

ORIGINAL INVESTIGATION

Raymond J. Peterson · David Goldman
Jeffrey C. Long

Nucleotide sequence diversity in non-coding regions of *ALDH2* as revealed by restriction enzyme and SSCP analysis

Received: 8 September 1998 / Accepted: 7 December 1998

Abstract The simultaneous analysis of closely linked nucleotide substitutions has recently become possible. However, it is not known whether the construction of molecular haplotypes will be a generally useful strategy for nuclear genes. Furthermore, whereas mobility-shift methods are widely used for the discovery of nucleotide substitutions, the yield of these methods has rarely been evaluated. This paper investigates these issues in non-coding regions of *ALDH2*, the gene that encodes aldehyde dehydrogenase 2 (*ALDH2*). Screening 20 Europeans, 20 native Americans, and 20 Asians by using restriction enzyme and single-strand conformation polymorphism (SSCP) analysis has revealed 16 variable sites. SSCP yields slightly fewer than the number of nucleotide substitutions predicted by the restriction enzyme digests. Estimates of nucleotide diversity are similar to those of other genes, suggesting that the pattern of polymorphism in *ALDH2* offers a preview of what can be expected in many human nuclear genes. Eight of the variable sites discovered here and four sites discovered by others have been genotyped in 756 people from 17 populations across five continents. An expectation-maximization method has been used to estimate haplotype states and frequencies. Only three haplotypes are common worldwide, and a fourth haplotype is common in, but private to, Asia. Although allele frequencies differ among sites, linkage disequilibrium is almost maximal across *ALDH2*. This suggests that haplotype construction at *ALDH2* is particularly successful. The *ALDH2* result, in conjunction with linkage disequilibrium results from oth-

er genes, indicates that haplotype construction will be a generally useful genomic strategy.

Introduction

The simultaneous analysis of closely linked DNA polymorphisms has recently been applied to diverse problems in human genetics. For example, the decay of linkage disequilibrium with physical distance has been used to aid the positional cloning of the genes for cystic fibrosis (Kerem et al. 1989) and diastrophic dysplasia (Hästbacka et al. 1992, 1994). Other studies have focused on nucleotide substitution rates in order to estimate the age and geographic location of the most recent common ancestor (MRCA) of non-recombining sequences. Mitochondrial data have been used to infer that the female MRCA lived in Africa 150,000–200,000 years ago (Cann et al. 1987; Vigilant et al. 1991) and Y chromosome data indicate that the male MRCA lived in Africa 125,000–188,000 years ago (Hammer 1995; Tavaré et al. 1997). β -Globin sequence data suggest that the MRCA of the non-recombining sequences lived in Africa some 800,000 years ago (Harding et al. 1997). As these examples illustrate, the simultaneous analysis of closely linked nucleotide substitutions is a promising general strategy.

Several issues need to be investigated before the potential of this strategy is fully known. First, no single method is best for the discovery of nucleotide substitutions. Although direct sequencing or chemical cleavage reveals all variable nucleotide positions, these methods are costly and labor-intensive (Aguadé et al. 1994). Restriction enzyme (*RE*) digests yield an unbiased estimate ($\hat{\pi}_{RE}$) of nucleotide diversity (π ; Hudson 1982), i.e., the average heterozygosity per nucleotide (Nei 1987), but they query only a small fraction of nucleotides. Moreover, mobility-shift methods, such as single-strand conformation polymorphism (SSCP; Orita et al. 1989), density gradient gel electrophoresis (Lerman et al. 1984), or denaturing high performance liquid chromatography (Huber et al. 1992), do not reveal all variable positions. However, they are

R. J. Peterson · D. Goldman · J. C. Long (✉)
Section on Population Genetics and Linkage,
Laboratory of Neurogenetics,
National Institute on Alcohol Abuse and Alcoholism,
National Institutes of Health, Park V Bldg. Rm. 451 MSC 8110,
12420 Parklawn Dr., Bethesda, MD 20892-8110, USA
Tel.: +1 443 5969, Fax: +1 301 443-8579,
e-mail: jcl@box-j.nih.gov

