

Article

The Aldehyde Dehydrogenase ALDH2*2 Allele, Associated with Alcohol Drinking Behavior, Dates Back to Prehistoric Times

Chih-Lang Lin ^{1,2,3,4}, Rong-Nan Chien ^{2,3,4}, Li-Wei Chen ^{1,3,4} , Ting-Shuo Huang ^{3,4,5} , Yu-Chiau Shyu ^{3,6,7} ,
Chau-Ting Yeh ^{2,4,*} and Kung-Hao Liang ^{8,9,10,*}

- ¹ Liver Research Unit, Keelung Chang Gung Memorial Hospital, Keelung 204, Taiwan; wn49792000@yahoo.com.tw (C.-L.L.); leiwei@cgmh.org.tw (L.-W.C.)
 - ² Liver Research Center, Chang Gung Memorial Hospital, Taoyuan 333, Taiwan; ronald@adm.cgmh.org.tw
 - ³ Community Medicine Research Center, Keelung Chang Gung Memorial Hospital, Keelung 204, Taiwan; huangts@adm.cgmh.org.tw (T.-S.H.); yuchiaushyu@gmail.com (Y.-C.S.)
 - ⁴ College of Medicine, Chang Gung University, Taoyuan 333, Taiwan
 - ⁵ Department of General Surgery, Chang Gung Memorial Hospital, Keelung 204, Taiwan
 - ⁶ Department of Nursing, Chang Gung University of Science and Technology, Taoyuan 333, Taiwan
 - ⁷ Institute of Molecular Biology, Academia Sinica, Taipei 115, Taiwan
 - ⁸ Department of Medical Research, Taipei Veterans General Hospital, Taipei 112, Taiwan
 - ⁹ Institute of Food Safety and Health Risk Assessment, National Yang-Ming Chiao-Tung University, Taipei 112, Taiwan
 - ¹⁰ Institute of Biomedical Informatics, National Yang-Ming Chiao-Tung University, Taipei 112, Taiwan
- * Correspondence: chaotingy@gmail.com (C.-T.Y.); kunghao@gmail.com (K.-H.L.);
Tel.: +886-3-3281200 (ext. 8129) (C.-T.Y.); +886-2-28712121 (ext. 1296) (K.-H.L.); Fax: +886-3-3282824 (C.-T.Y.)



Citation: Lin, C.-L.; Chien, R.-N.; Chen, L.-W.; Huang, T.-S.; Shyu, Y.-C.; Yeh, C.-T.; Liang, K.-H. The Aldehyde Dehydrogenase ALDH2*2 Allele, Associated with Alcohol Drinking Behavior, Dates Back to Prehistoric Times. *Biomolecules* **2021**, *11*, 1376. <https://doi.org/10.3390/biom11091376>

Academic Editor: Eric R. Gross

Received: 17 August 2021

Accepted: 15 September 2021

Published: 17 September 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: Human alcohol-consumption behavior is partly genetically encoded. The alcohol consumption of 987 residents in Keelung, Taiwan, was evaluated by using the Alcohol Use Disorder Identification Test (AUDIT). We assessed ~750,000 genomic variants of 71 residents who drank hazardously (AUDIT score ≥ 8) and 126 residents who did not drink in their daily lives (AUDIT score = 0), using high-density single nucleotide polymorphism (SNP) arrays. The rs671 G > A manifests the highest significance of the association with drinking behavior (Fisher's exact $P = 8.75 \times 10^{-9}$). It is a pleiotropic, non-synonymous variant in the aldehyde dehydrogenase 2 (ALDH2) gene. The minor allele "A", commonly known as ALDH2*2, is associated with non-drinkers. Intriguingly, identity-by-descent haplotypes encompassing genomic regions with a median length of 1.6 (0.6–2.0) million nucleotide bases were found in all study participants with either heterozygous or homozygous ALDH2*2 ($n = 81$ and 13 , respectively). We also analyzed a public-domain dataset with genome-wide genotypes of 2000 participants in Guangzhou, a coastal city in Southern China. Among them, 175 participants have homozygous ALDH2*2 genotype, and again, long ALDH2*2-carrying haplotypes were found in all 175 participants without exceptions. The median length of the ALDH2*2-carrying haplotype is 1.7 (0.5–2.8) million nucleotide bases. The haplotype lengths in the Keelung and Guangzhou cohorts combined indicate that the origin of the ALDH2*2 allele dates back to 7935 (7014–9381) years ago. In conclusion, the rs671 G > A is the leading genomic variant associated with the long-term drinking behavior among residents of Keelung, Taiwan. The ALDH2*2 allele has been in Asian populations since prehistoric times.

Keywords: aldehyde dehydrogenase; haplotype; metabolic disorder; gene-behavior interactions; germline variant

1. Introduction

Inventions of alcoholic beverages predate written history. Evidence of the earliest manmade alcoholic beverages, which date back to 9000 years ago, was found in the Henan province of China [1]. Since then, alcohol-containing beverages have made profound social,

economic and health impacts across the globe [2–8]. Alcohol can produce personally varied euphoria [9]. Human cognitive and sensory capabilities are also reduced under the influence of alcohol. The euphoria and other physiological effects altogether determine habitual drinking behaviors. Alcohol-use disorder is developed in some persons, causing burdens to families and societies. This disorder is influenced by genetics, with 60% heritability estimated in a twin study [10]. To pinpoint genetic variants associated with habitual drinking behaviors, many candidate-gene and genome-wide association studies (GWAS) have been conducted [11–17]. Interestingly, implications of drinking-behavior-related neurological and psychiatric genes remain unclear despite these explorations. Rather, genomic variants of metabolic genes, such as alcohol dehydrogenases (ADHs) and aldehyde dehydrogenases (ALDHs), have been reported frequently [11,14,15], suggesting the intimate connections between alcohol metabolisms and drinking behaviors. These genomic variants dictate personal metabolic rates of ethanol and aldehydes, including acetaldehyde [18] and 4-HNE [19]. A recent pan-ethnic group study in conjunction with a meta-analysis showed that the single nucleotide polymorphism (SNP) rs671, located at exon 12 of the ALDH2 gene, has the strongest association in East Asians [11].

Moreover, rs671 G > A is a non-synonymous, Asian-specific [20] genomic variant. The minor allele “A”, also known as the ALDH2*2 allele, encodes a glutamate-to-lysine substitution at the position 487 of the mature protein (Glu487Lys), which effectively disrupts enzymatic activities (Online Mendelian Inheritance in Man (OMIM) #100650) and causes the accumulation of aldehydes in the human body [21]. Apart from alcohol consumptions, this pleiotropic variant is associated with traits such as alcohol-induced facial flushing after drinking [22], blood pressure [23], chronic heart disease [24,25], post-stroke epilepsy [19], late-onset Alzheimer’s disease [26], fasting plasma glucose levels [27], gout [28], hip fracture and osteoporosis [29], lumbar disc herniation [30], breast cancer in postmenopausal Asian women [31], esophageal cancer [32] and hepatocellular carcinoma [33]. Considering the detrimental effects of ALDH2*2 on the enzymatic functions, which are apparently essential to humans, it remains mysterious why this allele persists in Asian populations without being purged by purifying selections. To date, the time of emergence of the ALDH2*2 allele continues to be elusive, apart from one reported value of 2000–3000 years ago shown in the literature [21,34–37]. This value first appeared in a study which investigated only five local genomic variants surrounding rs671 [34].

