doi: 10.1046/j.1529-8817.2003.00060.x

The evolution and population genetics of the *ALDH2* locus: random genetic drift, selection, and low levels of recombination

Hiroki Oota[1], Andrew J. Pakstis[1], Batsheva Bonne-Tamir[2], David Goldman[3], Elena Grigorenko[4], Sylvester L. B. Kajuna[5], Nganyirwa J. Karoma[5], Selemani Kungulilo[6], Ru-Band Lu[7], Kunle Odunsi[8], Friday Okonofua[9], Olga V. Zhukova[10], Judith R. Kidd[1] and Kenneth K. Kidd[1,*]

[1]*Department of Genetics, Yale University School of Medicine, 333 Cedar Street, P.O. Box 208005, New Haven, CT 06520-8005, USA*
[2]*Department of Human Genetics, Sackler School of Medicine, Tel Aviv University, Tel Aviv, Israel*
[3]*Laboratory of Neurogenetics, National Institute of Alcohol Abuse and Alcoholism, Rockville, MD 20852, USA*
[4]*Department of Psychology, Yale University, New Haven, CT 06520, USA*
[5]*The Hubert Kairuki Memorial University, Dar es Salaam, Tanzania*
[6]*Muhimbili University College of Health Sciences, Dar es Salaam, Tanzania*
[7]*Department of Psychiatry, Tri-Service General hospital, National Defense Medical Center, Taipei, Taiwan, R.O.C.*
[8]*Department of Gynecological Oncology, Roswell Park Cancer Institute, Buffalo, NY 14263, USA*
[9]*Department of Obstetrics and Gynecology, Faculty of Medicine, University of Benin, Benin City, Nigeria*
[10]*N.I. Vavilov Institute of General Genetics RAS, Moscow, Russia*

Summary

The catalytic deficiency of human aldehyde dehydrogenase 2 (ALDH2) is caused by a nucleotide substitution (G1510A; Glu487Lys) in exon 12 of the *ALDH2* locus. This SNP, and four non-coding SNPs, including one in the promoter, span 40 kb of *ALDH2*; these and one downstream STRP have been tested in 37 worldwide populations. Only four major SNP-defined haplotypes account for almost all chromosomes in all populations. A fifth haplotype harbours the functional variant and is only found in East Asians. Though the SNPs showed virtually no historic recombination, LD values are quite variable because of varying haplotype frequencies, demonstrating that LD is a statistical abstraction and not a fundamental aspect of the genome, and is not a function solely of recombination. Among populations, different sets of tagging SNPs, sometimes not overlapping, can be required to identify the common haplotypes. Thus, solely because haplotype frequencies vary, there is no common minimum set of tagging SNPs globally applicable. The F_{st} values of the promoter region SNP and the functional SNP were about two S.D. above the mean for a reference distribution of 117 autosomal biallelic markers. These high F_{st} values may indicate selection has operated at these or very tightly linked sites.

Introduction

Ethanol oxidization to acetaldehyde is catalyzed by alcohol dehydrogenase (ADH), and acetaldehyde is metabolized to acetate by aldehyde dehydrogenase (ALDH). Sixteen human ALDH genes have been identified, and the catalytic activities are known for 11 of the ALDH enzymes (Vasiliou & Pappa, 2000). Aldehyde dehydrogenase 2 (ALDH2 [MIM 100650]) is a mitochondrial enzyme present primarily in adult liver, kidney, muscle and heart (Stewart *et al*. 1996). Of the 16 ALDH gene products, ALDH2 has the highest affinity for acetaldehyde (Km $< 5\ \mu$M), and so is considered to be the main enzyme in acetaldehyde oxidization related to alcohol

*Address for correspondence and reprints: Dr. Kenneth K. Kidd, Yale University, SHM I-351, 333 Cedar street, New Haven, CT 06520. Tel: (203) 785 2654; Fax: (203) 785 6568. E-mail: kidd@biomed.med.yale.edu

© University College London 2004

metabolism (Goedde *et al*. 1979). The *ALDH2* gene encoding this enzyme is 44 kb long, located on the long arm of chromosome 12 (12q24.2 [GenBank accession number NT_009775]), and is composed of 13 exons encoding 517 amino acid residues (Hsu *et al*. 1988). The catalytic-deficient variant, which is associated with facial flushing in East Asians upon alcohol intake (Harada *et al*. 1981), was originally characterized as a protein polymorphism (Harada *et al*. 1980) and then as a DNA polymorphism (Yoshida *et al*. 1985): a G to A nucleotide substitution in exon 12 at mRNA nucleotide position (np) 1510 causes a Glu to Lys amino acid substitution at amino acid position 487. The G to A substitution that generates the deficient variant *ALDH2*487Lys* (previous symbol: *ALDH2*2*) has been reported in East Asian populations at high frequencies (as high as 30.0%), but has not been seen in other populations studied (Shibuya & Yoshida, 1988; Peterson *et al*. 1999a).

Seven human ADH genes - Class I (*ADH1A*, *ADH1B*, *ADH1C*), Class II (*ADH4*), Class III (*ADH5*), Class V (*ADH6*), and Class IV (*ADH7*) genes - have been identified and exist in a cluster extending >360 kb on the long arm of chromosome 4 (4q21). All ADH genes show tissue-specific expression patterns (Bilanchone *et al*. 1986) and different ethanol catalytic efficiencies (Edenberg & Bosron, 1997). Two Class I ADH genes (*ADH1B* and *ADH1C*) are primarily expressed in adult liver. A high activity variant encoded by *ADH1B*47His* (previous symbol: *ADH2*2*) is present at high frequency in East Asia (more than 59.0%) (Goedde *et al*. 1992; Osier *et al*. 2002a). Interestingly both variants, *ALDH2*487Lys* and *ADH1B*47His*, have functional differences from the "normal" that increase the transient level of acetaldehyde *in vivo* for *ALDH2*487Lys* and *in vitro* for *ADH1B*47His*. The high level of acetaldehyde, which is definitely toxic, causes facial flushing (Harada *et al*. 1981) among other symptoms, and results in a protective effect against alcoholism (Harada *et al*. 1982; Goldman & Enoch, 1990).

Haplotype and linkage disequilibrium (LD) analyses have been widely applied to disease gene mapping and understanding human population history (Castiglione *et al*. 1995; Jorde, 1995; Tishkoff *et al*. 1996, 1998; Laan & Pääbo, 1997; Kidd *et al*. 1998, 2000; Reich *et al*. 2001; DeMille *et al*. 2002). Most studies of haplotypes and LD in various loci have shown that African populations have more haplotypes and lower levels of LD than non-African populations; this is best explained by a founder effect in those modern humans (Tishkoff *et al*. 1996, 1998; Castiglione *et al*. 1995; Kidd *et al*. 1998, 2000; Reich *et al*. 2001; DeMille *et al*. 2002) who emerged from Africa around 100,000 years ago, known as the "out-of Africa" theory of human dispersal (Cann *et al*. 1987; Vigilant *et al*. 1991; Hammer 1995; Tishkoff *et al*. 1996). However, a global survey of haplotype frequencies and LD for the *ADH* gene cluster has shown an unusual global pattern of haplotypes and strong LD around the world, with only four major haplotypes in African as well as non-African populations (Osier *et al*. 2002a). Likewise, a previous study of *ALDH2* found that only three major haplotypes are common in all examined populations, including one African population (Biaka), with a fourth East Asian-specific haplotype distinguished by the deficiency variant (Peterson *et al*. 1999a). Moreover, in that study LD at *ALDH2* did not differ in populations from different regions (Peterson *et al*. 1999b). Hence, the previous haplotype analyses suggest that both the *ALDH2* gene and the *ADH* gene clusters depart from the haplotype frequency pattern and the LD patterns predicted by the out-of-Africa theory.

More recent haplotype-based studies have suggested that the human genome can be separated into haplotype blocks that show little evidence of substantial recombination in human history (Jeffreys *et al*. 2001; Daly *et al*. 2001; Patil *et al*. 2001; Gabriel *et al*. 2002). Gabriel *et al*. (2002) estimate that half of the human genome is organized in blocks of >22 kb and >44 kb in African and European/Asian samples, respectively, and propose "haplotype tag SNPs (tagging SNPs)," an approach to detecting haplotypes using the minimum number of SNPs. However, both the idea of haplotype blocks and the tagging-SNPs approach to detect disease genes are still controversial (Clark *et al*. 1998; Templeton *et al*. 2000; Wang *et al*. 2002). The size of the *ALDH2* locus (44 kb) could qualify it for pilot evaluation of the approach of tagging SNPs.

Various population samples from different regions of the world are required to study the evolutionary history of haplotypes at a locus. However, LD and haplotypes of the *ALDH2* locus have not been well studied except in East Asian populations, because it is well known that the functional variant is present only in East Asians.

 © University College London 2004

To rectify that limitation we have examined 37 population samples: eight African, two Southwest Asian, eight European, two Northwest Asian, seven East Asian, one Siberian, two Pacific, four North American, and three South American. Five single nucleotide polymorphism (SNP) sites span the *ALDH2* locus uniformly to eliminate bias of marker density, and one short tandem repeat polymorphism (STRP) downstream of the gene was typed to explore the extent of significant LD. This comprehensive study of haplotype frequency and LD of *ALDH2* illuminates the global evolutionary history of the *ALDH2* gene. The data also allow an examination of the usefulness of the tagging SNPs approach at this locus.

Material and Methods

Population Samples

We typed six markers in 1965 individuals from 37 world-wide human populations: seven African (Chagga, Biaka, Mbuti, Yoruba, Ibo, Hausa, Ethiopian Jews), one African American, two Southwest Asian (Yemenite Jews, Druze), eight European (Adygei, Chuvash, Russians, Ashkenazi Jews, Finns, Danes, Irish, European Americans), two from Northwest Asia (Komi Zyriane, Khanty), seven East Asian (Chinese from San Francisco, Taiwan Han Chinese, Hakka, Japanese, Ami, Atayal, Cambodians), one Siberian (Yakut), two Pacific Island (Nasioi, Micronesians), four North American (Cheyenne, Pima from Arizona, Pima from Mexico, Maya), and three South American (Ticuna, Rondonia Surui, Karitiana). Sample sizes ranged from 23 (Nasioi) to 116 (Irish) with most having close to 50 individuals. These populations were classified by the geographic region of current or recent origin.