R. J. Peterson
Department of Anthropology,
Pennsylvania State University, USA

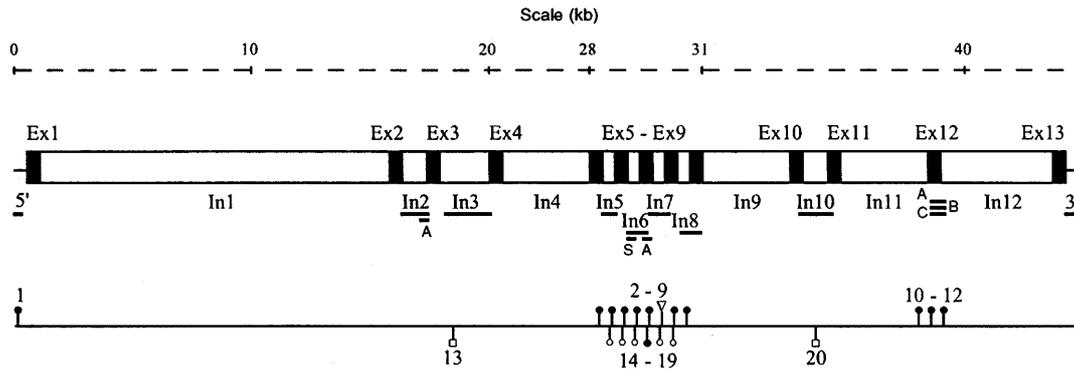


Fig. 1 *ALDH2* genomic structure, PCR products, and variable sites. Filled segments Exons, open segments introns, thick segments below names PCR products described in the text, numbers above the line variable sites genotyped in the worldwide survey, numbers below the line variable sites discovered in this study but not genotyped in the worldwide survey, filled circles nucleotide substitutions, upside-down triangle insertion, open circles SSCP sites, open squares restriction site polymorphisms

quicker, less costly, and query a large fraction of nucleotide positions (Sheffield et al. 1993).

An effective discovery strategy may be to combine restriction enzyme analysis with one of the mobility-shift methods, such as SSCP. An advantage of this approach is that an unbiased estimate of nucleotide diversity is fundamental to many population genetic analyses. For example, since estimates of nucleotide diversity are available for only a few nuclear genes, it is difficult to know how many nucleotides and chromosomes need to be searched in order to discover a target number of substitutions. Another advantage is that restriction enzyme data can be used to evaluate SSCP yield. This can be carried out by contrasting the number of nucleotide substitutions predicted from the restriction enzyme digests (Hudson 1982) with the number of SSCP sites that are discovered, or by contrasting $\hat{\pi}_{RE}$ with the nucleotide diversity estimated from the SSCP data ($\hat{\pi}_{SSCP}$). Finally, it is important to note that, if the combination of two well-known and easily applied methods is sufficient to construct a panel of closely linked nucleotide substitutions, then it should be feasible to accomplish the same goal with more sophisticated methods.

A second problem is that haplotype studies of nuclear genes have been limited because the phase of double heterozygotes is ambiguous. Recently, maximum-likelihood methods have been developed to estimate haplotype frequencies from diploid genes in random samples of people (Long et al. 1995; Excoffier and Slatkin 1995; Hawley and Kidd 1995). The utility of inferences with maximum likelihood depends on the extent of linkage disequilibrium and the ability to detect it (Thompson et al. 1988). Levels and patterns of linkage disequilibrium vary across the genome (Kwiatkowski et al. 1992; Watkins et al. 1994) and the power to detect this disequilibrium depends on a host of factors including allele frequencies and sample sizes. The limits of this approach with respect to sample sizes and nucleotide sequence length are presently unknown

and can only be fully revealed from knowledge of the genome structure.

We have investigated these issues in *ALDH2*, the gene for aldehyde dehydrogenase 2 (*ALDH2*). *ALDH2* oxidizes acetaldehyde to acetate and is the second enzyme in the two-step ethanol metabolic pathway. *ALDH2* is of interest, because it has a deficiency allele that is common in, but private to, Asia (Goedde et al. 1979; Teng 1981; Hsu et al. 1985). Because *ALDH2*-deficient people become ill when they consume alcohol (Wolff 1972), the deficiency allele is strongly protective against alcoholism (Harada et al. 1982). An investigation into the levels and patterns of polymorphism in *ALDH2* is an important first step toward a better understanding of the evolutionary forces that have operated on this gene. Additional advantages to studying *ALDH2* include known coding sequence, intron/exon boundaries, and approximate intron lengths (Fig. 1; Hsu et al. 1985, 1988). Seven introns, representing 20% of the gene, can be amplified by using standard polymerase chain reaction (PCR) conditions and primers complementary to the flanking exons. Comprising 13 exons across 44 kb (Hsu et al. 1988) of chromosome 12q24.2 (Ragunathan et al. 1988), the size of *ALDH2* is not atypical of many nuclear genes.

In this paper, an unbiased estimate ($\hat{\pi}_{RE}$) of nucleotide diversity in *ALDH2* non-coding regions has been obtained. Further, the *ALDH2* SSCP yield has been evaluated by contrasting $\hat{\pi}_{RE}$ with $\hat{\pi}_{SSCP}$, and by contrasting the number of nucleotide substitutions predicted by the restriction enzyme digest with the number of SSCP sites discovered. In addition, 12 *ALDH2* variable sites have been genotyped in 756 randomly sampled people from 17 populations across five continents. From these data, extended haplotypes have been estimated by using an expectation-maximization (E-M) method, and the extent of two-site linkage disequilibrium has also been investigated.