Community-based studies are investigations of the general populations in certain defined geographical regions. Community-based studies are suitable for investigating issues pertaining to food, lifestyle, culture, behavior patterns and occupations. The Keelung City, located in the northern coast of Taiwan, is a historic harbor city once colonized by the Spanish, Dutch and Japanese. Keelung is at a pivotal location connecting East and Southeast Asia and has been the gateway of international trades for centuries. The distance from Keelung to Mainland China is ~180 km across the Taiwan Strait. The population size is ~370,000, of whom most are of Chinese ancestry. We investigated the alcohol consumption demographics of Keelung residents by using the original, unmodified definition of the Alcohol Use Disorder Identification Test (AUDIT) score developed by the World Health Organization (WHO) in 1993 [38]. Based on prior statistics in Taiwan and worldwide, AUDIT scores ≥ 8 were indicative of hazardous drinking [39]. Thus, residents with the AUDIT scores ≥ 8 were selected for a genome-wide comparison with controls who did not drink in their daily life.

2. Materials and Methods

2.1. Study Subjects

This study was approved by the institutional review board of Chang Gung Memorial Hospital, Taiwan, and conducted according to the ethical guidelines of the 1975 Declaration of Helsinki. Written informed consents were obtained from all participants in this community study. All of the study subjects were adults at the time of enrollment. An AUDIT questionnaire in Chinese was used, which was directly translated from the WHO AUDIT

questionnaire without modifications [38]. The standard drink is defined as 10 g, consistent with WHO [38] but lower than the definition in USAUDIT (i.e., 14 g), which is optimized for residents in United States [40]. In this community study, 987 participants were enrolled into the Northeastern Taiwan Community Medicine Research Cohort (NTCMRC, ClinicalTrials.gov Identifier: NCT04839796) between February and November 2019, and screened using the AUDIT questionnaire. Among them, 24 individuals were found to drink extremely heavily, reaching the criteria of AUDIT score ≥ 16 (Supplementary Materials Figure S1). Additionally, 194 individuals were found to have $8 \leq$ AUDIT score < 16 . We selected all 24 patients with the AUDIT score ≥ 16 , and randomly selected 47 patients with $8 \leq$ AUDIT score < 16 as the heavy-drinker group in this case-control study. Most of these 71 patients are male (91.55%). To select suitable non-drinker controls for a genome-wide comparison, we performed the propensity score matching by using two basic demographic variables, namely age and gender, in the case:control ratio of 1:2, in participants who did not drink in their daily life (AUDIT score = 0). This procedure was performed by using the “MatchIt” package of the R statistical software. A total of 142 control samples were found matched with case patients. We then randomly selected 126 patients from the 142 patients as the non-drinker group, due to the remaining processing capacity of our microarray plates, each of which can handle 96 samples. The case (71 individuals who drank heavily; AUDIT score ≥ 8) and the control (126 residents who did not drink at all; AUDIT score = 0) participants were selected for a subsequent genome-wide comparison. Demographic parameters were documented and evaluated, including gender, age, BMI, coffee- and tea-consumption behavior, education levels and exercise habits (Table 1). Peripheral bloods were collected from these participants and centrifuged for separating the blood cells from plasma. Total DNA was extracted from the peripheral blood cells for subsequent genome-wide genotyping.

Table 1. Demographic information of subjects in this genome-wide association study.

	AUDIT Score ≥ 8	AUDIT Score = 0	P
Patient number	71	126	
Gender			
Male	65 (91.55%)	114 (90.48%)	0.802 *
Female	6 (8.45%)	12 (9.52%)	
Age, years, Mean \pm SD	53.72 \pm 9.93	56.97 \pm 12.53	0.062 ‡
BMI, kg/m ² , Mean \pm SD	25.6 \pm 4.0	25.8 \pm 3.7	0.643 ‡
<24,	25 (35.2%)	41 (32.5%)	0.597 *
24–27	24 (33.8%)	37 (29.4%)	
>27	22 (31.0%)	48 (38.1%)	
Coffee			
never	24 (33.8%)	36 (28.6%)	0.483 *
social	19 (26.8%)	44 (34.9%)	
more than sometimes	28 (39.4%)	46 (36.5%)	
Tea			
never	15 (21.1%)	36 (28.6%)	0.401 *
social	24 (33.8%)	44 (34.9%)	
more than sometimes	32 (45.1%)	46 (36.5%)	
Education level			
No education	1 (1.4%)	4 (3.2%)	0.609 ‡
Elementary school	8 (11.3%)	23 (18.3%)	
Junior high school	15 (21.1%)	19 (15.1%)	
Senior high school	34 (47.9%)	34 (27.0%)	
College	11 (15.5%)	38 (30.2%)	
Higher education	2 (2.8%)	8 (6.3%)	
Exercise			
little	23 (33.8%)	41 (31.9%)	0.699 *
Sometimes (≤ 2 days per week)	17 (25.0%)	26 (23.7%)	
Frequent (≥ 3 days per week)	28 (41.2%)	59 (44.4%)	

* χ^2 tests; ‡ Welsh *t*-tests.

2.2. Genome-Wide Genotype Calling and Assessment of Associations

The Affymetrix Axiom Genome-Wide TWB 2.0 array plates, designed jointly by the Taiwan Biobank, the National Center for Genome Medicine in Taiwan and Thermo Fisher Scientific (Waltham, MA, USA), were used for assessing the genome-wide variants. Genome-wide genotyping was performed by the National Center for Genome Medicine in Taiwan, where a rigorous quality-control pipeline has been installed to process the data before release. The Affymetrix APT base-calling software was used. Associations between genotypes and phenotypes were evaluated by using the Fisher's exact test, based on major and minor allele counts in case and control subjects. We used the Golden Helix SVS software v.8.8.3 for genome-wide comparisons of genotypes (Golden Helix, London, UK).

2.3. Obtaining Genotypes from a Public-Domain Database to Augment This Study

We searched the European Nucleotide Archive (<https://www.ebi.ac.uk/eva/?Home> accessed on 9 September 2021), an important public-domain database of genomic variants. We found the genome-wide genotypes of a participant cohort (ID: PRJEB42554, published 21 January 2021) which are useful for the augmentation of this study. This cohort has a decent sample size ($n = 2000$) with participants recruited in Guangzhou, a coastal city in Southern China. The genotypes were assayed by using SNP arrays. We downloaded the genotypes in the format of compressed VCF files, where the SNP locations have been annotated based on the GRCh37 reference genome. The numbers of participants with rs671 "GG", "GA" and "AA" genotypes are 937, 888 and 175, respectively. The high number of "AA" carriers was an important merit which motivated us to choose this dataset. The minor allele frequency is 30.95%.