The detailed information on the individual populations and samples is in ALFRED (the ALelle FREquency Database) (Osier *et al*. 2001, 2002b). All individuals were apparently healthy volunteers with no diagnoses of alcoholism or related disorders performed, except in the Taiwan Han Chinese, Ami, and Atayal, as described by Osier *et al*. (1999). All samples have been collected with appropriate informed consent and IRB approval. We studied the samples anonymously.

DNA samples were extracted from lymphoblastoid cell lines that have been established and/or maintained in the laboratory of J.R.K and K.K.K. at Yale University. The methods of transformation, cell culture, and DNA purification have been described elsewhere (Anderson & Gusella, 1984; Sambrook *et al*. 1989; Kidd *et al*. 1991; Chang *et al*. 1996).

SNP and STRP typing

We selected five SNP sites to span the *ALDH2* locus uniformly, and a STRP site located 80 kb downstream of the *ALDH2* locus (Figure 1). The four non-coding SNPs, *Sac*I, *Hae*IIIc, *Rsa*I, and *Hae*IIIa sites, were typed as PCR-based RFLPs (restriction fragment length polymorphisms). The functional variant, the Glu487Lys (G1510A) site, was typed by the fluorescence polarization (FP) method (Chen et al. 1999). The STRP, *D12S1344*, was typed on an ABI PRISM 377 DNA sequencer with fragment size analysis using the program GENOTYPER.

All markers examined in this study have been reported in previous studies (Yoshida *et al*. 1984; Peterson *et al*. 1999a; Harada *et al*. 1999; Chou *et al*. 1999; Koch *et al*. 2000). The PCR primers for the *Sac*I, *Rsa*I, *Hae*IIIa, and *D12S1344* sites were modified from those previously reported (Peterson *et al*. 1999a; Koch *et al*. 2000). We also designed new primers for the SNP in intron 1 (*Hae*IIIc) originally reported in a database of Japanese Single Nucleotide Polymorphisms (JSNP). For the Glu487Lys site, we designed PCR primers appropriate to the FP method. The program "mfold" (SantaLucia, 1998) predicted a secondary structure that would likely inhibit the primer extension reaction. Therefore, we introduced an artificial mismatch in the downstream primer to disrupt the secondary structure. The upstream PCR primer was used as a detection primer for the single nucleotide base extension (SBE), giving very tight homo- and hetero- zygote genotype clusters.

All PCR conditions were optimized using gradient PCR in 96-well plates, and typing done in 384-well plates (total volume: 10μl). The genomic DNA as well as PCR and restriction enzyme reaction mixtures were dispensed by a Biomek 2000 Laboratory Automation Workstation (BECKMAN), and the reactions were carried out on a PTC-225 Peltier Thermal Cycler (MJ Research). The PCR products were digested with the

© University College London 2004

appropriate enzyme following the manufacturers' protocols. The digestion patterns at *Hae*IIIc, *Rsa*I, and *Hae*IIIa sites were detected using 2% regular agarose gels, whereas the digest patterns at the *Sac*I site were detected using 4% NuSieve GTG agarose gels. The FP genotyping for the Glu487Lys site was read on a LJL BioSystem Analyst. The detailed typing protocols including the primer information are available in ALFRED. We repeated the typing of failed or unclear typings until the proportion of typed individuals was >95% in each population.

Several DNA samples from our laboratory were previously examined by Peterson *et al*. (1999a). We replicated these typings in our laboratory and our results show reproducibility for all duplicated samples.

Ancestral-type Inference

We sequenced the regions, including the five SNPs and the STRP, for non-human primates - a common chimpanzee (*Pan troglodytes*), two gorillas (*Gorilla gorilla*), and two orangutans (*Pongo pygmaeus*) - to infer the ancestral state of the polymorphisms, using the typing primers for PCR and sequencing. The PCR products were purified by QIAquick PCR Purification Kit (QIAGEN); sequencing was done using ABI PRISM BigDye Terminator cycle sequencing and the ABI PRISM 377 DNA sequencer.

Statistical Analyses

Genotype and allele frequencies at the individual sites were determined by gene counting, assuming co-dominant inheritance. Agreement with Hardy-Weinberg ratios was tested at the separate sites in each sample by means of an auxiliary program, FENGEN, which also creates the input file for the program HAPLO (Hawley & Kidd, 1995) from raw data records.

Wright's F_{st} (Wright, 1969) was calculated by the program DISTANCE (Kidd & Cavalli-Sforza, 1974). We used 32 out of the 37 standard populations for the F_{st} calculation to compare with 117 reference sites on the other chromosomes that have been examined in our laboratory in the same 32 population samples.

Maximum likelihood estimates of haplotype frequencies and the standard errors (jackknife method) were calculated from the individual multi-site phenotypes of individuals in each population using the program HAPLO (Hawley & Kidd, 1995), which implements the EM algorithm (Dempster *et al*. 1977). Overall and pairwise measures of linkage disequilibrium were evaluated using the ξ coefficient by the HAPLO/P program (Zhao *et al*. 1999). Pairwise linkage disequilibrium values, D' (Lewontin, 1964) and Δ^2 (Devlin & Risch, 1995), were calculated by the program LINKD (Kidd *et al*. 2000).

Results

Map and Site Description

Figure 1 shows the location, sequences, and inferred ancestral states for the polymorphic sites. The four non-coding SNPs, *Sac*I, *Hae*IIIc, *Rsa*I, and *Hae*IIIa sites, are located approximately 40, 30, 20 and 10 kb upstream of the Glu487Lys site in exon 12 whereas the STRP, *D12S1344*, is 83 kb downstream of this functional variant.

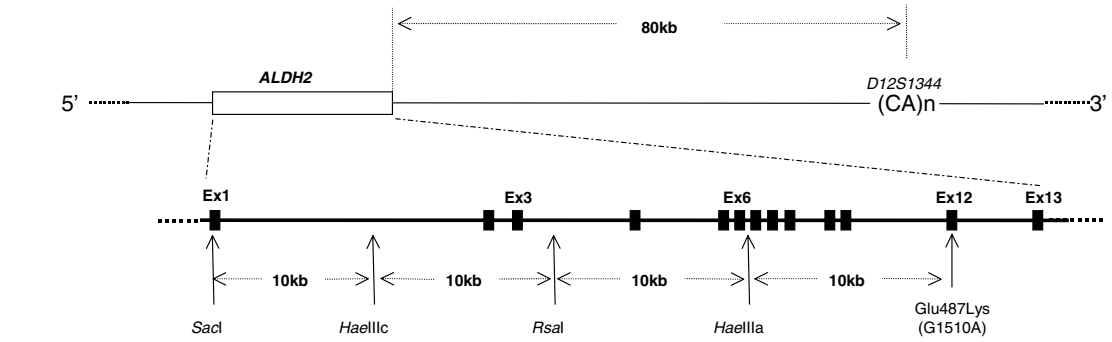
We designate the site-absent and the site-present alleles as "1" and "2," respectively, for the non-coding SNPs, and "G" and "A" for the bases of the Glu and Lys alleles, respectively, at the functional variant. For the *D12S1344* alleles, we designate the sizes of the alleles as called by GENOTYPER as the names of the alleles. For the haplotype, we use these designations from the 5′ to 3′ ends in order. For example, a 5-SNP haplotype described as 1111G indicates that all non-coding SNPs are the restriction site-absent alleles and the functional variant is the G allele.

The common chimpanzee, gorilla, and orangutan sequences at the sites of the human SNPs were consistent (Figure 1) and following the logic in Iyengar *et al*. (1998) provide unambiguous indication of which allele is ancestral: G (site present) at the *Sac*I site, C (site present) at the *Hae*IIIc site, C (site absent) at the *Rsa*I site, T (site absent) at the *Hae*IIIa site, and G (Glutamic acid) at the exon 12 site.

Individual-Site Results

All individual-site allele and haplotype frequencies for all populations are given in ALFRED under the locus and site UIDs. Figure 2 graphs the frequencies of the ancestral alleles for the five SNPs in 37 populations.

 © University College London 2004



Human	G**G**/**A**AGCTC	GGC**C**/**T**	G**C**/**T**AC	GG**T**/**C**C	CACT**G**/**A**AAG
Chimp	GGAGCTC	GGCC	GCAC	GGTC	CACTGAAG
Gorilla	GGAGCTC	GGCC	GCAC	GGTC	CACTGAAG
Orangutan	GGAGCTC	GGCC	GCAC	GGTC	CACTGAAG
Ancestor	G (2)	C (2)	C (1)	T (1)	G
ALFRED UID	SI000732M	SI000746R	SI000733N	SI000717P	SI000734O

Figure 1 Relative map for five SNPs and one STRP examined. All non-coding SNPs are named by restriction enzymes, whereas the coding SNP is named by the amino acid change and the position. The enzyme recognition sequences are shown below the map with non-human primates sequence. The inferred ancestral states are G at the *Sac*I site ("site-present" is represented as "2"); C at the *Hae*IIIc site, "2; " C at the *Rsa*I site ("site-absent" is represented as "1"); T at the *Hae*IIIa site, "1; " and G at the Glu487Lys (G1510A) site.