Materials and methods

Variable site search strategy

A search for *ALDH2* variable sites was performed in three population samples: 20 native Americans, 20 Europeans, and 20 Asians. The geographic diversity of the search sample was selected in order to reduce ascertainment bias with respect to ethnicity.

For our purposes, *ALDH2* has been defined as the transcribed segment of DNA from which *ALDH2* is produced. The interven-

Table 1 *ALDH2* PCR product names and primers (*In* intron, *Ex* exon)

Name	5'-Forward primer-3'	5'-Reverse primer-3'
5'	GCAGTGCCGTCTGCCCCATCCATGT	GGCCCGAGCCAGGGCGACCCTGAGCT
In2	CCCACCGTCAATCCGTCCACT	GCACTCACCGCCAGGTAGGTC
In2	ACATTGCTGAAGTCTGGTGCTC	TCACTGCCTTGTCCACATCTT
In3	CCGGACCTACCTGGCGGTGAG	CCGGAGACATTTGAGGACCAT
In5	CCCATTGACGGAGACTTCTT	CAGCTACCTTATCACAACCA
In6	TGGAATTTCCCGCTCCTGATG	GGCACAATGTTGACCACACCA
In6S	TGGAATTTCCCGCTCCTGATG	TGGGGTCTTGCTATGTTGTTC
In6A	AAATATTGCTCTAGGCCAGGC	TGGGAATTCTAAATGGGACGG
In7	GGACAAAGTGGCATTACAGG	TGATGATGTTGGGGCTCTTCC
In8	GGAAGAGCCCCAACATCATCA	GCTCCGCTCCACAAACTCATC
In10	TTTAAGAAGATCCTCGGTAC	CTTCCCAACAACCTCTCTA
Ex12A	CAAATTACAGGGTCAACTGCT	GCCCAACTCCCGCCAGTCCC
Ex12B	"	TCAGTGTATGCCTGCAGCCCGACT
Ex12C	"	CCACACTCACAGTTTTCTCTT
3'	CACAGTCAAAGTGCCTCAGAA	TAGCCCAGAATACAAAGCAGG

ing sequence was chosen because the goal was to discover polymorphic markers, and because introns are expected to harbor more variable sites than exons (Li and Sadler 1991). All seven of the PCR products were scanned by using a battery of 13 restriction enzymes. Four introns (In5–In8) were also intensively screened by using SSCP. To aid these analyses, the complete nucleotide sequence was determined for In5–In8. Hereafter, the PCR products of these introns are referred to as the reference PCR products.

To maximize SSCP sensitivity, two strategies were used. First, each reference PCR product was cut at different locations by using different restriction enzymes (Liu and Sommer 1994). Each digestion was separately electrophoresed through a standard gel matrix. Second, one restriction enzyme condition for each reference PCR product was electrophoresed through two additional matrices (Sheffield et al. 1993; Ravnik-Glavac et al. 1994).

Allelic variants detected by restriction enzymes and SSCP were confirmed by three methods: 1) direct nucleotide sequencing; 2) transmission in families; and 3) comparison of restriction enzyme digest patterns with the patterns predicted by the reference and variant nucleotide sequences.

Molecular methods

Polymerase chain reaction

Table 1 lists each PCR product by mnemonic and primer sequences. PCR products were amplified in a 15- μ l final volume with 40 ng genomic DNA, 0.33 μ M each primer, 0.250 mM dNTPs, 0.6 U AmpliTaq polymerase (Perkin Elmer), and either 1 \times PCR buffer (Perkin Elmer) or 1 \times Taq Extender PCR buffer and 0.6 U proprietary enzyme (Stratagene). The GeneAmp 9600 (Perkin Elmer) standard thermal cycling profile was modified as needed for each PCR product (Peterson 1998).

Restriction enzymes

The battery of restriction enzymes included *AluI*, *AvaII*, *BstUI*, *DdeI*, *DpnII*, *EcoRI*, *HaeIII*, *HhaI*, *HindIII*, *HinfI*, *MspI*, *RsaI*, and *TaqI* (NEB). For each digestion, 15 μ l PCR product was digested with 15 U restriction enzyme at a final volume of 25 μ l following the manufacturer's instructions. Digested PCR products were electrophoresed through 2% agarose with 1 \times TBE buffer (TBE buffer = 0.09 M TRIS-borate, 0.002 M EDTA, pH 8.3) and sized by using DNA ladders of known length. Migration patterns were visualized under UV-light after ethidium-bromide staining.

SSCP analysis

For SSCP analysis, the PCR final volume was scaled down to 5 μ l. The addition of 2.25 μ Cl of 33 P dNTPs during the PCR allowed the incorporation of radioactive label, and 5 U restriction enzyme were used in a final volume of 10 μ l. The electrophoretic loading buffer comprised 95% formamide, 10 mM NaOH, 0.05% xylene cyanol, and 0.05% bromophenol blue. Buffered samples were denatured at 94°C for 3 min, and 4 μ l of each sample was loaded into each gel lane. Standard electrophoretic conditions were 1 \times MDE gel, 0.6 \times TBE at 500 V for 12–20 h in a 4°C cold-room. In addition, two other gel matrices were used: 1 \times MDE gel with 10% glycerol, or 10% polyacrylamide gel with a 50:1 polyacrylamide:bis ratio and 10% glycerol. After electrophoresis, each gel was blotted, dried, and exposed to autoradiography film at –70°C for 2–6 days.