2.4. Obtaining Allele-Carrying Haplotypes from Densely Genotyped Persons

Allele-carrying haplotypes in individuals with homozygous alleles at a variant site were first searched. Starting from the closest variant at both sides of the allele, the haplotype was extended at both the upstream and downstream directions, one at a time, as long as the genotype of the next variant was still homozygous and was identical to the genotype of all the other individuals (Supplementary Materials Figure S2). A core region of haplotype was found when the consecutive genotypes identical across all the individuals in both upstream and downstream directions were linked together. The disagreement of genotypes in different persons began to emerge at the boundary of the core region. The haplotypes were extended further as long as this individual who has the homozygous allele in the next variant, which was also the major allele in the remaining members. Otherwise, individuals with the other allele were designated to be no longer extending in this direction, and their haplotype boundary was found. When the number of people with the "extending" status was reduced to 2, and the subsequent variants of the two people differed, so that no major allele could be determined, the extending process in this direction stopped. Finally, the longest haplotype joining the left and right arms was found. This longest haplotype provided a template for phasing the individuals with the heterozygous allele. The algorithm for finding allele-carrying haplotypes in persons with the homozygous allele are presented in a flowchart in Supplementary Materials Figure S2.

2.5. Mutation Age Estimation Using Haplotype Length Distributions

The age of the ALDH2*2 allele founder was estimated by using the gamma method, utilizing the linkage-disequilibrium decay effect caused by prior recombination events [41–43]. This method is based on the premise that the allele-carrying haplotypes are copies of the ancient chromosome where the founder allele emerges. The haplotype initially encompasses the entire haploid chromosome, yet is gradually shortened by the subsequent recombination events, causing a linkage-disequilibrium decay [41,42]. Nevertheless, the remaining allele-carrying haplotypes still harbor hundreds of variants genotyped in a dense genotyping microarray, indicating identity-by-descent. The lengths offer a means of calculating

the ages. The shorter the haplotype is, the older the founder allele and allele-carrying haplotype are.

To estimate the time, the genomic locations of the haplotype boundaries were first cross-referenced with the genetic map offered by the international HapMap project for quantifying the haplotype lengths in the units of Morgans. The lengths would manifest a gamma distribution defined by a shape parameter and a scale parameter (α, τ). The shape parameter is set as 2 [42]. The number of generations (τ) is the scale parameter, which can be calculated by using the maximum-likelihood estimations with a bias-correction [42] defined in Equation (1):

$$\tau = \frac{2n - 1}{\sum_{i=1}^n l_i} \quad (1)$$

where the $l_1, l_2, l_3 \dots, l_n$ are the haplotype lengths in the unit of Morgans; n is the number of haplotypes investigated. The age of the founder mutation is then inferred from the number of generations, using the time per generation, which is assumed to be 25 years in this study:

$$\text{Age} = (\text{time per generation}) \times (\tau) \quad (2)$$

We wrote a Python script implementing the above calculations in Equations (1) and (2) with a bootstrapping method [44]. The script was executed in the Python 3.6 environment. Bootstrapping is a random sampling with replacement method that, for each step of sampling, n haplotypes were drawn from the pool of n haplotypes, regardless of whether the haplotype have or have not been drawn, for calculating one age estimate. The sampling process was repeated to produce more age estimates, creating a distribution of estimated ages.

3. Results

3.1. The Genomic Variant rs671 G > A Is the Leading Variant Associated with Alcohol Consumption Behavior

We assessed ~750,000 SNPs in 197 residents in Keelung, including 71 heavy drinkers (AUDIT score ≥ 8) and 126 non-drinkers (AUDIT score = 0). The two groups did not manifest disparity in terms of gender, age, body mass index (BMI), coffee- and tea-consumption behavior, education levels and exercise frequencies [41] (Table 1). Among all genomic variants, rs671 G > A in chromosome 12 manifested the strongest association ($P = 8.75 \times 10^{-9}$, Figure 1A). There are 103, 81 and 13 persons with the GG, GA and AA genotypes at rs671, respectively (Table 2). The minor allele A, also known as the ALDH2*2, is associated with non-drinkers, and the minor allele frequency (MAF) is 0.272. None of the heavy drinkers had the homozygous "AA" genotype. All the variants with p -values smaller than 5×10^{-7} are summarized in Table 2. The genotype counts of all these variants do not deviate from Hardy-Weinberg equilibrium. In Table 2, the variants in the vicinity of rs671 showed similar genotype counts, reflecting the linkage disequilibrium effects among them. A zoom-in Manhattan plot of this neighborhood showed a typical multiple-variant association in the presence of linkage disequilibrium (Figure 1B).

Table 2. Leading variants associated with alcohol consumption levels in this GWAS.

SNP	Chr	Gene	Allele	Case				Control		MAF	Fisher's Exact P	Hardy-Weinberg Equilibrium P	
				Homozygous Major	Heterozygous	Homozygous Minor	Homozygous Major	Heterozygous	Homozygous Minor			Case	Control
rs116369005	6	Intergenic	G/C	38	26	0	113	7	0	0.090	7.54×10^{-8}	0.125	0.947
rs671	12	ALDH2	G/A	56	15	0	47	66	13	0.272	8.75×10^{-9}	0.609	0.345
rs77768175	12	HECTD4	A/G	56	15	0	48	65	13	0.269	1.56×10^{-8}	0.609	0.416
rs78069066	12	MAPKAPK5/TMEM116	G/A	54	17	0	45	68	13	0.282	2.96×10^{-8}	0.519	0.225
rs3782886	12	BRAP	T/C	53	18	0	46	67	13	0.282	1.20×10^{-7}	0.473	0.282
rs11066280	12	HECTD4	T/A	52	18	1	43	70	13	0.294	3.05×10^{-7}	0.923	0.136
rs541300736	16	TEKT5	G/A	36	25	0	112	8	0	0.091	2.62×10^{-7}	0.132	0.931