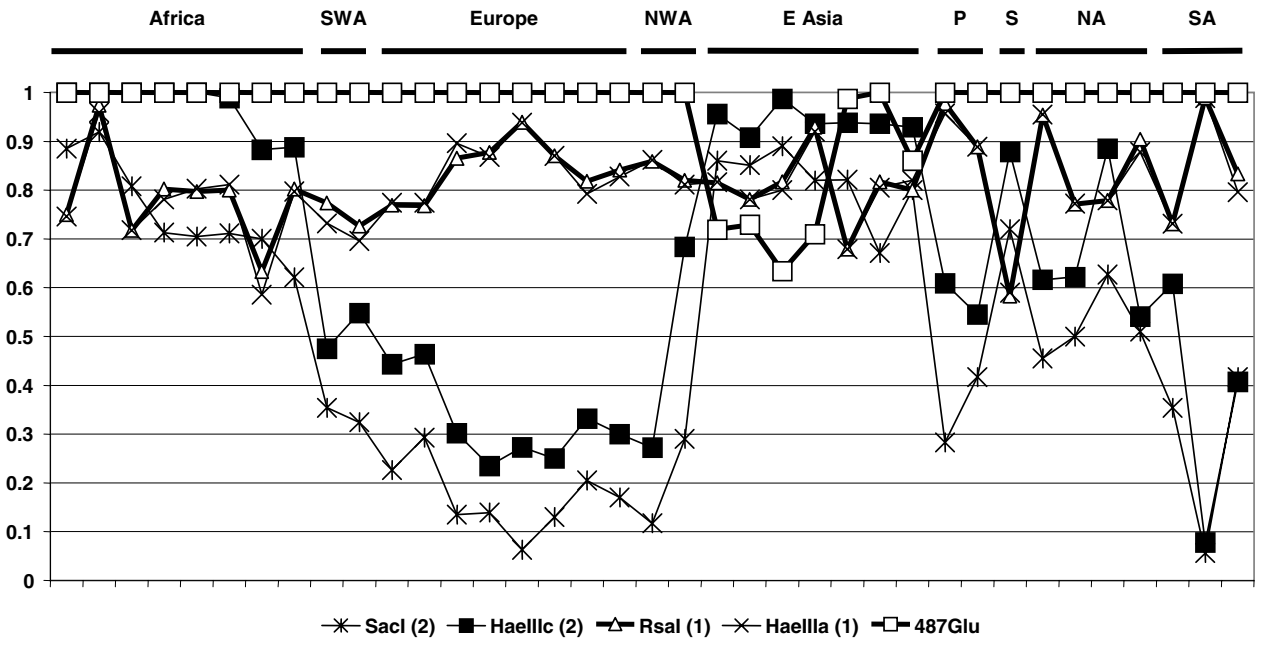


Figure 2 *ALDH2* 5-SNP ancestral allele frequencies. Geographic regions are classified as described in "Methods" by abbreviations: "SWA" for Southwest Asia, "NWA" for Northwest Asia, "E Asia" for East Asia, "P" for Pacific, "S" for Siberia, "NA" for North America, "SA" for South America. The population order on the X axis from left to right is as follows: Africa (Biaka, Mbuti, Yorba, Ibo, Hausa, Chagga, Ethiopia, African American), SW Asia (Yemenities, Druze), Europe (Adgygei, Chuvash, Russians, Ashenazi, Finns, Danes, Irish, European American), NW Asia (Komi Zyriane, Khanty), E Asian (San Francisco Chinese, Taiwan Han Chinese, Hakka, Japanese, Ami, Atayal, Cambodians), Pacific (Nasioi, Micronesians), Siberia (Yakut), N America (Cheyenne, Arizona Pima, Mexico Pima, Maya), S America (Ticuna, Rondonia Surui, Karitiana).

© University College London 2004

Table 1 Average expected heterozygosity and range of allele frequencies for 5 SNPs and 1 STRP

Region[a]		No. of Populations	*Sac*I[b]	*HaeIIIc*	*Rsa*I	*Hae*III*a*	Glu487Lys	*D12S1344*
Africa	H[c]	8	349	53	322	330	0	788
	Frq.[c]	8	621–919	883–1000	633–974	586–959	1000	
	F_{st}	8 (7)[d]	0.065	0.071	0.051	0.053	0	0.027
SW Asia	H	2	450	497	374	405	0	769
	Frq.	2	324–354	475–548	726–774	696–732	1000	
	F_{st}	2	0.003	0.006	0.001	0.001	0	0.012
Europe	H	8	273	426	258	261	0	711
	Frq.	8	063–293	235–464	768–939	774–938	1000	
	F_{st}	8 (6)	0.031	0.028	0.022	0.020	0	0.022
NW Asia	H	2	310	414	269	275	0	688
	Frq.	2	117–290	272–684	820–860	811–862	1000	
	F_{st}	2 (0)	0.046	0.170	0.003	0.005	0	0.046
E Asia	H	7	291	109	304	309	276	555
	Frq.	7	671–890	908–987	679–929	679–920	634–1000	
	F_{st}	7	0.036	0.010	0.034	0.025	0.118	0.121
Pacific	H	2	450	486	121	140	0	498
	Frq.	2	583–717	545–609	889–977	889–957	1000	
	F_{st}	2	0.001	0.004	0.018	0.015	0	0.026
Siberia	H	1	400	214	486	480	0	662
	Frq.	1	720	878	583	590	1000	
	F_{st}	nc (1)[e]	nc	nc	nc	nc	nc	nc
N America	H	4	493	411	240	248	0	698
	Frq.	4	455–627	541–885	771–954	772–955	1000	
	F_{st}	4	0.023	0.070	0.041	0.042	0	0.053
S America	H	3	350	368	231	243	0	657
	Frq.	3	056–417	078–608	731–989	731–989	1000	
	F_{st}	3	0.110	0.110	0.092	0.093	0	0.118
Global pop.[f]	F_{st}	32	0.301	0.371	0.058	0.060	0.258	0.026

[a]Africa includes African Americans, and Europe includes European Americans.
[b]Average Heterozygosities (H), allele frequencies (Frq.) for ancestral alleles described in figure 1, and F_{st} values are shown for each SNP.
[c]Heterozygosity and allele frequency are given x 1000.
[d]Number of populations in parentheses for F_{st} values indicates number involved in global comparison at bottom; values shown for each region involve all populations in those regions.
[e]nc: Not calculable.
[f]The F_{st} value for global populations is based on the same 32 populations used for the computation of F_{st} at 117 biallelic (mostly SNPs, some insertions and deletions) reference sites.

Table 1 shows the average expected heterozygosity, the range of the allele frequencies, and F_{st} values for the five SNPs and one STRP in each geographic region. Two of the SNPs, the *Rsa*I site and the *Hae*IIIa site, show relatively little variation in frequency among the populations and their F_{st} values (Table 1) are both about .06. In contrast, the other SNP sites show highly significant allele-frequency variation among the geographical regions. For the *Sac*I site, the frequencies of the ancestral (site-present) allele are always higher than those of the derived (site-absent) allele in all African and all East Asian populations (the range of the site-present allele frequencies: .621 − .919 and .671 − .890, respectively), whereas the opposite ratio exists in all European and Southwest Asian populations (the range of the site-present allele frequencies: .063 − .354). The same pattern of the allele frequencies is observed at the *Hae*IIIc site, which is 10 kb downstream of the *Sac*I site. All individuals from four African populations (Biaka, Mbuti, Yoruba, Ibo) and almost all individuals from two other African populations (Chagga, Hausa) have only the site-present (ancestral) allele at the *Hae*IIIc site. Populations in East Asia similarly have very high frequencies of the ancestral allele. In the remaining populations the derived

 © University College London 2004

allele is more common and is the most frequent allele in the European populations. The frequencies of the site-present allele in East Asia are as high as those of African populations (the range: .908 − .987), but the site-present allele shows low frequencies in all European and Southwest Asian populations (the range: .235 − .548). For Native Americans, the frequencies of the *Sac*I site-present and the *Hae*IIIc site-present alleles span the ranges (.056 − .627 and .078 − .885, respectively) between Africans/East Asians and Europeans/Southwest Asians. The Glu487Lys site is polymorphic only in six East Asian populations (Chinese from San Francisco, Taiwan Han Chinese, Hakka, Japanese, Ami, Cambodians) and not in the other populations, which agrees with a previous study (Peterson *et al*. 1999a). Thus, the allele frequencies for these three sites vary greatly among populations as shown by the large F_{st} values, .258 − .371 (Table 1) and show strong geographic patterns, evident in Figure 2. The similarities/differences in the allele frequency patterns across populations (Figure 2) can be quantified as correlation coefficients between sites (Table 2). These values show two pairs of sites are highly correlated: *Sac*I with *Hae*IIIc at r = 0.95 and *Rsa*I with *Hae*IIIa at 0.99. Neither of these two patterns corresponds to the pattern of population variation shown by the functional site.

Table 2 Correlation coefficient matrix[a]

	*Sac*I	*Hae*IIIc	*Rsa*I	*Hae*IIIa
*Sac*I				
*Hae*IIIc	0.951			
*Rsa*I	− 0.353	− 0.394		
*Hae*IIIa	− 0.346	− 0.379	0.985	
Glu487Lys	− 0.485	− 0.375	− 0.054	− 0.063

[a]Pearson product moment correlation of allele frequencies across 37 populations, as graphed in figure 2.

To provide a better context for the different F_{st} values, we calculated F_{st} values for a subset consisting of 32 populations on which data from 117 reference sites at other loci exist in our lab (Pakstis *et al*. 2002). The F_{st} values of the *Sac*I, *Hae*IIIc, and Glu487Lys sites were .30, .37, and .26, respectively, which were about two SD (standard deviation) above the average for the reference sites: 0.140 ± .068. Thus, the global survey in this study has revealed that not only the functional polymorphism (Glu487Lys) but also the upstream polymorphisms (*Sac*I and *Hae*IIIc) are outliers for the F_{st} values, showing more variation among populations than most random SNPs.

Haplotype Frequencies

Figure 3 shows the 5-SNP haplotype frequencies for nine geographic regions (Africa, Southwest Asia, Europe, Northwest Asia, East Asia, Pacific, Siberia, North

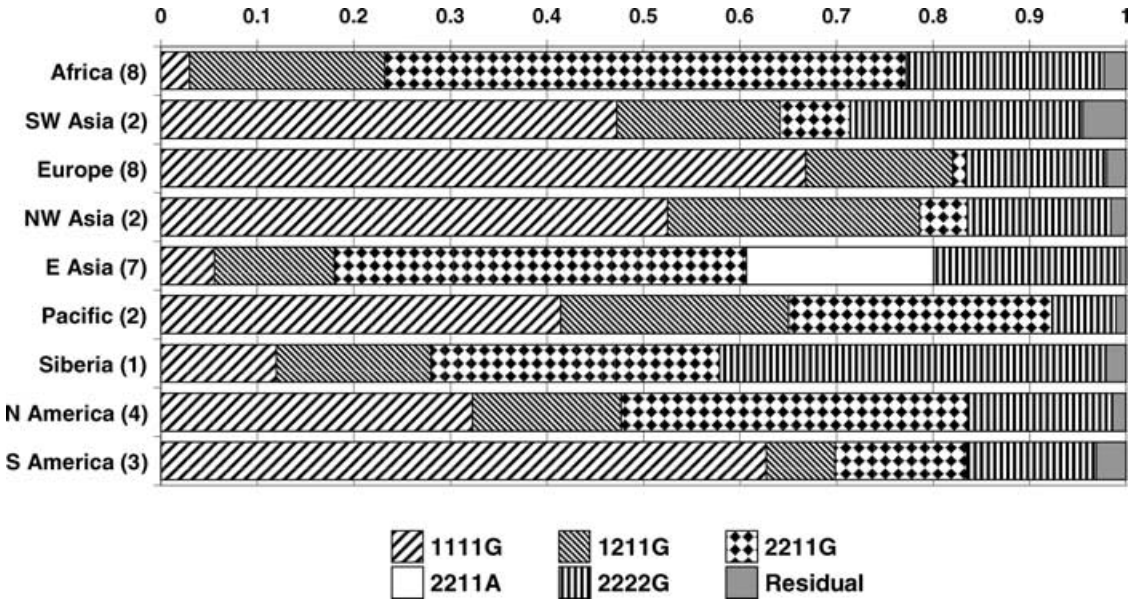


Figure 3 *ALDH2* 5-SNP haplotype frequencies in each geographic region. Abbreviations are as follows, SW: Southwest, NW: Northwest, E: East, S: Siberia, N: North, S: South. The numbers of populations are shown in parentheses. "Africa" and "Europe" include African Americans and European Americans, respectively.