Nucleotide sequencing

PCR products were purified by using Centricon-100 (Amicon) spin columns, and both strands were cycle sequenced with the ABI Prism Dye Terminator Ready Reaction Kit following the manufacturer's instructions. The resulting sequencing ladders were purified on Centri-Sep (Princeton Separations) spin columns and electrophoresed on an ABI 373A DNA Sequencer. Data collection and sequence analysis were performed by using ABI software.

Worldwide population survey

Populations

DNA samples were collected and provided by a number of investigators. These samples have been described elsewhere; Table 2 lists citations that provide more details. Briefly, the Biaka, who were sampled in the Central African Republic, represent Africans. Asian samples include Cambodian Khmers who live in Los Angeles and Chinese Han who live in San Francisco or New Haven. Japanese were collected in the USA, South Koreans were collected in Seoul, and Taiwanese of Chinese descent were collected in Taipei. Black Thai are ethnic Khmers and were collected in Nakhon Pathom Province, Thailand. Europeans are represented by 16 pairs of French and Utah Centre d'Etude Polymorphisme Humain (CEPH) parents, whereas Finns and Swedes are cosmopolitan samples collected in Helsinki and Stockholm, respectively. Cheyenne were ascertained through the Cheyenne-Arapaho tribal enrollment list and verified by voter registration records. All other native North American samples (Pima, Maya, and Navajo) and native South American samples (Karitiana, Rondonian Surui, and Ti-

Table 2 Populations used in the worldwide survey

Continent	Population	No.	Source	Citation
Africa	Biaka	51	Cavalli-Sforza/Kidd	Bowcock et al. 1987
Asia	Cambodians	24	Dumars	Barr and Kidd 1993
	Chinese	47	Cavalli-Sforza/Kidd	Bowcock et al. 1987
	Japanese	49	Cavalli-Sforza/Kidd	Barr and Kidd 1993
	Korea	40	Park/Song	A.W. Bergen et al., in preparation
	Taiwanese	43	Tsai	Novoradovsky et al. 1995
Europe	Black Thai	50	Chandanayingyong	Chandanayingyong et al. 1994
	CEPH	32	Dausset	Dausset et al. 1990
	Finns	41	Linnoila	Urbanek et al. 1996
N. America	Swedes	45	Taskman	Urbanek et al. 1996
	Cheyenne	51	Goldman	Urbanek et al. 1996
	Mayan	50	Weiss	Kidd et al. 1991
S. America	Navajo	46	Long	Urbanek et al. 1996
	Pima	45	Goldman	Urbanek et al. 1996
	Karitiana	49	Black	Kidd et al. 1991
	R. Surui	44	Black	Kidd et al. 1991
	Ticuna	49	Wallace/Lawrence	Neel et al. 1980

Table 3 Variable site genotyping conditions (*RSP* restriction site polymorphism)

Site	Location ^a	Genotype method	PCR product ^b	Restriction enzyme	Fragment length	
					Reference	Variant
1	G-355A	RSP	5'	<i>Sac</i> I	113, 21	134
2	T348C	RSP	In6A	<i>Msp</i> I	40	1251, 150
3	T483C	RSP	In6A	<i>Hae</i> III	342, 44, 15	226, 116, 44, 15
4	A48G	SSCP	In8	<i>Alw</i> NI	–	–
5	G52C	SSCP	In8	<i>Alw</i> NI	–	–
6	G69A	RSP	In8	<i>Dpn</i> II	485, 104	589
7	C79CC	SSCP	In8	<i>Alw</i> NI	–	–
8	G316C	SSCP	In8	–	–	–
9	G910A	SSCP	In8	–	–	–
10	G1464A	SSCP	Ex12C	–	–	–
11	G1486A	RSP	Ex12B	<i>Hin</i> II	90, 25	115
12	G1519A	SSCP	Ex12C	–	–	–

^a For 5' flanking and exons, +1=A in cDNA start codon; for introns, +1= the first nucleotide of the intron; the reference nucleotide is left

^b 5' 5' flanking sequence, *Ex* exon, *In* intron

cuna) were collected either from a village, reservation, or nearby metropolitan hospital.

Genotyping

Twelve variable sites were genotyped (Table 3), including the site of the Asian deficiency allele (Yoshida et al. 1984). Sites 1 and 11 were genotyped as artificial restriction site polymorphisms (RSPs) created by designing single base mismatches into the PCR primers. Sites 2, 3, and 6 were genotyped as naturally occurring RSPs. Sites 4, 5, 7, 8, and 9 were genotyped as naturally occurring SSCP. Sites 10 and 12 were genotyped as artificial SSCPs by incorporating a single base mismatch into a PCR primer.