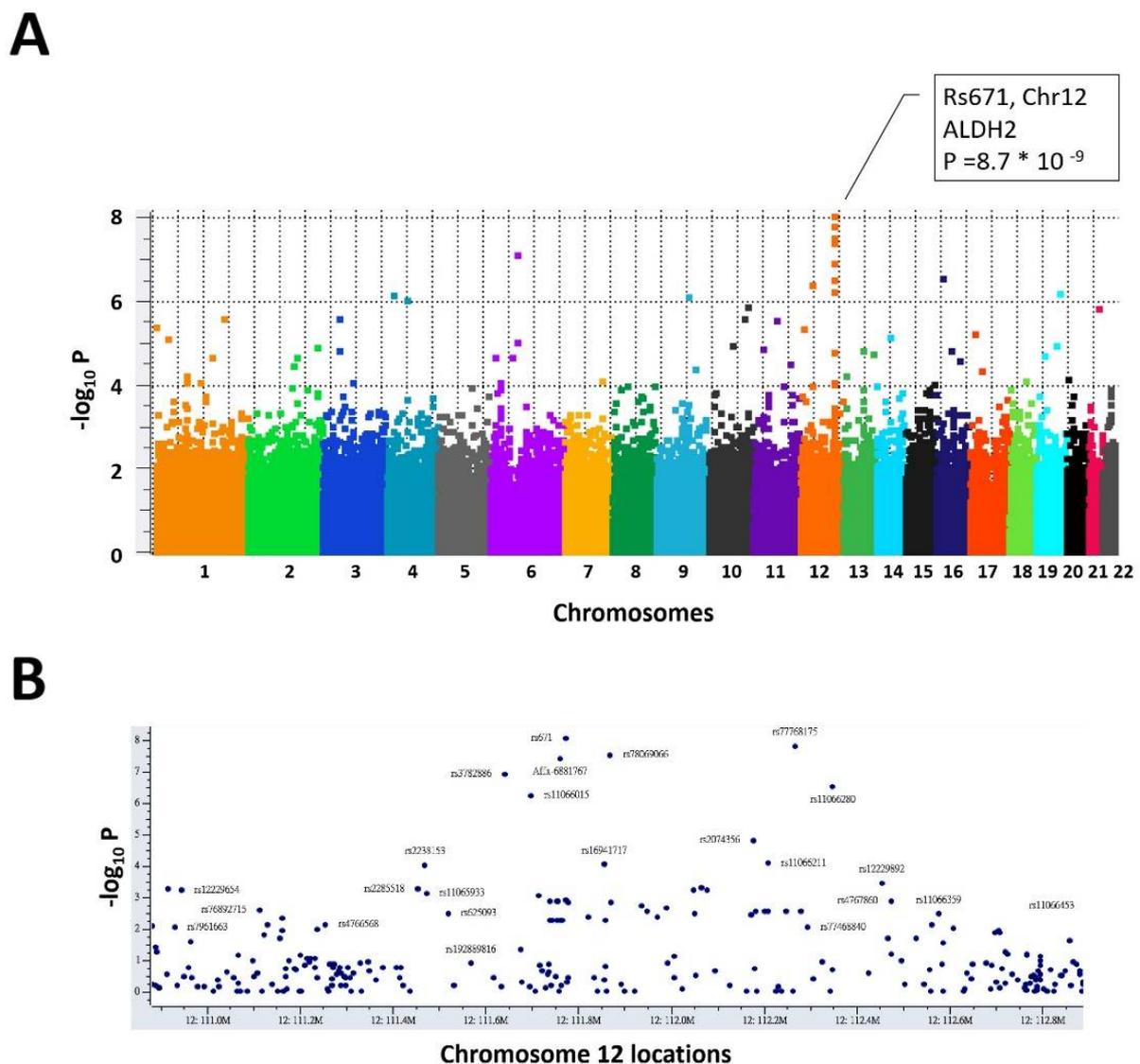


Figure 1. Genome-wide comparisons of residents in Keelung who drank hazardously and who did not drink in their daily life. **(A)** Manhattan plot of autosomal chromosomes. **(B)** Zoom-in plot of association in the vicinity of rs671 (2 million nucleotide bases; Chr12:110,912,851–112,927,816 in GRCh38).

3.2. A Haplotype Shared by All ALDH2*2-Carrying Individuals Indicates Its Identity by Descent

Identical haplotypes encompassing 138 consecutive, densely genotyped genomic variants, spanning a region of ~600,000 nucleotide bases surrounding rs671, were found, with no exceptions, in all 13 persons with homozygous ALDH2*2 (shown within the green rectangular box of Figure 2A). In this region, genotypes of all variants are consistently homozygous, showing clearly the existence of two identical haplotypes, each of which carries one ALDH2*2 allele. This region is referred to as the “core” region of the haplotype. The “core” haplotype can be extended further in some but not all 13 persons, and it can finally encompass a maximum of 439 consecutive genomic variants, spanning a genomic region of two million nucleotide bases (Chr12:110,912,851–112,927,816 in GRCh38, Figure 2A). The longest haplotype represents a replica of an ancient chromosomal segment where the “G” allele at this mutation site changed to “A”. According to the linkage-disequilibrium decay theory, the longer the haplotype is, the younger the allele is [41–43]. We performed the same analysis in 103 residents with the homozygous “GG” genotypes for a comparison. Much shorter shared haplotypes were observed (Figure 2B). This suggests an earlier emergence

of the “G” allele. These allele-carrying haplotypes in residents with rs671 “AA” genotype could be identical by descent, considering the hundreds of identical variant types shared by all people.

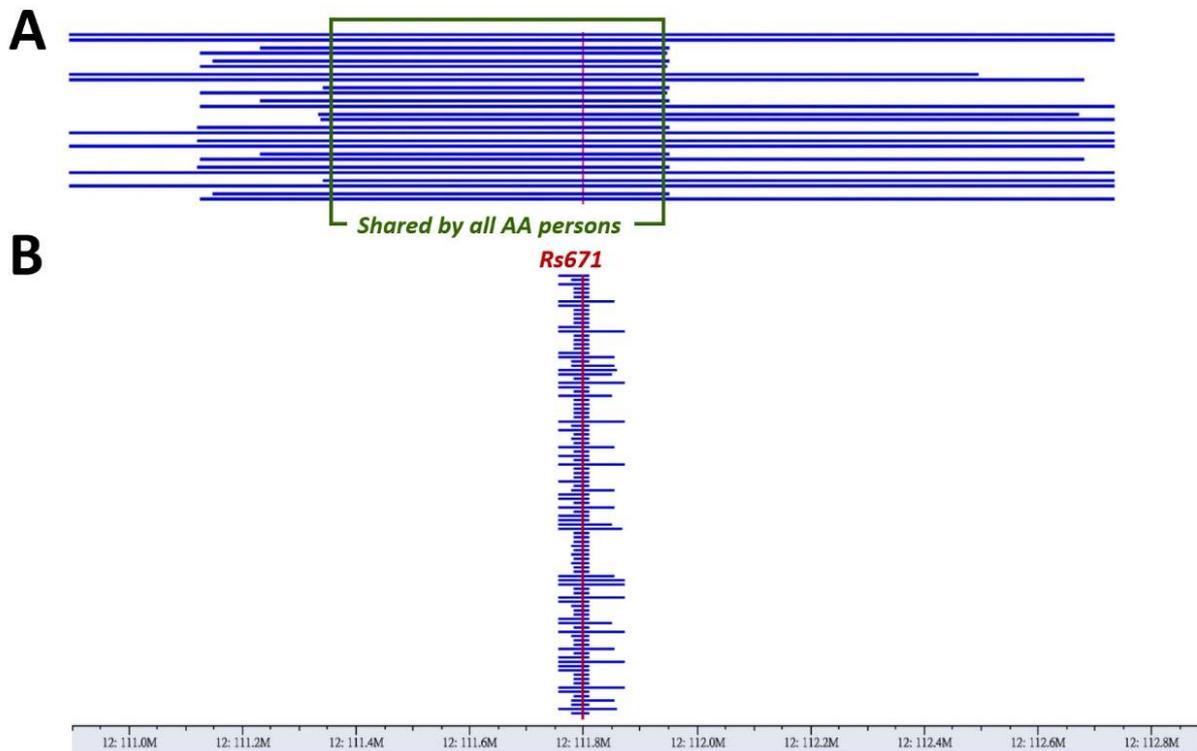


Figure 2. The haplotypes inferred in the vicinity of rs671 (Chr12:110,912,851–112,927,816), shown as blue horizontal lines. The vertical red lines depict the spot of rs671. **(A)** The rs671 “A”-allele-carrying haplotypes in persons with homozygous “AA” genotypes. Each person has two haplotypes shown as two consecutive blue horizontal lines (13 persons, 26 haplotypes). In each aligned genomic location of the blue horizontal lines, the nucleotide bases are identical, without exceptions. Within the green box, all persons have identical haplotypes encompassing 38 consecutive variants. **(B)** The rs671 “G”-allele-carrying haplotypes, calculated by using the same method as in **(A)**, in persons with homozygous “GG” genotypes.