© University College London 2004

America, South America). Out of 32 possible haplotypes, 15 haplotypes were estimated to have non-zero values, and 10 out of these 15 haplotypes were definitely present in at least one individual in our samples, while five had inferential evidence for existing. Out of the 10 haplotypes, only four major haplotypes account for almost all chromosomes (97.8%) in all but the East Asian populations. The *ALDH2* 487Lys* allele defines a fifth haplotype (2211A) that was present only in East Asians. Except for this 2211A haplotype, the other four major haplotypes were observed in all eight geographical regions. However, the haplotype frequencies showed marked differences among the geographic regions. For Africans, three haplotypes, 1211G, 2211G, and 2222G, were common (averaging 20.3%, 53.9%, 20.5%, respectively), and for African Americans it was almost the same. The haplotype 1111G was rare among Africans (3.0%) and East Asians (5.6%), and somewhat more common among African Americans (11.3%). These were much lower frequencies than exist among Europeans and Southwest Asians (66.8 and 47.3%, respectively). In contrast, the haplotype 2211G was quite rare in Europe and Southwest Asia (1.4% and 7.2%, respectively) but common (10.0% – 53.9%) elsewhere. The haplotypes 1211G and 2222G were observed at relatively similar frequencies all over the world (ranges: 12.4% – 26.1% and 13.4% – 24.2%, respectively), except the frequencies are slightly lower in South America (7.1%) and the Pacific (6.7%) for 1211G and 2222G, respectively. The combined frequencies of the remaining haplotypes (Residual) were less than 5.0% in all geographic regions, indicating that each haplotype in the residuals was extremely uncommon among the samples we examined.

The geographic variation in frequencies was more pronounced when the STRP, *D12S1344*, was included in the haplotypes. The graphs in Figure 4 show the distributions of the 5-SNP haplotypes according to *D12S1344* allele. Overall, the 236 and 240 alleles were the most common alleles in all regions of the world. In Africa, East Asia, and the Americas, allele 240 was the most common allele and occurred primarily in conjunction with 2211G, except in South America where allele 240 occurred mainly in conjunction with 1111G. In Europe and Southwest Asia, the distribution patterns were quite similar to each other, but different from those of Africa, East Asia, and America: allele 236 was the most common allele, occurring almost exclusively with the 1111G haplotype, and allele 240 was the second most common allele mainly occurring with the 1111G and 1211G haplotypes. Alleles 226 and 228 were observed in all regions, except in South America, and almost always with the 2222G haplotype. In East Asia, the frequency of allele 240 was considerably higher (more than 50%) than in Africa and America, and the East Asian specific haplotype, 2211A, was observed mainly in conjunction with allele 240 and occasionally with allele 238, both of which also occurred with 2211G. Thus, the *D12S1344* allele frequency and the 5-SNP haplotype distribution patterns were quite distinct between Africa/East Asia/America and Europe/Southwest Asia, and the 5-SNP haplotype distribution patterns were different among Africa, East Asia and the Americas.

Linkage Disequilibrium

The overall ξ coefficient, which is a measure of the overall deviation from random association (Zhao *et al*. 1999), showed very strong LD across the *ALDH2* locus. The ξ values for four non-coding SNPs were uniformly high around the world (the ξ range: 1.20 – 3.85) (Figure 5). Mbuti, Nasioi, and R. Surui had slightly lower values, though still with statistical significance ($p < 0.01$) ($\xi =$ .27, .74, and .69, respectively). We also found high ξ values that extend over 123 kb using a segment test (cf. Zhao *et al*. 1999; Kidd *et al*. 2000) between the 5-SNP haplotypes and *D12S1344* (the ξ range: .47 – 2.55) in all the populations with statistical significance ($p < 0.01$), except for the Biaka, Mbuti, Ibo, Finns, and Nasioi ($p = 0.01$, 0.49, 0.01, 0.01, and 0.10, respectively). Thus, the high overall ξ values indicate strong LD across 44 kb of the *ALDH2* locus and extending downstream for 123 kb to *D12S1344* in all geographic regions.

Pairwise LD has been evaluated with three coefficients – D', Δ^2, and ξ – with significance evaluated by a permutation test. In most instances the values are statistically significant except when heterozygosity is very low for one of the sites. The values show a complicated pattern, illustrated in Figure 6, for three of the six pairwise combinations that exist worldwide. D' is

 © University College London 2004

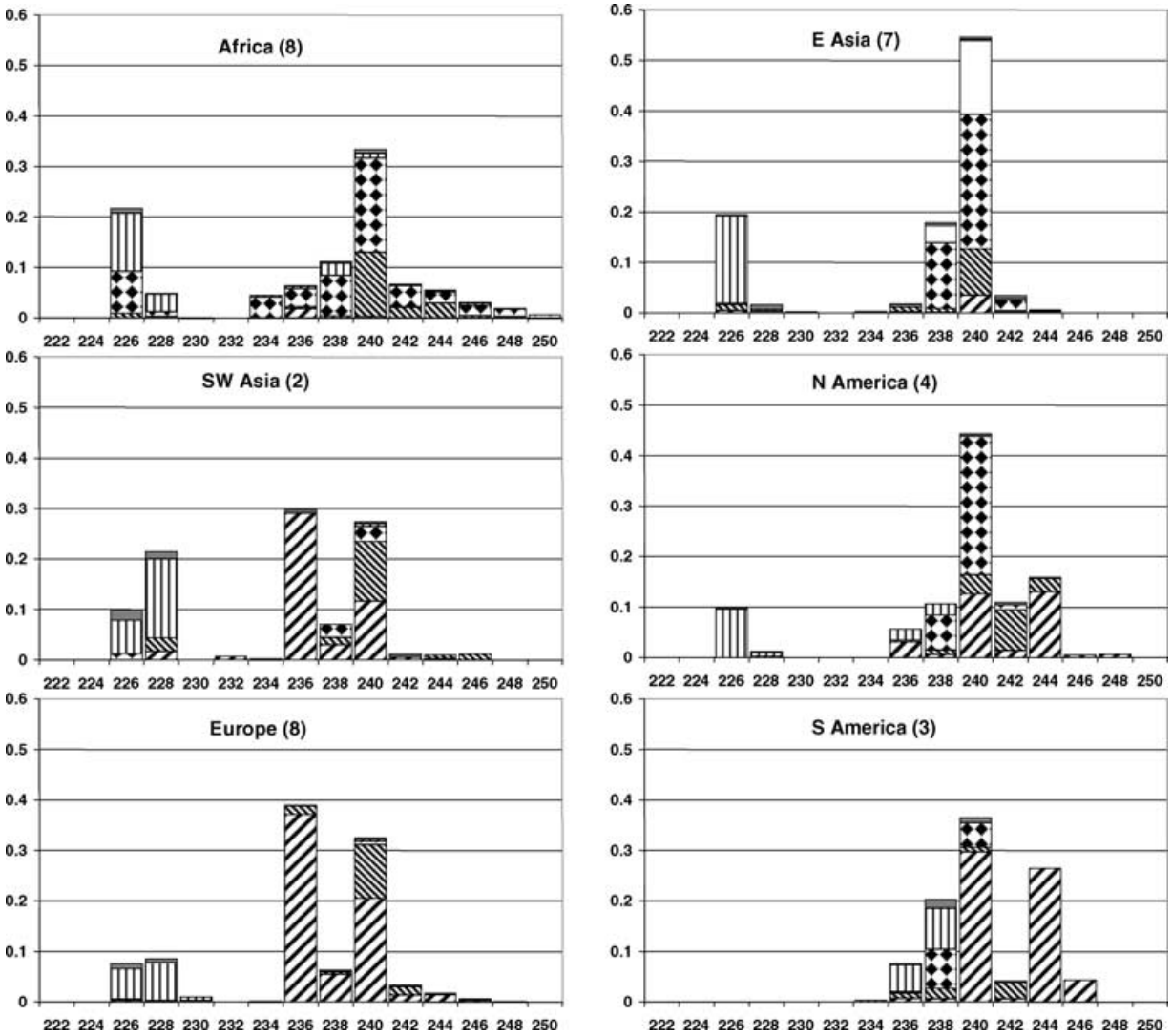


Figure 4 *ALDH2* 5-SNP haplotype frequencies with *D12S1344* alleles in major geographic regions. The numbers on the X axes are the allele sizes of the di-nucleotide repeat polymorphism. The patterns correspond to haplotypes shown in figure 3.

not illustrated because most pairwise comparisons had values of 1.0, providing no information on relative levels of LD. Figure 6a shows pairwise LD values (ξ) between *Sac*I and *Rsa*I (20 kb), and between *Rsa*I and *Hae*IIIa (10 kb), respectively. Pairwise ξ values between *Sac*I and *Rsa*I showed high LD in Europe and Southwest Asia, very low LD in Africa and East Asia, and low LD in Northeast Asia and Americas. Meanwhile, pairwise ξ values between the *Rsa*I and *Hae*IIIa sites showed reasonably high LD around the world with statistical significance ($p < 0.01$), except for the Nasioi ($p = 0.09$). Figure 6b shows pairwise LD values between the *Hae*IIIc and *Hae*IIIa sites with Δ^2 and ξ in order to compare two different statistical values. The patterns of Δ^2 and ξ were quite similar to one another: pairwise LD values between the *Hae*IIIc and *Hae*IIIa sites were high in Europe and Southwest Asia, low in East Asia, and intermediate in Pacific, Northeast Asia, and Americas. For five sub-Saharan Africans (Biaka, Mbuti, Yorba, Ibo, and Hausa), pairwise LD values between the *Hae*IIIc and *Hae*IIIa sites could not be calculated because the *Hae*IIIc site is not polymorphic in those populations. Similarly, the pairwise ξ values between the *Hae*IIIc and *Rsa*I sites, and the *Sac*I and *Hae*IIIa sites show very strong LD in Europe and Southwest Asia, and less LD in Africa, East Asia, and Americas, while the pairwise ξ values between