Statistical analyses

Expected number of substitutions and nucleotide diversity

The data used in these analyses were from the reference PCR products of the 60 people in the search sample, for which a complete reference nucleotide sequence had been determined. Estimates of nucleotide diversity, which can depend on allele frequency, were

obtained for each population separately and for the pooled sample. The proportion of predicted nucleotide substitutions (p) was estimated from the restriction enzyme digests by Hudson's (1982) method, $\hat{p} = k/(2m - k)j$, where k is the number of variable cut sites, m is the total number of cut sites, and j is the average recognition sequence length. The evolutionary variance of \hat{p} is \hat{p}^2/k . The expected number of nucleotide substitutions was obtained by multiplying p by $L = 2250$, the length of the reference PCR products (less primers).

An unbiased estimate ($\hat{\pi}_{RE}$) of nucleotide diversity (π) was obtained by using the restriction enzyme data: $\hat{\pi}_{RE} = \hat{p} / \ln(n)$, where n is the number of sampled chromosomes (e.g., Hudson 1982). The evolutionary variance of $\hat{\pi}_{RE}$ is $(\hat{\pi}_{RE})^2/k$. Nucleotide diversity was also estimated from the SSCP data ($\hat{\pi}_{SSCP}$). $\hat{\pi}_{SSCP}$ is only an unbiased estimator of π under the assumption that SSCP discovers all substitutions and that each SSCP is attributable to a single substitution. $\hat{\pi}_{SSCP}$ was estimated by using Nei and Tajima's (1981) method, viz., $\hat{\pi}_{SSCP} = n/(n-1) \sum_{i \neq j} x_i x_j \hat{\pi}_{ij}$. Here, x_i and x_j are the frequencies of the i th and j th unique haplotypes and $\hat{\pi}_{ij}$ is an estimate of the proportion of site-differences among them. The evolutionary variance of $\hat{\pi}_{SSCP}$ is $1/3L\hat{\pi}_{SSCP} + 2f/9(\hat{\pi}_{SSCP})^2$.

Knowledge of the percentage of nucleotides queried in the reference PCR products enhanced the contrast of the restriction en-

zyme and SSCP analyses. For the restriction enzymes, this percentage was calculated by summing the recognition lengths associated with each cut site, adjusting for overlapping sites, and dividing by L . For SSCP, it is believed that each SSCP condition reveals 70%–90% of variable sites (Ravnik-Glavac et al. 1994).

Allele frequencies, haplotype estimation and linkage disequilibrium

Each genotyped site had two co-dominant alleles. For each site, the allele with the higher worldwide frequency was assigned to be the reference allele. Allele frequencies in each sample were determined by direct gene counting. The genotype distribution for each site in each sample was evaluated for departure from the Hardy-Weinberg equilibrium by using a contingency table χ^2 test (Weir 1996).

Since phase-unknown genotypes were collected, haplotype states and frequencies were estimated by maximum likelihood with an E-M method (Dempster et al. 1977; Long et al. 1995). Two approaches were used. First, the E-M method was applied, and all haplotype states that had a frequency greater than 1/200 were retained. This arbitrary frequency limit was appropriate given that sample sizes were ~ 100 chromosomes. Next, a three-step procedure was used. First, the multi-site genotype data were examined, and all directly observed haplotypes were tabulated. A directly observed haplotype was one that occurred in at least one multi-site homozygote or single-site heterozygote. Second, each multi-site heterozygote was examined to determine whether some pair of observed haplotypes could explain it. The third step was to estimate haplotype state and frequency by constraining the E-M method to allow only the directly observed haplotypes or by introducing a set of new haplotypes that required the fewest changes by mutation or recombination from the directly observed set. The standard errors of the haplotype frequency estimates were obtained by using a jackknife procedure (Weir 1996).

Within each population, the two-site haplotype frequencies were obtained from the 12-site haplotype frequencies by summing frequencies of all haplotypes with each specific combination of alleles at the two sites. The two-site linkage disequilibrium parameters were estimated as $D_{A_iB_i} = P_{A_iB_i} - p_{A_i}q_{B_i}$ (Weir 1996). For convenience, D will hereafter be given without any subscripts when the argument pertains to a pair of alleles at any two sites. Because D depends on allele frequency (Lewontin 1964), the effects of allele frequency were reduced by normalizing D to the maximum that it could have been, given the allele frequencies (Lewontin 1964). This measure, D' , is D/pq when $D < 0$ and $D/\min((1-p)q, p(1-q))$ when $D > 0$. D' is a useful quantitative measure of the extent of pair-wise linkage disequilibrium. Moreover, a D' value between -1.0 and $+1.0$ indicates the presence of a recombinant haplotype.