The longest haplotype observed in residents with rs671 “AA” also provided a reference template for haplotype phasing of the 81 residents with the “GA” genotype. The median length of all the ALDH2*2-carrying haplotypes found in the Keelung cohort is 1.6 million nucleotide bases, corresponding to 0.361 centiMorgans in this chromosomal region. The minimum and maximum lengths are 0.6 and 2.0 million bases.

We also analyzed the genome-wide genotypes of the Guangzhou cohort, particularly the 175 participants with homozygous ALDH2*2 genotypes. Long ALDH2*2-carrying haplotypes were found in all 175 participants without exceptions. The median (min, max) lengths is 1.7 (0.5–2.8) million nucleotide bases. The fact that all of the homozygous ALDH2*2 participants in Keelung and Guangzhou have ALDH2*2-carrying haplotypes longer than 0.5 million bases supports the idea that this allele has one single origin. This allele is less likely to have arisen recurrently at different ancient chromosomes.

3.3. Age of the ALDH2*2 Allele Was Estimated to Be 7935 Years Ago

The left and right boundaries of the haplotypes represent the historical recombination sites. Based on the lengths of the ALDH2*2-carrying haplotypes between the two boundaries, the age of the most recent common ancestor can be inferred [42]. To ensure an adequate sample size for age estimation, participants with the homozygous ALDH2*2 allele in the Keelung and Guangzhou cohorts ($n = 13$ and 175 respectively) were included and combined together. We then transformed the physical lengths of these haplotypes into

genetic lengths in the units of centiMorgans (cM), utilizing the map of conversion offered by the international HapMap project. The length distribution of these ALDH2*2-carrying haplotypes is shown in Figure 3A. Subsequently, we made 100,000 bootstrap estimations of the age, and they are shown in Figure 3B. In this distribution, the median age is 7935 (min–max; 7014–9381) years ago.

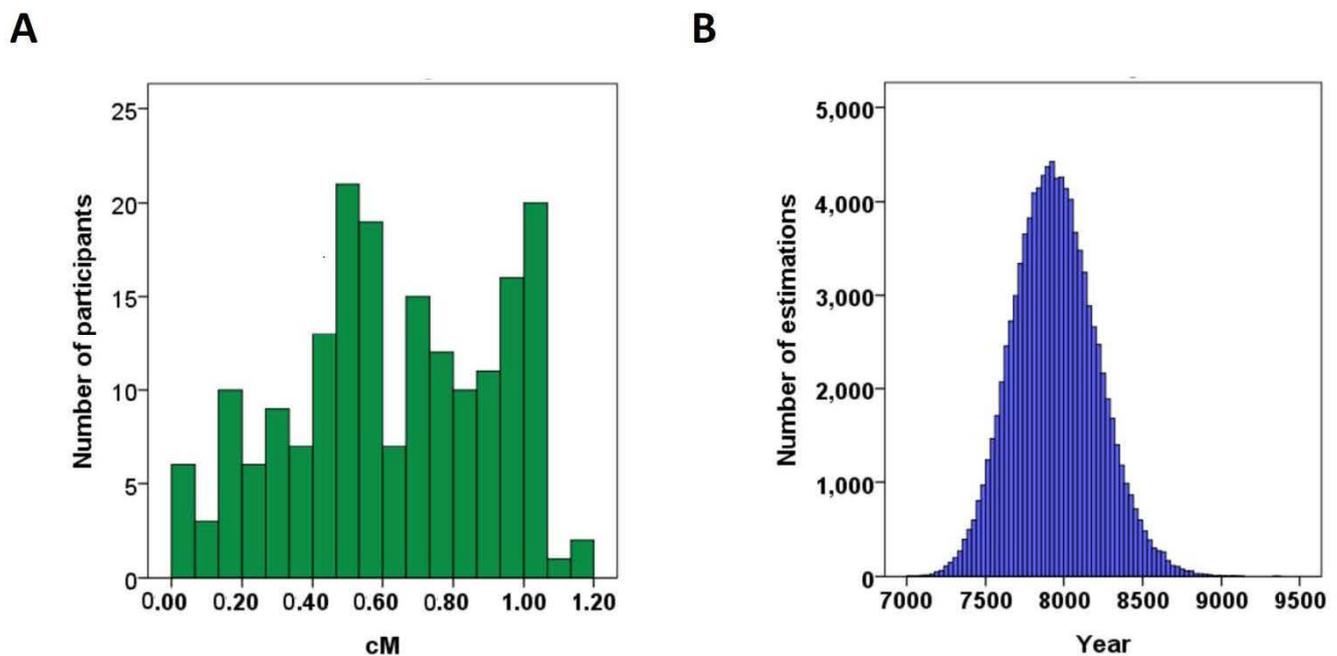


Figure 3. (A) Histogram of lengths of ALDH2*2 haplotypes in centiMorgans of the participants with homozygous ALDH2*2 in the Keelung and Guangzhou cohorts ($n = 183$). (B) Histogram of the estimated age (years) of ALDH2*2 allele by 100,000 bootstrapping estimations. The minimum, median and maximum of this distribution are 7014, 7935 and 9381 years, respectively. The arithmetic mean of the distribution is 7944 years ago.

4. Discussion

Young-age alleles tend to have small minor allele frequencies (e.g., <0.01) in a population, as time is required for them to populate [41]. Compared with common alleles, variants of such small MAFs are less likely to be detected in a typical GWAS, where variants are assessed in the same way with a fixed sample size regardless of their wide diversity in MAFs [45]. Overall, rs671 is a common variant in Asia. A recent study revealed the wide geographical distribution of this allele in East and Southeast Asia [46]. The distribution centered in the Fujian and Guangdong, two coastal provinces of China, where the MAF could be as high as 0.409 and 0.357, respectively [46]. Similar distributions were shown in one other study, where the region with the highest MAF was assumed the geographical origin of ALDH2*2 [34]. The authors further inferred the time origin of this allele by using the time of emergence of the “Pai-Yuei” ethnic people who have populated this geographical region since 2000–3000 years ago [34]. This value was subsequently cited in the literature [21,35–37]. However, the region with the highest MAF does not always co-localize with the geographical origin. Thus, the assumption behind the estimation of 2000–3000 years may be undermined. In human migratory history, a small subset from a large genetically diverse population may migrate to and then populate a new territory. The minor alleles enriched in the immigrants may cause a sudden increase of corresponding MAFs. This is called the founder effect. The wide geographical distribution and the high MAF of ALDH2*2 prompted us to ask whether this allele could have such a short history.