© University College London 2004

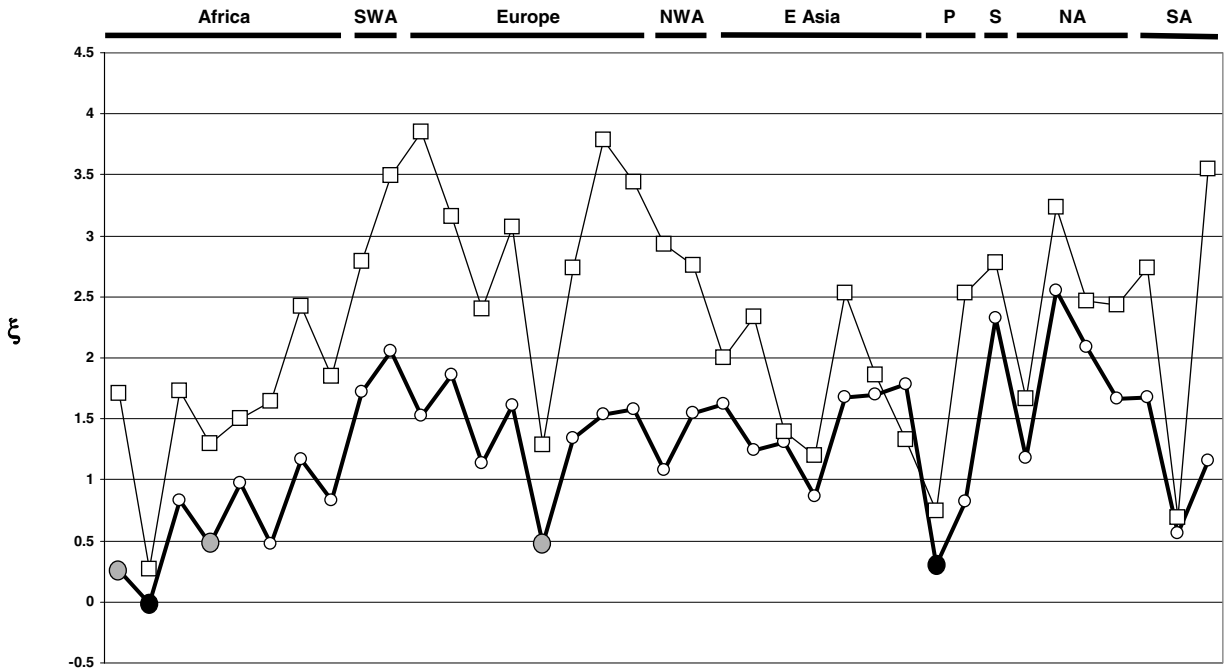


Figure 5 Overall LD for the four non-coding SNPs and segment tests between the *ALDH2* locus and *D12S1344*. The LD values are the ξ coefficient (Zhao et al. 1999). The population names are omitted at the bottom of the graph and the geographic regions are shown at the top. The order of the populations is the same as that of figure 2. The open squares represent the overall ξ coefficient values, whereas the open circles represent the segment tests. The filled and gray circles, corresponding to Mbuti and Nasioi, and Biaka, Ibo and Finns, are not statistically significant ($p \geq 0.02$, $0.02 > p \geq 0.01$, respectively); all other values are significant at $p < 0.01$.

the *Sac*I and *Hae*IIIc sites increase from Africa to Europe, East Asia, and the Americas (data not shown).

Because the Glu487Lys site showed variation only in East Asia, pairwise LDs with the functional variant could not be calculated for the other populations. All pairwise ξ values including the *ALDH2* Glu487Lys site were relatively low (the ξ range: $-.04$ to $+.28$), and lower values were observed between the *Hae*IIIc and the *ALDH2* Glu487Lys sites (30 kb) (the ξ range: $-.04$ to $+.04$) than for the other pairs. Thus, there was no relationship between LD values and the physical distance between the sites.

Discussion

Haplotype Evolution and Geographic Distributions

The *Hae*IIIc site is not polymorphic in five sub-Saharan Africans, and the *ALDH2* Glu487Lys site is polymorphic only in East Asian populations. The results indicate the *Hae*IIIc site and the *ALDH2* Glu487Lys site are relatively young polymorphisms. However, the ages of the other polymorphic sites we examined are as old as modern humans' expansion, because all of them have sufficiently high heterozygosity in all 37 populations.

We found four common and one East Asian-specific haplotype, based on five SNPs examined in 37 worldwide human populations. Figure 7 shows a phylogenetic network for the major haplotypes with the haplotype frequencies in each geographic population. The circles represent the haplotypes and the areas of the circles represent the relative global frequencies of the haplotypes. The segments of the circles show the proportions of the haplotypes that occurred in each geographic region. The ancestral 2211G is the root of this network tree. Four of the five common haplotypes can be linked by single mutation events. The common 1111G haplotype is two mutations away from the ancestral haplotype; the intermediate 1211G haplotype is also among the common haplotypes. This sequential pattern of mutations also indicates that the *Hae*IIIc polymorphism is relatively young despite the derived allele being the most common haplotype globally. In contrast, the

 © University College London 2004

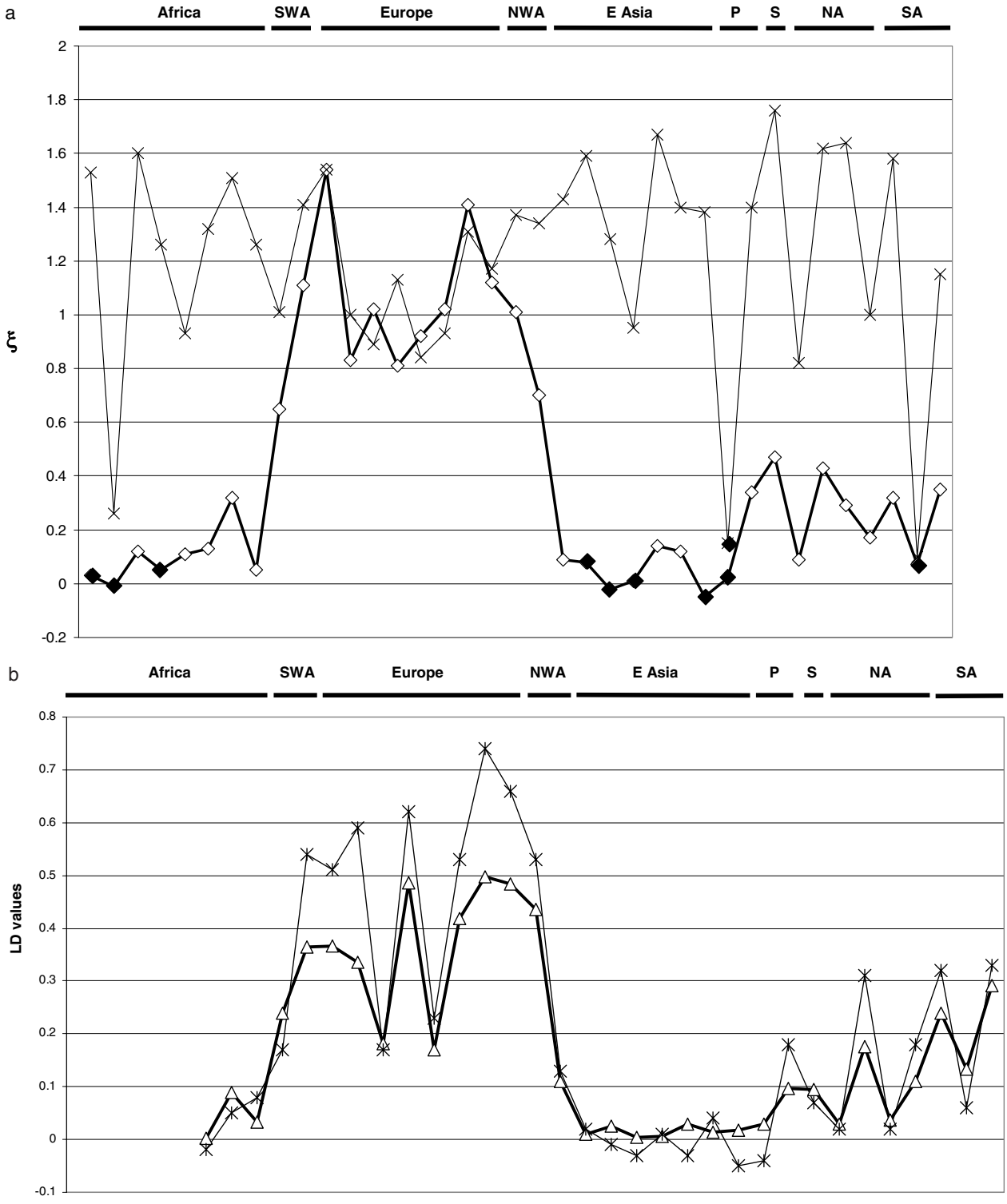


Figure 6 a: Pairwise LD values (ξ) between *Sac*I and *Rsa*I (20 kb), and between *Rsa*I and *Hae*IIIa (10 kb). The diamonds represent ξ values between *Sac*I and *Rsa*I, whereas the Xs represent ξ values between *Rsa*I and *Hae*IIIa. Filled diamonds indicate no statistic significance ($p > 0.01$) between *Sac*I and *Rsa*I for Biaka, Mbuti, Ibo, Taiwan Han Chinese, Hakka, Japanese, Cambodians, Nasioi, and Rondonia Surui, and between *Rsa*I and *Hae*IIIa for Nasioi. b: Pairwise LD values (ξ and Δ^2) between *Hae*IIIc and *Hae*IIIa. The asterisks represent the ξ coefficient values, whereas the triangles represent Δ^2 (Devlin & Risch, 1995). The order of the populations is the same as in figure 2.

