri M, Fujimaki T, Horibe H, Yamada Y (2017) Longitudinal exome-wide association study to identify genetic susceptibility loci for hypertension in a Japanese population. *Experimental and Molecular Medicine* 49.

Ye KX, Gu ZL (2011) Recent Advances in Understanding the Role of Nutrition in Human Genome Evolution. *Advances in Nutrition* 2: 486-496.

You M, Fischer M, Cho WK, Crabb D (2002) Transcriptional control of the human aldehyde dehydrogenase 2 promoter by hepatocyte nuclear factor 4: Inhibition by cyclic AMP and COUP transcription factors. *Archives of Biochemistry and Biophysics* 398: 79-86.

Zhang H, Fu LW (2021) The role of ALDH2 in tumorigenesis and tumor progression: Targeting ALDH2 as a potential cancer treatment. *Acta Pharmaceutica Sinica B* 11: 1400-1411.

Zhang H, Gong DX, Zhang YJ, Li SJ, Hu SS (2012) Effect of mitochondrial aldehyde dehydrogenase-2 genotype on cardioprotection in patients with congenital heart disease. *European Heart Journal* 33: 1606-1614.

Figures

Figure 1

(a.) iHS p-values plotted across the human chromosomal region 12.q24.12 for the population GBR (European genetic ancestry); red/green lines: threshold for significant ($p < 0.01$; Bonferroni correction $p < 1 \times 10^{-5}$) iHS scores; (b.) pairwise F_{ST} (GBR–LWK); red line: significant outlier loci with $F_{ST} > 0.3$. Bottom: position of genes and SNPs from Table 1.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementaryfile1.docx](#)
- [Supplementaryfile2.xlsx](#)
- [Supplementaryfile3.xlsx](#)