Results

Variable sites

Description:

Twenty *ALDH2* variable sites are now known (Fig. 1). Site 1 was discovered by M. Stewart (personal communication), sites 10 and 11 were previously discovered in our lab (Novoradovsky et al. 1995), site 12 is the well-known deficiency allele (Yoshida et al. 1984), and the remaining sixteen sites were discovered during this study. In the search sample and in the reference PCR products, SSCP detected the variants at sites 5, 16, 17, 18, and 19, whereas both SSCP and the restriction enzymes detected the variants at sites 2, 3, and 6. Sites 14 and 15 were detected by restriction enzymes in PCR products In3 and In10, respectively. In the worldwide survey, SSCP genotyping of

variable sites in PCR product In8 revealed sites 4, 7, 8, 9, 13, and 20.

Nucleotide sequencing revealed that sites 2, 3, 4, 5, 6, 8, and 9 were attributable to nucleotide substitution and that site 7 resulted from a single base insertion. Only site 7 could not be confirmed by restriction enzyme analysis. The variants at sites 1, 2, 3, and 6 were confirmed by transmission. For example, in one family, one parent was (A_2, A_2), (B_1, B_1), (C_1, C_1), and (F_2, F_2) and the other was (A_1, A_1), (B_2, B_2), (C_2, C_2), and (F_1, F_1). As expected, all eight genotyped children were heterozygous at each site. The nucleotide sequence was obtained for introns In5–In8. The respective GenBank accession numbers are AF073511–AF073514.

Sites 2–8 were located in the intervening sequence. Site 9 was located in the coding sequence and was predicted to change the amino acid at position 287 of the polypeptide from valine to methionine. This is the third amino acid substitution recorded for *ALDH2*.

Expected number of substitutions and nucleotide diversity

The battery of restriction enzymes had an average recognition length (j) of 4.09 bp, and computer analysis of the reference nucleotide sequences revealed 94 (m) cut sites. Three cut sites were polymorphic, yielding $k = 3$. Since these sites were polymorphic in all three samples (our variable sites 2, 3, and 6), the three samples yielded identical estimates of p and π . The expected proportion of substitutions (\hat{p}) was estimated to be 0.0040 ± 0.0023 , whereas the number of nucleotide substitutions that the search sample was expected to harbor in the reference PCR products ($L^* \hat{p}$) was estimated to be 8.9 ± 5.2 . Nucleotide diversity ($\hat{\pi}_{RE}$) was estimated to be 0.0008 ± 0.0005 . Hudson's equation does not account for allele frequency. Because of this, the equation probably masks differences in $\hat{\pi}_{RE}$ among samples, because of allele frequency differences among the samples. However, since allele frequencies at sites 2, 3, and 6 were similar among samples, it appeared that Hudson's equation masked only minor differences in $\hat{\pi}_{RE}$.

The equation for $\hat{\pi}_{SSCP}$ accounts for allele frequency. SSCP discovered variable site 5, in addition to variable sites 2, 3, and 6. The variant at this site was completely linked to the deficiency allele and, thus, was common in, but private to, Asia. Because of this, the Asian sample yielded the highest estimate of $\hat{\pi}_{SSCP}$ at 0.0007 ± 0.0002 . Europeans were somewhat lower at 0.0005 ± 0.0002 , whereas native Americans were lower still at 0.0003 ± 0.0002 . Pooling these three samples yielded $\hat{\pi}_{SSCP} = 0.0005 \pm 0.0003$.

Worldwide population survey

Allele and haplotype frequencies

The worldwide allele frequencies for sites 1–12 (Table 4) assume that each RSP and SSCP is attributable solely to

Table 4 Frequency of the variant allele ($\times 1000$) at 12 sites in 17 worldwide populations

Populations	N	Site											
		1	2	3	4	5	6	7	8	9	10	11	12
Biaka	102	98	225	225	59	0	235	0	0	0	0	0	0
Africa ^a	102	98	225	225	59	0	235	0	0	0	0	0	0
Cambodian	48	146	188	188	0	146	188	21	0	0	0	0	146
Chinese	94	106	170	170	0	309	170	0	0	0	0	0	298
Japanese	98	163	102	102	0	276	102	0	0	0	0	0	286
S. Korean	80	162	175	175	0	113	175	0	0	0	0	0	113
Taiwanese	86	116	233	233	0	267	233	0	0	0	0	12	267
Black Thai	100	140	230	230	0	60	230	0	0	0	0	0	60
Asia ^a	506	138	182	182	0	200	182	2	0	0	0	2	200
CEPH	64	766	219	219	0	0	219	0	0	0	0	0	0
Finn	82	793	195	195	0	0	195	0	0	0	0	0	0
Swede	90	844	144	144	0	0	144	0	0	0	0	0	0
Europe ^a	236	805	182	182	0	0	182	0	0	0	0	0	0
Cheyenne	102	647	88	88	0	0	88	0	0	2	0	0	0
Mayan	100	510	110	110	0	0	110	0	0	0	0	0	0
Navajo	92	815	76	76	0	0	76	0	0	0	0	0	0
Pima	90	556	233	233	0	0	233	0	0	0	56	0	0
N. America ^a	384	630	125	125	0	0	125	0	0	5	13	0	0
Karitiana	98	571	194	194	0	0	194	0	20	0	0	0	0
R. Surui	88	943	11	11	0	0	11	0	0	0	0	0	0
Ticuna	98	541	235	235	0	0	235	0	0	0	0	0	0
S. America ^a	284	676	151	151	0	0	151	0	7	0	0	0	0
World ^b	1512	465	165	165	4	67	165	1	1	1	3	1	67