© University College London 2004

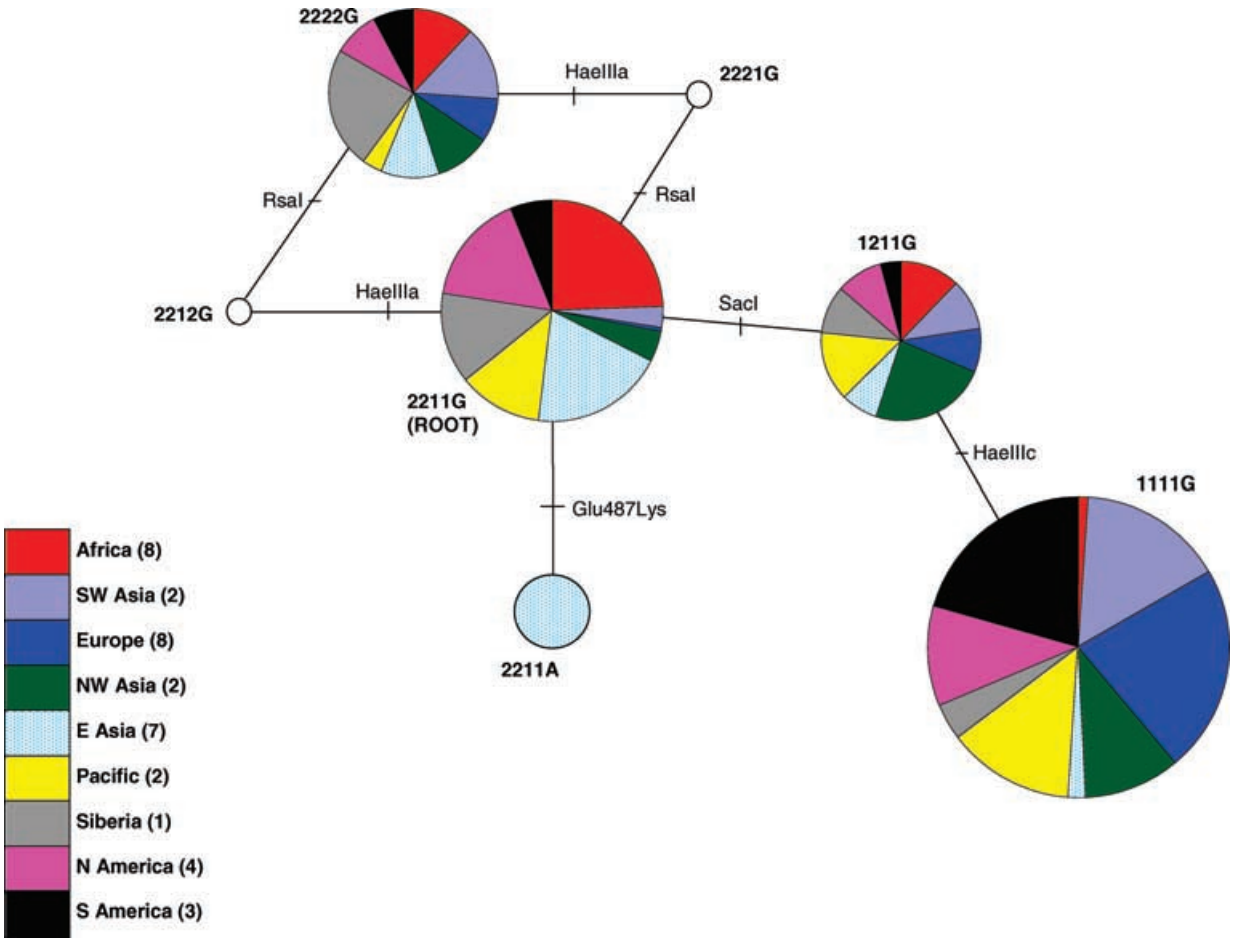


Figure 7 Phylogenetic network for *ALDH2* 5-SNP. The seven circles represent 5-SNP haplotypes observed in our samples. The circle size is proportional to the world average of frequencies for the four major haplotypes with subdivisions showing the frequencies of the haplotypes in each geographic region. The East Asian-specific and two minor haplotypes are included in the network. The ambiguity in the evolution of the 2222G haplotype is indicated by two pathways from the ancestral haplotype.

2222G haplotype is two mutational steps away from the ancestral haplotype and both of the two possible intermediates, 2212G and 2221G, exist in various populations but are not common anywhere. Since the 2222G haplotype is ubiquitously distributed, it would appear to have arisen in Africa and drifted to an appreciable frequency prior to the expansion of modern humans out of Africa. Of the 16 possible haplotypes for the globally ubiquitous polymorphisms (not counting the East Asian-specific functional variant), nine have been seen somewhere in the world. All of these can be explained by single crossover events involving two haplotypes present in those populations. Because they occur sporadically around the world, it seems likely that they represent relatively recent crossover products from multiple independent events, rather than ancient lineages from a single recombination event. However, only studying additional polymorphisms and additional populations will enable resolution of that question.

In Africa the ancestral haplotype has multiple STRP alleles associated with it, in accordance with it representing an ancient lineage on which multiple mutational events at the STRP could have arisen. Outside of Africa, the association of particular STRP alleles with particular haplotypes is much stronger and is the basis for the strong linkage disequilibrium seen across the interval from *ADLH2* to *D12S1344*. These data argue both that the STRP has a low mutation rate relative to random genetic drift since populations left Africa, and that recombination across this 83 kb interval is quite low relative to this time span of roughly 100,000 years.

Compared to the Peterson *et al*. (1999a,b) results, we have subdivided their H2 haplotype using the *Hae*IIIc SNP site into haplotypes 1211G and 1111G. They

 © University College London 2004

commented that they did not see a strong out-of-Africa effect at *ALDH2* but this new SNP does show such an effect. Their haplotype H1 corresponds to the ancestral haplotype (root in figure7). Their H3 haplotype corresponds to our 2222G with the *Rsa*I site in our study providing duplicate information to the *Hae*IIIa site that they also studied. The other sites that Peterson *et al*. (1999a) studied provided essentially redundant information with the three sites studied in common. While we identified one significant subdivision of their H2 into two different common haplotypes, it seems unlikely that a much more complicated evolutionary tree for *ADLH2* haplotypes will be found without investigation of many more SNPs across the gene.

A Haplotype Block

We found strong overall LD (ξ) values across the *ALDH2* locus as well as significant LD between the *ALDH2* locus and *D12S1344*, a physical distance of 123kb. In our analysis, and in the Peterson *et al*. (1999a,b) analyses, only the same few haplotypes account for almost all chromosomes in populations from all regions. Other haplotypes are rare and sporadic. It appears that this region would qualify as a "haplotype block," irrespective of the origin of the population. However, as is obvious from Figure 3, the haplotype frequencies differ considerably. Part of the rationale for undertaking the "hap map" project is that an entire block can be studied by testing only the few "tagging SNPs" that discriminate between the common haplotypes, thereby saving typing effort in future association studies of common disorders. For the *ALDH2* gene, there are four haplotypes to be distinguished (five in East Asia). In Africa, 95% of the chromosomes are 1211G, 2211G, and 2222G. The first site (*Sac*I) is necessary; adding either the third or fourth site (*Rsa*I or *Hae*IIIa) gives a pair that is sufficient to discriminate among these three haplotypes. In Europe and Southwest Asia, over 90% of the chromosomes are 1111G, 1211G and 2222G. The second site (*Hae*IIIc) is necessary and either the first, third, or fourth site gives a pair that is sufficient to distinguish the predominant chromosomes. In East Asia, 1211G, 2211G, 2211A, and 2222G account for 90% of all chromosomes. Both the first site and the functional variant *ALDH2* 487Lys* (fifth site) are required, as well as either the third or fourth site to discriminate among these four common haplotypes. In the other regions, 1111G, 1211G, 2211G, and 2222G account for 94% of the chromosomes. Again both the first site and the second site are required, as well as either the third or fourth site. In summary, in different parts of the world different subsets of the SNPs are needed to discriminate among common haplotypes. Collectively for all regions of the world, four of the five sites are necessary. Only sites three and four are equivalent, and while one is required the other can be omitted.

The additional SNPs studied by Peterson *et al*. (1999a) do not appear, by inference, to increase significantly the number of common haplotypes. Thus, there may be many SNPs in this region that are redundant and savings are possible. However, the most relevant point is that the minimum set of SNPs required to distinguish between the common haplotypes in one population may be insufficient/inadequate for a different population. Peterson *et al*. (1999a) identified a SNP that was completely associated with the functional variant in their samples. That SNP would suffice for distinguishing the haplotype with the functional variant, but if it and the functional variant were not identified the relevant haplotype 2211A would be pooled with the frequent 2211G haplotype, greatly reducing the power to find an effect in an association study.

As noted by Peterson *et al*. (1999a,b) the variation in LD among populations, however measured, is complex. While some of that variation can be attributed to the East Asian-specific haplotype and possible involvement of selection (see below), it is clear from all of the analyses that recombination is not a relevant factor. The complex pattern of LD is solely the result of the relative frequencies of the few common haplotypes, and probably the result of random genetic drift in most cases. This emphasizes the fact that LD is a statistical abstraction based on haplotype frequencies and not, *per se*, a fundamental aspect of the genome. Clearly relative rates of recombination are not responsible for the different levels of LD among the SNPs in *ALDH2*.

An East Asian-Specific Allele in Exon 12

We confirmed that the functional variant at the *ALDH2* Glu487Lys site is present only in East Asian populations. Three SNPs are known in exon 12 of the *ALDH2* gene: two G to A substitutions at np1464 and np1486 have

© University College London 2004

been found by Novoradovsky *et al*. (1995), and a G to A substitution at np1510 was reported by Yoshida *et al*. (1984). The coding SNP, Glu487Lys, corresponds to the nucleotide substitution at np1510. The nucleotide substitution at np1464 is a silent change found in Native Americans (Novoradovsky *et al*. 1995), whereas those at both np1486 and np1510 result in the amino acid change Glu to Lys at amino acid positions 479 and 487. The nucleotide substitution at np1486 occurs on the same haplotype as the deficient enzyme (Novoradovsky *et al*. 1995), and it has also been observed only in East Asian populations. Some previous studies have reported that inactive ALDH2 enzymes have been found in South Americans, Atacammenos, Mapuche and Shuara (at frequencies of around 40%) (Goedde *et al*. 1986). That we did not find the functional variant *ALDH2*487Lys* in Native Americans could be explained in two ways: one is simply that the North and South American samples we examined do not include people who have this variant, and another is that Native American deficient enzyme(s) is/are caused by unknown nucleotide substitutions elsewhere in the gene. More samples from various tribes of Native Americans must be genotyped to clarify the discrepancy.