Thus, we investigated the allele-carrying haplotypes by using our densely genotyped data, which are critical for inferring the historical genomic recombination events surrounding the allele. As people migrated, they took the alleles and the haplotypes along. We

discovered an identity-by-descent allele-carrying haplotype spanning 138 consecutive, densely genotyped genomic variants, covering ~600,000 nucleotide bases in the vicinity of rs671 in the Keelung cohort. The median haplotype length is 1.6 (0.6–2.0) million nucleotide bases. We also utilized a public-domain dataset of Guangzhou, the capital of the Guangdong province where the ALDH2*2 frequency is high. Again, we found long ALDH2*2-carrying haplotypes in all 175 participants with the homozygous ALDH2*2 allele without exceptions. The median haplotype length is 1.7 (0.5–2.8) million nucleotide bases. In contrast, the previous study [34] only employed five local variants spanning a genomic region of 40,000 nucleotide bases, which would not be sufficient to offer high-resolution time estimates. We had two discoveries. First, an identity-by-descent relatively long ALDH2*2-carrying haplotype was found consistently in all investigated individuals with this allele. Second, the ALDH2*2 was ascertained to have arisen in a pre-historical, possibly Neolithic time of 7935 years ago, which matched the time of the earliest known alcohol-containing beverages in China [1]. What we have estimated is the most recent common ancestor of the included individuals with ALDH2*2. The birth of the allele might have happened at an even more ancient time. The ALDH2*2 allele was not observed in the residents of the Americas; thus, it is reasonable to infer that the allele did not emerge before humans emigrated from Asia to the Americas through the Bering sea ~16,500 years ago [47,48].

After the emergence of this allele, it stably persisted, rather than have come and gone, in Asian populations. It is still mysterious why this enzymatic-function disrupting allele was not purged by purifying selection. One possibility is that the enzyme dysfunction produced a face flushing and discomfort phenotype of the allele carriers whenever they drank, notifying them to stop drinking excessively. The consequences of this behavior counteracted purifying selections.

5. Conclusions

The SNP rs671 G > A is the leading genomic variant associated with the drinking behavior of people in Keelung, Taiwan. The origin of the ALDH2*2 allele dates back to 7935 years ago.

Supplementary Materials: The following are available online at <https://www.mdpi.com/article/10.3390/biom11091376/s1>. Supplementary Figure S1. The AUDIT score distribution of screened participants in this community study. Supplementary Figure S2. The algorithm used for exploring the allele carrying haplotypes in a group of people with the homozygous allele.

Author Contributions: C.-L.L. and C.-T.Y. designed and supervised the study; K.-H.L. drafted the manuscript; R.-N.C., L.-W.C., T.-S.H. and Y.-C.S. collected and analyzed the data; K.-H.L. performed statistical analysis. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Chang Gung Memorial Hospital Research Projects (CRRPG2H0044, CRRPG2H0082, CLRPG2L0051, CMRPG2J0011 and CMRPG2K0081). The funding body provided support in the design of the study; in the collection, analysis and interpretation of data; and in writing the manuscript.

Institutional Review Board Statement: This study was approved by the institutional review board of Chang Gung Memorial Hospital, Taiwan, and conducted according to the ethical guidelines of the 1975 Declaration of Helsinki.

Informed Consent Statement: Written informed consents were obtained from all participants in this community study. All of the study subjects were adults at the time of enrollment.

Data Availability Statement: The genome-wide genotypes of 197 study subjects will be provided as a CSV file by the co-corresponding authors to academic scientists on reasonable request.

Acknowledgments: The authors would like to thank Yi-Ting Liao, Mei Lu Tsai and Chien-Chih Wang for their excellent technical and administrative assistance. The genome-wide genotyping assay was performed and technically supported by the National Center for Genome Medicine, Taiwan, with outstanding quality, which is highly appreciated.

Conflicts of Interest: The authors have declared that no conflict of interest exists.