^a Continental total (N) or average (site)

^b World-wide total (N) or average (site)

the DNA variant given in Table 3. The standard errors of the allele frequency estimates can be calculated from the binomial distribution as $\sqrt{p_i(1-p_i)/2n_i}$, where n_i is the number of people in the i th sample. Sites 13–20 were not genotyped in the worldwide sample. Five of these variants were observed in single copy, C252T in In8 co-occurred with the Biakan-specific variant at site 4, and an *RsaI* site in In3 co-occurred with the variants at sites 2, 3, and 6. These variants added too little information to justify proper confirmation and genotyping. The haplotype frequencies, states, and associated jackknife standard errors are given in Table 5. Since there was an almost one to one correspondence between variant allele frequency and haplotype frequency, it was convenient to present the allele frequencies in the context of the haplotype states and frequencies. Further, the unconstrained and constrained E-M procedures yielded nearly identical results. Because of this, the haplotype results are presented in the context of the constrained procedure, and the exceptions are noted.

Worldwide, eleven haplotype states were directly observed, and two additional haplotype states were inferred by the maximum-likelihood estimation procedure. Unless otherwise specified, all haplotypes were assumed to have been directly observed. For brevity, a number and the letter H designate each haplotype state.

Only H1, H2, and H3 were common worldwide. H1 carried the reference allele at all twelve sites and ranged in frequency from 1% in Swedes to 61% in African Biaka.

H2 differed from H1 by carrying the variant at site 1. Since H2 haplotypes accounted for all but eight copies of the variant at site 1, the geographic distribution of H2 and the variant at site 1 were nearly identical, i.e., H2 and the variant at site 1 ranged in frequency from 10% in the African Biaka to 94% in the South American R. Surui, H3 differed from H1 by carrying the variants at sites 2, 3, and 6, and H3 accounted for all but two copies of the variants at sites 2 and 3 and three copies of the variant at site 6. Thus, the geographic distribution of H3 and the variants at sites 2, 3, and 6 were nearly identical. Specifically, H3 and the variants at sites 2, 3, and 6, varied in frequency from 1% in the R. Surui to 23% in the Taiwanese of Chinese descent.

H4, which carried the variant at site 5 and the deficiency allele, was private to Asia. H4 haplotypes accounted for all but two copies of the variant at site 5 and all but two copies of the deficiency allele. Thus, the geographic distribution of H4, the variant at site 5, and the deficiency allele was almost identical. H4, the variant at site 5, and the deficiency allele attained a frequency of 27% to 30% in the Chinese, Taiwanese of Chinese descent, and Japanese.

In the Biaka, H5 carried the variant at site 4, and H6 carried the variant at site 6, but not at sites 2 and 3. H6 may have arisen by mutation or by recombination between H1 and H3. H7 carried the variant at site 7 and was inferred to be present only in the Cambodians. Chinese

Table 5 Estimated frequency of unique haplotypes ($\times 1000$) and jackknife standard errors in 17 worldwide populations