Hypotheses for Natural Selection on ALDH2

The F_{st} values of three out of five SNPs at the *ALDH2* locus are obviously unusual, as are the haplotype frequencies and LD values. The F_{st} values of *Sac*I, *Hae*IIIc, and Glu487Lys are remarkable departures (.30, .37, and .26) from the average of 117 reference sites (.14) in other loci. It is interesting that the F_{st} value of the *Sac*I site in the regulatory region is higher than the F_{st} of the East Asian-specific *ALDH2* Glu487Lys site. The *Sac*I and *Hae*IIIc sites are highly correlated in their allele frequencies across populations. This is not too surprising since they correspond to one of the "arms" of the evolutionary tree of the haplotypes (Figure 7). The pattern, however, is quite different from that of the Glu487Lys functional site (Figure 2 & Table 2). Thus, the high F_{st} values shared by the *Sac*I and *Hae*IIIc sites and the Glu487Lys site are not caused by hitchhiking of those non-coding SNPs with the Glu487Lys site. The interesting question of whether separate selection forces are operating arises, especially since the *Sac*I site is in the promoter region. The result of a transfection assay shows that the G allele (*Sac*I site present: 2) is about 3-fold more active than the A allele (*Sac*I site absent: 1) in hepatoma cells (Chou *et al*. 1999). Furthermore, some individual sites in the Class I ADH cluster have also shown extremely high F_{st} values (Osier *et al*. 2002a). These high F_{st} values may imply that selection has operated at these sites, or closely linked loci, in both the *ADH* and *ALDH2* genes in modern humans. The high F_{st} at the *ALDH2 Sac*I site is largely attributable to the greatly reduced frequency of this ancestral and more active regulatory allele. If selection has operated, it has been most efficacious in Europe.

Deficiency of the ALDH2 enzyme induces a high concentration of acetaldehyde in humans following ingestion of alcohol. The ALDH2 enzyme is a tetramer; the deficient allele product dramatically reduces the stability of the structure of the ALDH2 tetramer, resulting in greatly reduced activity for all hetero-tetramers in the heterozygote. The enzyme in the deficient allele homozygote has no catalytic activity. Acetaldehyde is generated from ethanol by alcohol dehydrogenase. Osier *et al*. (2002a) have reported that particular alcohol dehydrogenase haplotypes, 221221 and 221112, (see Table 4 in Osier *et al*. (2002a)) exist at high frequency only in East Asians (average: 65.0%) and in Africa (average: 16.1%), respectively. These haplotypes, 221221 and 221112, are characterized by functional variants in *ADH1B*, 47His and 369Cys, respectively. Interestingly, both the *ADH1B*47His* and *ADH1B*369Cys* alleles demonstrate high activity for catalyzing ethanol digestion, resulting in increased concentration of acetaldehyde with alcohol intake.

Why are enzymes leading to a high level of acetaldehyde common in East Asia? It is puzzling that *ADH1B* and *ALDH2* alleles, presumably leading to high levels of a toxic substance, acetaldehyde, should rise to high frequencies. There might be two hypothetical explanations for the paradox, both involving selection. The first hypothesis is that variant(s) of ADH and ALDH2 have alternative functions, besides alcohol metabolism, that are more essential than the risk of a high level of acetaldehyde. The Class I *ADH* genes are expressed not only in adult liver but also fetal liver, intestine and lung (Bilanchone *et al*. 1986), whereas the *ALDH2* gene is expressed in adult/fetal liver, kidney, adult muscle, heart,

 © University College London 2004

and fetal lung (Stewart *et al.* 1996). Such expression in a large number of tissues implies these enzymes might have other unknown functions. However, it does not give a clear explanation for the "East Asian-specific" variants. Another hypothesis is that a higher concentration of acetaldehyde has advantage(s) for some endemic disease in East Asia, past or present. This may be related to protection against parasite(s) infection, such as *Entamoeba histolytica*, which causes significant mortality due to ulcerative dysentery and extra intestinal abscesses (Goldman & Enoch, 1990). It is well known that nitroimidazole, a specific ALDH inhibitor, is effective against a number of anaerobes and microaerophiles, implying that a high level of acetaldehyde inhibits growth of such parasites. Though there are no data for either hypothesis, they are both plausible explanations.

Alternatively, selection might have operated at closely linked loci around the *ADH* gene cluster and the *ALDH2* locus. However, because both the *AD*H gene cluster and *ALDH2* are on different chromosomes (chromosomes 4 and 12, respectively) this possibility seems less likely than other hypotheses. At any rate, we emphasize again that these two alleles, *ALDH2* 487Lys* and *ADH1B* 47His*, at unlinked loci, both occur at high frequencies in East Asia. This pattern is difficult to explain by genetic drift alone. The data presented here confirm this pattern for a larger number of populations than previously studied.

More haplotype and LD analyses of the *ADH* gene cluster and *ALDH2* gene will increase our understanding of possible selection involving these genes. Furthermore, investigations involving the *ADH* gene cluster and *ALDH2* gene in non-human primates will give more information about selection in humans, which could yield a good animal model for studies of alcohol metabolism and alcoholism. Such studies will provide a key to the evolutionary history of the *ADH* and *ALDH2* genes.

Acknowledgements

This work was funded in part by National Institute of Health grant AA09379 and GM57672 to K.K.K. and NSF BCS-9912028 to J.R.K., and in part by National Health Research Institute, Taiwan, ROC, Grant NHRI-EX91-8939SP to R.B.L., and National Science Council, Taiwan, ROC, Grant NSC 90-2314-B-016-081 to R.B.L. We thank William C. Speed, Roy Capper, Andrew R. Dyer, and Valeria Ruggeri, for their excellent technical assistance. We are indebted to the following people who helped assemble the diverse population collection used in this study: F.L. Black, L.L. Cavalli-Sforza, K. Dumars, J. Friedlaender, K. Kendler, W. Knowler, F. Oronsaye, J. Parnas, L. Peltonen, L.O. Schulz and K. Weiss. In addition, some of the cell lines were obtained from the National Laboratory for the Genetics of Israeli Populations at Tel Aviv University, Israel, and the African American samples were obtained from the Coriell Institute for Medical Research, Camden, NJ. Special thanks are due to the many hundreds of individuals who volunteered to give blood samples for studies such as this. Without such participation of individuals from diverse parts of the world we would be unable to obtain a true picture of the genetic variation in our species.

Electronic-Database Information

Accession numbers and URLs for data presented herein are as follows:

ALFRED (Alelle Frequency Database), http://alfred.med.yale.edu/alfred/index.asp

JSNP (a database of Japanese Single Nucleotide Polymorphism), http://snp.ims.u-tokyo.ac.jp

Mfold (RNA and DNA folding applications), http://bioinfo.math.rpi.edu/~mfold/dna

GenBank, http://www.ncbi.nlm.nih.gov/Genbank/ (for 12q24.2 [accession number NT_009775])

Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/(ALDH2 [MIM 100650]).

References

Anderson, M. A. & Gusella, J. F. (1984) Use of cyclosporin A in establishing Epstein-Barr virus-transformed human lymphoblastoid cell lines. *In Vitro* **20**, 856–858.

Bilanchone, V., Duester, G., Edwards, Y. & Smith, M. (1986) Multiple mRNAs for human alcohol dehydrogenase (ADH): developmental and tissue specific differences. *Nucleic Acids Res* **14**, 3911–3926.

Cann, R. L., Stoneking, M. & Wilson, A. C. (1987) Mitochondrial DNA and human evolution. *Nature* **325**, 31–36.

Castiglione, C. M., Deinard, A. S., Speed, W. C., Sirugo, G., Rosenbaum, H. C., Zhang, Y., Grandy, D. K., Grigorenko, E. L., Bonne-Tamir, B. & Pakstis, A. J. *et al*. (1995) Evolution of haplotypes at the DRD2 locus. *Am J Hum Genet* **57**, 1445–1456.

Chang, F. M., Kidd, J. R., Livak, K. J., Pakstis, A. J. & Kidd, K. K. (1996) The world-wide distribution of allele frequencies at the human dopamine D4 receptor locus. *Hum Genet* **98**, 91–101.

Chen, X., Levine, L. & Kwok, P. Y. (1999) Fluorescence polarization in homogeneous nucleic acid analysis. *Genome Res* **9**, 492–498.

© University College London 2004

Chou, W. Y., Stewart, M. J., Carr, L. G., Zheng, D., Stewart, T. R., Williams, A., Pinaire, J. & Crabb, D. W. (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. *Alcohol Clin Exp Res* **23**, 963–968.

Clark, A. G., Weiss, K. M., Nickerson, D. A., Taylor, S. L., Buchanan, A., Stengard, J., Salomaa, V., Vartiainen, E., Perola, M., Boerwinkle, E. & Sing, C. F. (1998) Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *Am J Hum Genet* **63**, 595–612.

Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J. & Lander, E. S. (2001) High-resolution haplotype structure in the human genome. *Nat Genet* **29**, 229–232.

DeMille, M. M., Kidd, J. K., Ruggeri, V., Palmatier, M. A., Goldman, D., Odunsi, A., Okonofua, F. Grigorenko, E., Schulz, L. O. & Bonne-Tamir, B. *et al*. (2002) Population variation in linkage disequilibrium across the COMT gene considering promoter region and coding region variation. *Hum Genet* **111**, 521–537.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977) Maximum likelihood from incomplete data via EM algorithm. *J Roy Statis Soc Ser B* **39**, 1–22.

Devlin, B. & Risch, N. (1995) A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29**, 311–322.

Edenberg, H. J. & Bosron, W. F., (1997) Alcohol dehydrogenase. In: *Biotransformation* (ed. Guengreich, F. P.). Vol. 3 in: *Comprehensive toxicology* (eds. Sipes, I. G., McQueen, C. A., Gandolfi, A. J.). pp 119 – 131. Pergamon, New York.

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A. & Faggart, M. *et al*. (2002) The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229.