References

1. McGovern, P.E.; Zhang, J.; Tang, J.; Zhang, Z.; Hall, G.R.; Moreau, R.A.; Nunez, A.; Butrym, E.D.; Richards, M.P.; Wang, C.-S.; et al. Fermented beverages of pre- and proto-historic China. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 17593–17598. [[CrossRef](#)]
2. Hines, L.M.; Rimm, E.B. Moderate alcohol consumption and coronary heart disease: A review. *Postgrad. Med. J.* **2001**, *77*, 747–752. [[CrossRef](#)]
3. Goldwater, D.; Karlamangla, A.; Merkin, S.S.; Seeman, T. Compared to non-drinkers, individuals who drink alcohol have a more favorable multisystem physiologic risk score as measured by allostatic load. *PLoS ONE* **2019**, *14*, e0223168. [[CrossRef](#)]
4. Koppes, L.L.; Dekker, J.M.; Hendriks, H.F.; Bouter, L.; Heine, R.J. Moderate Alcohol Consumption Lowers the Risk of Type 2 Diabetes: A meta-analysis of prospective observational studies. *Diabetes Care* **2005**, *28*, 719–725. [[CrossRef](#)]
5. Mukamal, K.J.; Maclure, M.; Muller, J.E.; Sherwood, J.B.; Mittleman, M. Prior Alcohol Consumption and Mortality Following Acute Myocardial Infarction. *JAMA* **2001**, *285*, 1965–1970. [[CrossRef](#)]
6. Mukamal, K.J.; Chung, H.; Jenny, N.S.; Kuller, L.H.; Longstreth, W.T.; Mittleman, M.; Burke, G.L.; Cushman, M.; Psaty, B.M.; Siscovick, D.S. Alcohol Consumption and Risk of Coronary Heart Disease in Older Adults: The Cardiovascular Health Study. *J. Am. Geriatr. Soc.* **2005**, *54*, 30–37. [[CrossRef](#)]
7. Muntwyler, J.; Hennekens, C.H.; Buring, J.E.; Gaziano, J.M. Mortality and light to moderate alcohol consumption after myocardial infarction. *Lancet* **1998**, *352*, 1882–1885. [[CrossRef](#)]
8. Solomon, C.G.; Hu, F.B.; Stampfer, M.J.; Colditz, G.A.; Speizer, F.E.; Rimm, E.B.; Willett, W.C.; Manson, J.E. Moderate alcohol consumption and risk of coronary heart disease among women with type 2 diabetes mellitus. *Circulation* **2000**, *102*, 494–499. [[CrossRef](#)]
9. Esch, T.; Stefano, G.B. The neurobiology of pleasure, reward processes, addiction and their health implications. *Neuro Endocrinol. Lett.* **2004**, *25*, 235–251.
10. Mbarek, H.; Milaneschi, Y.; Fedko, I.O.; Hottenga, J.-J.; de Moor, M.; Jansen, R.; Gelernter, J.; Sherva, R.; Willemsen, G.; Boomsma, D.I.; et al. The genetics of alcohol dependence: Twin and SNP-based heritability, and genome-wide association study based on AUDIT scores. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.* **2015**, *168*, 739–748. [[CrossRef](#)]
11. Jorgenson, E.; Thai, K.K.; Hoffmann, T.J.; Sakoda, L.C.; Kvale, M.N.; Banda, Y.; Schaefer, C.; Risch, N.; Mertens, J.; Weisner, C.; et al. Genetic contributors to variation in alcohol consumption vary by race/ethnicity in a large multi-ethnic genome-wide association study. *Mol. Psychiatry* **2017**, *22*, 1359–1367. [[CrossRef](#)]
12. Sun, Y.; Chang, S.; Wang, F.; Sun, H.; Ni, Z.; Yue, W.; Zhou, H.; Gelernter, J.; Malison, R.T.; Kalayasiri, R.; et al. Genome-wide association study of alcohol dependence in male Han Chinese and cross-ethnic polygenic risk score comparison. *Transl. Psychiatry* **2019**, *9*, 249. [[CrossRef](#)]
13. Clarke, T.-K.; Adams, M.J.; Davies, G.; Howard, D.; Hall, L.S.; Padmanabhan, S.; Murray, A.; Smith, B.; Campbell, A.; Hayward, C.; et al. Genome-wide association study of alcohol consumption and genetic overlap with other health-related traits in UK Biobank (N = 112,117). *Mol. Psychiatry* **2017**, *22*, 1376–1384. [[CrossRef](#)]
14. Macgregor, S.; Lind, P.A.; Bucholz, K.K.; Hansell, N.K.; Madden, P.A.; Richter, M.M.; Montgomery, G.W.; Martin, N.G.; Heath, A.C.; Whitfield, J.B. Associations of ADH and ALDH2 gene variation with self report alcohol reactions, consumption and dependence: An integrated analysis. *Hum. Mol. Genet.* **2009**, *18*, 580–593. [[CrossRef](#)] [[PubMed](#)]
15. Quillen, E.; Chen, X.-D.; Almasry, L.; Yang, F.; He, H.; Li, X.; Wang, X.-Y.; Liu, T.-Q.; Hao, W.; Deng, H.-W.; et al. 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. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.* **2014**, *165*, 103–110. [[CrossRef](#)] [[PubMed](#)]
16. Sanchez-Roige, S.; Palmer, A.A.; Fontanillas, P.; Elson, S.L.; Adams, M.J.; Howard, D.; Edenberg, H.J.; Davies, G.; Crist, R.C.; Deary, I.J.; et al. Genome-Wide Association Study Meta-Analysis of the Alcohol Use Disorders Identification Test (AUDIT) in Two Population-Based Cohorts. *Am. J. Psychiatry* **2019**, *176*, 107–118. [[CrossRef](#)] [[PubMed](#)]
17. Kranzler, H.R.; Zhou, H.; Kember, R.; Smith, R.V.; Justice, A.C.; Damrauer, S.; Tsao, P.S.; Klarin, D.; Baras, A.; Reid, J.; et al. Author Correction: Genome-wide association study of alcohol consumption and use disorder in 274,424 individuals from multiple populations. *Nat. Commun.* **2019**, *10*, 4050. [[CrossRef](#)]
18. Budas, G.R.; Disatnik, M.-H.; Mochly-Rosen, D. Aldehyde Dehydrogenase 2 in Cardiac Protection: A New Therapeutic Target? *Trends Cardiovasc. Med.* **2009**, *19*, 158–164. [[CrossRef](#)]
19. Yang, H.; Song, Z.; Yang, G.-P.; Zhang, B.-K.; Chen, M.; Wu, T.; Guo, R. The ALDH2 rs671 Polymorphism Affects Post-Stroke Epilepsy Susceptibility and Plasma 4-HNE Levels. *PLoS ONE* **2014**, *9*, e109634. [[CrossRef](#)] [[PubMed](#)]
20. Hemani, G.; Bowden, J.; Smith, G.D. Evaluating the potential role of pleiotropy in Mendelian randomization studies. *Hum. Mol. Genet.* **2018**, *27*, R195–R208. [[CrossRef](#)]
21. Chen, C.-H.; Ferreira, J.C.B.; Gross, E.R.; Mochly-Rosen, D. Targeting Aldehyde Dehydrogenase 2: New Therapeutic Opportunities. *Physiol. Rev.* **2014**, *94*, 1–34. [[CrossRef](#)] [[PubMed](#)]
22. Peng, G.-S.; Chen, Y.-C.; Wang, M.-F.; Lai, C.-L.; Yin, S.-J. ALDH2*2 but not ADH1B*2 is a causative variant gene allele for Asian alcohol flushing after a low-dose challenge. *Pharm. Genom.* **2014**, *24*, 607–617. [[CrossRef](#)]