Goedde, H. W., Harada, S. & Agarwal, D. P. (1979) Racial differences in alcohol sensitivity: a new hypothesis. *Hum Genet* **51**, 331–334.

Goedde, H. W., Agarwal, D. P., Harada, S., Rothhammer, F., Whittaker, J. O. & Lisker, R. (1986) Aldehyde dehydrogenase polymorphism in North American, South American, and Mexican Indian populations. *Am J Hum Genet* **38**, 395–399.

Goedde, H. W., Agarwal, D. P., Fritze, G., Meier-Tackmann, D., Singh, S., Beckmann, G., Bhatia, K., Chen, L. Z. Fang, B. & Lisker, R. *et al*. (1992) Distribution of ADH2 and ALDH2 genotypes in different populations. *Hum Genet* **88**:344–346.

Goldman, D. & Enoch, M. A. (1990) Genetic epidemiology of ethanol metabolic enzymes: a role for selection. *World Rev Nutr Diet* **63**, 143–160.

Hammer, M. F. (1995) A recent common ancestry for human Y chromosomes. *Nature* **378**, 376–378.

Harada, S., Agarwal, D. P. & Goedde, H. W. (1980) Isozymes of alcohol dehydrogenase and aldehyde dehydrogenase in Japanese and their role in alcohol sensitivity. *Adv Exp Med Biol* **132**, 31–39.

Harada, S., Agarwal, D. P. & Goedde, H. W. (1981) Aldehyde dehydrogenase deficiency as cause of facial flushing reaction to alcohol in Japanese. *Lancet* **2**, 982.

Harada, S., Agarwal, D. P., Goedde, H. W., Tagaki, S. & Ishikawa, B. (1982) Possible protective role against alcoholism for aldehyde dehydrogenase isozyme deficiency in Japan. *Lancet* **2**, 827.

Harada, S., Okubo, T., Nakamura, T., Fujii, C., Nomura, F., Higuchi, S. & Tsutsumi, M. (1999) A novel polymorphism (-357 G/A) of the ALDH2 gene: linkage disequilibrium and an association with alcoholism. *Alcohol Clin Exp Res* **23**, 958–962.

Hawley, M. E. & Kidd, K. K. (1995) HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. *J Hered* **86**, 409–411.

Hsu, L. C., Bendel, R. E. & Yoshida, A. (1988) Genomic structure of the human mitochondrial aldehyde dehydrogenase gene. *Genomics* **2**, 57–65.

Iyengar, S., Seaman, M., Deinard, A. S., Rosenbaum, H. C., Sirugo, G., Castiglione, C. M., Kidd, J. R. & Kidd, K. K. (1998) Analyses of cross species polymerase chain reaction products to infer the ancestral state of human polymorphisms. *DNA Seq* **8**, 317–327.

Jeffreys, A. J., Kauppi, L. & Neumann, R. (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat Genet* **29**, 217–222.

Jorde, L. B. (1995) Linkage disequilibrium as a gene-mapping tool. *Am J Hum Genet* **56**, 11–14.

Kidd, K. K. & Cavalli-Sforza, L. (1974) The role of genetic drift in the differentiation of Icelandic and Norwegian cattle. *Evolution* **28**, 381–395.

Kidd, J. R., Black, F. L., Weiss, K. M., Balazs, I. & Kidd, K. K. (1991) Studies of three Amerindian populations using nuclear DNA polymorphisms. *Hum Biol* **63**, 775–794.

Kidd, K. K., Morar, B., Castiglione, C. M., Zhao, H., Pakstis, A. J., Speed, W. C., Bonne-Tamir, B., Lu, R. B., Goldman, D. & Lee, C. *et al*. (1998) A global survey of haplotype frequencies and linkage disequilibrium at the DRD2 locus. *Hum Genet* **103**, 211–227.

Kidd, J. R., Pakstis, A. J., Zhao, H., Lu, R. B., Okonofua, F. E., Odunsi, A., Grigorenko, E., Tamir, B. B., Friedlaender, J., Schulz, L. O., Parnas, J. & Kidd, K. K. (2000) Haplotypes and linkage disequilibrium at the phenylalanine hydroxylase locus, PAH, in a global representation of populations. *Am J Hum Genet* **66**, 1882–1899.

Koch, H. G., McClay, J., Loh, E. W., Higuchi, S., Zhao, J. H., Sham, P., Ball, D. & Craig, I. W. (2000) Allele association

 © University College London 2004

studies with SSR and SNP markers at known physical distances within a 1 Mb region embracing the ALDH2 locus in the Japanese, demonstrates linkage disequilibrium extending up to 400 kb. *Hum Mol Genet* **9**, 2993–2999.

Laan, M. & Pääbo, S. (1997) Demographic history and linkage disequilibrium in human populations. *Nat Genet* **17**, 435–438.

Lewontin, R. C. (1964) The interaction of selection and linkage. I. General considerations: heterotic models. *Genetics* **49**, 49–67.

Novoradovsky, A., Tsai, S. J., Goldfarb, L., Peterson, R., Long, J. C. & Goldman, D. (1995) Mitochondrial aldehyde dehydrogenase polymorphism in Asian and American Indian populations: detection of new ALDH2 alleles. *Alcohol Clin Exp Res* **19**, 1105–1110.

Osier, M., Pakstis, A. J., Kidd, J. R., Lee, J. F., Yin, S. J., Ko, H. C., Edenberg, H. J., Lu, R. B. & Kidd, K. K. (1999) Linkage disequilibrium at the ADH2 and ADH3 loci and risk of alcoholism. *Am J Hum Genet* **64**, 1147–1157.

Osier, M. V., Cheung, K. H., Kidd, J. R., Pakstis, A. J., Miller, P. L. & Kidd, K. K. (2001) ALFRED: an allele frequency database for diverse populations and DNA polymorphisms–an update. *Nucleic Acids Res* **29**, 317–319.

Osier, M. V., Pakstis, A. J., Soodyall, H., Comas, D., Goldman, D., Odunsi, A., Okonofua, F., Parnas, J., Schulz, L. O. & Bertranpetit, J. *et al*. (2002a) A global perspective on genetic variation at the ADH genes reveals unusual patterns of linkage disequilibrium and diversity. *Am J Hum Genet* **71**, 84–99.

Osier, M. V., Cheung, K. H., Kidd, J. R., Pakstis, A. J., Miller, P. L. & Kidd, K. K. (2002b) ALFRED: An allele frequency database for anthropology. *Am J Phys Anthropol* **119**, 77–83.

Pakstis, A. J., Kidd, J. R. & Kidd, K. K. (2002) A reference distribution of Fst values for biallelic DNA markers. *Am J Hum Genet* 71 Suppl **4**, 371.

Patil, N., Berno, A. J., Hinds, D. A., Barrett, W. A., Doshi, J. M., Hacker, C. R., Kautzer, C. R., Lee D. H., Marjoribanks, C. & McDonough, D. P. *et al*. (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**, 1719–1723.

Peterson, R. J., Goldman, D. & Long, J. C. (1999a) Nucleotide sequence diversity in non-coding regions of ALDH2 as revealed by restriction enzyme and SSCP analysis. *Hum Genet* **104**, 177–187.

Peterson, R. J., Goldman, D. & Long, J. C. (1999b) Effects of worldwide population subdivision on ALDH2 linkage disequilibrium. *Genome Res* **9**, 844–852.

Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., Lavery, T., Kouyoumjian, R., Farhadian, S. F. & Ward, R. *et al*. (2001) Linkage disequilibrium in the human genome. *Nature* **411**, 199–204.

SantaLucia, J. Jr. (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc Natl Acad Sci USA* **95**, 1460–1465.

Shibuya, A. & Yoshida, A. (1988) Frequency of the atypical aldehyde dehydrogenase-2 gene (ALDH2(2)) in Japanese and Caucasians. *Am J Hum Genet* **43**, 741–743.

Stewart, M. J., Malek, K. & Crabb, D. W. (1996) Distribution of messenger RNAs for aldehyde dehydrogenase 1, aldehyde dehydrogenase 2, and aldehyde dehydrogenase 5 in human tissues. *J Investig Med* **44**, 42–46.

Templeton, A. R., Clark, A. G., Weiss, K. M., Nickerson, D. A., Boerwinkle, E. & Sing, C. F. (2000) Recombinational and mutational hotspots within the human lipoprotein lipase gene. *Am J Hum Genet* **66**, 69–83.

Tishkoff, S. A., Dietzsch, E., Speed, W., Pakstis, A. J., Kidd, J. R., Cheung, K., Bonne-Tamir, B., Santachiara-Benerecetti, A. S., Moral, P. & Krings, M. *et al*. (1996) Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science* **271**, 1380–1387.

Tishkoff, S. A., Goldman, A., Calafell, F., Speed, W. C., Deinard, A. S., Bonne-Tamir, B., Kidd, J. R., Pakstis, A. J., Jenkins, T. & Kidd, K. K. (1998) A global haplotype analysis of the myotonic dystrophy locus: implications for the evolution of modern humans and for the origin of myotonic dystrophy mutations. *Am J Hum Genet* **62**, 1389–1402.

Vasiliou, V. & Pappa, A. (2000) Polymorphisms of human aldehyde dehydrogenases. Consequences for drug metabolism and disease. *Pharmacology* **61**, 192–198.

Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K. & Wilson, A. C. (1991) African populations and the evolution of human mitochondrial DNA. *Science* **253**, 1503–1507.

Wang, N., Akey, J. M., Zhang, K., Chakraborty, R. & Jin, L. (2002) Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. *Am J Hum Genet* **71**, 1227–34.

Wright, S. (1969) *Evolution and the genetics of populations: the theory of gene frequencies.* Vol 2: *The theory of gene frequencies*. University of Chicago Press, Chicago.

Yoshida, A. (1984) Genetic polymorphisms of alcohol metabolizing enzymes related to alcohol sensitivity and alcoholic diseases. *Alcohol Alcohol* **29**, 693–696.

Yoshida, A., Ikawa, M., Hsu, L. C. & Tani, K. (1985) Molecular abnormality and cDNA cloning of human aldehyde dehydrogenases. *Alcohol* **2**, 103–106.

Zhao, H., Pakstis, A. J., Kidd, J. R. & Kidd, K. K. (1999) Assessing linkage disequilibrium in a complex genetic system. I. Overall deviation from random association. *Ann Hum Genet* **63**, 167–179.

Received: 26 February 2003
Accepted: 7 July 2003

© University College London 2004