23. Kato, N.; Takeuchi, F.; Tabara, Y.; Kelly, T.N.; Go, M.J.; Sim, X.; Tay, W.T.; Chen, C.-H.; Zhang, Y.; Yamamoto, K.; et al. Meta-analysis of genome-wide association studies identifies common variants associated with blood pressure variation in east Asians. *Nat. Genet.* **2011**, *43*, 531–538. [[CrossRef](#)]
24. Mizuno, Y.; Hokimoto, S.; Harada, E.; Kinoshita, K.; Nakagawa, K.; Yoshimura, M.; Ogawa, H.; Yasue, H. Variant Aldehyde Dehydrogenase 2 (ALDH2*2) Is a Risk Factor for Coronary Spasm and ST-Segment Elevation Myocardial Infarction. *J. Am. Heart Assoc.* **2016**, *5*. [[CrossRef](#)]
25. Xia, C.-L.; Chu, P.; Liu, Y.-X.; Qu, X.-L.; Gao, X.-F.; Wang, Z.-M.; Dong, J.; Chen, S.-L.; Zhang, J.-X. ALDH2 rs671 polymorphism and the risk of heart failure with preserved ejection fraction (HFpEF) in patients with cardiovascular diseases. *J. Hum. Hypertens.* **2019**, *34*, 16–23. [[CrossRef](#)]
26. Chen, J.; Huang, W.; Cheng, C.-H.; Zhou, L.; Jiang, G.-B.; Hu, Y.-Y. Association Between Aldehyde dehydrogenase-2 Polymorphisms and Risk of Alzheimer’s Disease and Parkinson’s Disease: A Meta-Analysis Based on 5315 Individuals. *Front. Neurol.* **2019**, *10*, 290. [[CrossRef](#)]
27. Dakeishi, M.; Murata, K.; Sasaki, M.; Tamura, A.; Iwata, T. Association of alcohol dehydrogenase 2 and aldehyde dehydrogenase 2 genotypes with fasting plasma glucose levels in Japanese male and female workers. *Alcohol Alcohol.* **2008**, *43*, 143–147. [[CrossRef](#)] [[PubMed](#)]
28. Sakiyama, M.; Matsuo, H.; Nakaoka, H.; Yamamoto, K.; Nakayama, A.; Nakamura, T.; Kawai, S.; Okada, R.; Ooyama, H.; Shimizu, T.; et al. Identification of rs671, a common variant of ALDH2, as a gout susceptibility locus. *Sci. Rep.* **2016**, *6*, 25360. [[CrossRef](#)] [[PubMed](#)]
29. Takeshima, K.; Nishiwaki, Y.; Suda, Y.; Niki, Y.; Sato, Y.; Kobayashi, T.; Miyamoto, K.; Uchida, H.; Inokuchi, W.; Tsuji, T.; et al. A missense single nucleotide polymorphism in the ALDH2 gene, rs671, is associated with hip fracture. *Sci. Rep.* **2017**, *7*, 428. [[CrossRef](#)] [[PubMed](#)]
30. Dong, Q.; Ren, G.; Zhang, K.; Liu, D.; Dou, Q.; Hao, D. Genetic polymorphisms of ALDH2 are associated with lumbar disc herniation in a Chinese Han population. *Sci. Rep.* **2018**, *8*, 10379. [[CrossRef](#)]
31. Park, B.; Kim, J.-H.; Lee, E.-G.; Jung, S.-Y.; Lee, S.-Y.; Kang, H.-S.; Han, J.H. Role of aldehyde dehydrogenases, alcohol dehydrogenase 1B genotype, alcohol consumption, and their combination in breast cancer in East-Asian women. *Sci. Rep.* **2020**, *10*, 6564. [[CrossRef](#)] [[PubMed](#)]
32. Zhao, T.; Wang, C.; Shen, L.; Gu, D.; Xu, Z.; Zhang, X.; Xu, Y.; Chen, J. Clinical significance of ALDH2 rs671 polymorphism in esophageal cancer: Evidence from 31 case-control studies. *OncoTargets Ther.* **2015**, *8*, 649–659. [[CrossRef](#)]
33. Chien, J.; Liu, J.; Lee, M.; Jen, C.; Batrla-Utermann, R.; Lu, S.; Wang, L.; You, S.; Yang, H.; Chen, C. Risk and Predictors of Hepatocellular Carcinoma for Chronic Hepatitis b Patients with Newly Developed Cirrhosis. *J. Gastroenterol. Hepatol.* **2016**, *31*, 1971–1977. [[CrossRef](#)]
34. Luo, H.-R.; Wu, G.-S.; Pakstis, A.J.; Tong, L.; Oota, H.; Kidd, K.K.; Zhang, Y.-P. Origin and dispersal of atypical aldehyde dehydrogenase ALDH2*487Lys. *Gene* **2009**, *435*, 96–103. [[CrossRef](#)] [[PubMed](#)]
35. Yan, S.; Wu, G. Could ALDH2*2 be the reason for low incidence and mortality of ovarian cancer for East Asia women? *Oncotarget* **2017**, *9*, 12503–12512. [[CrossRef](#)]
36. Chang, J.S.; Hsiao, J.-R.; Chen, C.-H. ALDH2 polymorphism and alcohol-related cancers in Asians: A public health perspective. *J. Biomed. Sci.* **2017**, *24*, 19. [[CrossRef](#)] [[PubMed](#)]
37. Chen, C.-H.; Ferreira, J.C.B.; Joshi, A.U.; Stevens, M.C.; Li, S.-J.; Hsu, J.H.-M.; Maclean, R.; Ferreira, N.D.; Cervantes, P.R.; Martinez, D.D.; et al. Novel and prevalent non-East Asian ALDH2 variants; Implications for global susceptibility to aldehydes’ toxicity. *EBioMedicine* **2020**, *55*, 102753. [[CrossRef](#)]
38. Saunders, J.B.; Aasland, O.G.; Babor, T.F.; De La Fuente, J.R.; Grant, M. Development of the Alcohol Use Disorders Identification Test (AUDIT): WHO Collaborative Project on Early Detection of Persons with Harmful Alcohol Consumption-II. *Addiction* **1993**, *88*, 791–804. [[CrossRef](#)] [[PubMed](#)]
39. Tsai, M.-C.; Tsai, Y.-F.; Chen, C.-Y.; Liu, C.-Y. Alcohol Use Disorders Identification Test (AUDIT): Establishment of Cut-off Scores in a Hospitalized Chinese Population. *Alcohol. Clin. Exp. Res.* **2005**, *29*, 53–57. [[CrossRef](#)] [[PubMed](#)]
40. Higgins-Biddle, J.C.; Babor, T.F. A review of the Alcohol Use Disorders Identification Test (AUDIT), AUDIT-C, and USAUDIT for screening in the United States: Past issues and future directions. *Am. J. Drug Alcohol. Abus.* **2018**, *44*, 578–586. [[CrossRef](#)]
41. Slatkin, M.; Rannala, B. Estimating Allele Age. *Ann. Rev. Genom. Human Genet.* **2020**, *1*, 225–249. [[CrossRef](#)] [[PubMed](#)]
42. Gandolfo, L.C.; Bahlo, M.; Speed, T.P. Dating Rare Mutations from Small Samples with Dense Marker Data. *Genetics* **2014**, *197*, 1315–1327. [[CrossRef](#)]
43. Liang, K.-H.; Lu, Y.-H.; Niu, C.-W.; Chang, S.-K.; Chen, Y.-R.; Cheng, C.-Y.; Hsu, T.; Yang, C.-F.; Nakamura, K.; Niu, D.-M. The Fabry disease-causing mutation, GLA IVS4+919G>A, originated in Mainland China more than 800 years ago. *J. Hum. Genet.* **2020**, *65*, 619–625. [[CrossRef](#)]
44. Efron, B. *Bootstrap Methods: Another Look at the Jackknife*; Springer Series in Statistics; Springer: Berlin/Heidelberg, Germany, 1992; pp. 569–593.
45. Hong, E.P.; Park, J.W. Sample Size and Statistical Power Calculation in Genetic Association Studies. *Genom. Inf.* **2012**, *10*, 117–122. [[CrossRef](#)]

46. Li, H.; Borinskaya, S.; Yoshimura, K.; Kal'Ina, N.; Marusin, A.; Stepanov, V.; Qin, Z.; Khaliq, S.; Lee, M.-Y.; Yang, Y.; et al. Refined Geographic Distribution of the OrientalALDH2*504Lys(nce487Lys) Variant. *Ann. Hum. Genet.* **2009**, *73*, 335–345. [[CrossRef](#)] [[PubMed](#)]
47. Goebel, T.; Waters, M.R.; O'Rourke, D.H. The Late Pleistocene Dispersal of Modern Humans in the Americas. *Science* **2008**, *319*, 1497–1502. [[CrossRef](#)]
48. Gruhn, R. Evidence grows that peopling of the Americas began more than 20,000 years ago. *Nat. Cell Biol.* **2020**, *584*, 47–48. [[CrossRef](#)]