Populations	2N	Haplotypes ^a												
		H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	H12	H13
Biaka	102	608 (52)	98 (32)	225 (40)	0 (0)	59 (26)	10 (10)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Africa ^b	102	608 (48)	98 (29)	225 (41)	0 (0)	59 (23)	10 (10)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Cambodian 48	542	125 (114)	167 (92)	146 (81)	0 (48)	0 (0)	20 (0)	0 (21)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Chinese	94	415 (49)	106 (37)	170 (39)	298 (47)	0 (0)	0 (0)	0 (0)	11 (11)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Japanese	98	449 (51)	163 (36)	102 (28)	276 (46)	0 (0)	0 (0)	0 (0)	0 (0)	10 (10)	0 (0)	0 (0)	0 (0)	0 (0)
S. Korean	80	550 (60)	163 (46)	175 (40)	112 (36)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Taiwanese	86	384 (52)	116 (35)	232 (46)	256 (47)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	12 (12)	0 (0)	0 (0)	0 (0)
Black Thai	100	570 (51)	140 (35)	230 (46)	60 (23)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Asia ^b	506	479 (22)	135 (15)	178 (17)	199 (18)	0 (0)	0 (0)	1 (2)	2 (2)	2 (2)	2 (2)	0 (0)	0 (0)	0 (0)
CEPH	64	16 (16)	766 (54)	218 (49)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Finn	82	12 (13)	793 (47)	195 (47)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Swede	90	11 (11)	844 (34)	145 (36)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Europe ^b	236	13 (7)	802 (26)	185 (25)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Cheyenne	102	265 (49)	647 (53)	69 (28)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	19 (14)	0 (0)	0 (0)
Mayan	100	380 (52)	510 (55)	110 (32)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Navajo	92	109 (34)	815 (39)	76 (27)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Pima	90	211 (43)	500 (53)	233 (44)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	56 (24)	0 (0)
N. America ^b	384	246 (22)	617 (25)	119 (17)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	5 (4)	13 (6)	0 (0)
Karitiana	98	235 (37)	551 (47)	194 (40)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	20 (19)
R Surui	88	45 (23)	943 (23)	12 (12)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Ticuna	98	224 (34)	541 (44)	235 (41)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
S. America ^b	284	172 (22)	671 (28)	151 (21)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	6 (5)
World	1512	298 (12)	460 (13)	162 (9)	67 (6)	3 (2)	1 (1)	1 (1)	1 (1)	1 (1)	1 (1)	1 (1)	2 (1)	1 (1)

^a With '1' the reference allele, the haplotype configurations are: H1: 111111111111; H2: 211111111111; H3: 122112111111; H4: 111121111112; H5: 111211111111; H6: 111112111111; H7: 111112111111 or 122112211111; H8: 111121111111; H9:

111111111112; H10: 111121111122; H11: 122112112111 or 111111112111; H12: 211111111211; H13: 211111121111

^b Continental total (N) or average (site)

^c World-wide total (N) or average (site)

H8 carried the variant at site 5 but not the deficiency allele. Japanese H9 carried the deficiency allele but not the variant at site 5. H8 and H9 could have arisen by mutation or recombination between H1 and H4. H10, observed in the Taiwanese of Chinese descent, carried the variant at site 11. H11, inferred to be present only in the Cheyenne, carried the variant at site 9. H12 and the variant at site 10

were observed in the Pima. Karitiana H13 carried the variant at site 8.

Only H7 and H11 were not directly observed. These haplotypes accounted for just three of the 1512 chromosomes and represented the discrepancies between the unconstrained and constrained procedures. Whereas the existence of H7 was clear from the presence of the variant at

site 7, the Cambodian carrying this variant was otherwise an H2/H3 heterozygote. Because of this, it could not be determined whether the variant occurred on H2 or H3, and both haplotype states are given. Similarly, the variant at site 9 that defines H11 occurred in two Cheyenne who were otherwise H1/H3 heterozygotes. Here too, both haplotype states are given to reflect that the variant at site 9 occurred on either an H1 or H3 background.

Genotypes

The number of segregating sites in each population ranged from four to seven. The χ^2 test for departure from the single-site Hardy-Weinberg expectation was applied to each segregating site in each population. Altogether, 86 tests were performed. Four tests had p -values of less than 5%. These tests lacked independence because of the correlation of alleles among sites. Despite this, it is reasonable to conclude that this number of departures from the Hardy-Weinberg expectation reflects sampling fluctuation under the null hypothesis.

Linkage disequilibrium analysis

Remarkably, D' was -1.0 or $+1.0$ for all two-site comparisons. Whereas maximal D' values indicate high levels of linkage disequilibrium, they do not mean that evidence of recombination is completely absent. H6, H8, and H9 were identified as possible recombinants. The D' analysis did not identify H6 in the Biaka, H8 in the Chinese, and H9 in the Japanese as recombinants, because only three of the four possible two-site haplotypes segregated in each population. The finding that H6, H8, and H9 accounted for just three of the 1512 chromosomes is consistent with a low rate of recombination across *ALDH2*.

Discussion

The scanning of the reference PCR products in the search sample, by using both restriction enzymes and SSCP, allows the evaluation of the SSCP yield. The restriction enzymes have queried 14.6% of the nucleotides, revealed sites 2, 3, and 6, and suggested that the search sample harbors 8.9 ± 5.2 substitutions. SSCP has revealed sites 2, 3, 5, 6, 16, 17, 18, and 19. Eight discovered SSCPs lie well within the confidence interval of $\hat{\pi}_{RSP}$. Differences in $\hat{\pi}_{RSP}$ among native Americans, Europeans, and Asians, masked by Hudson's equation, are revealed by $\hat{\pi}_{SSCP}$. However, these differences are small in contrast to the large evolutionary variance associated with the point estimates.

$\hat{\pi}_{SSCP}$ is lower than $\hat{\pi}_{RSP}$, perhaps because SSCP has not identified all nucleotide substitutions or because an SSCP allele may have been attributable to more than one substitution. In any event, since the confidence intervals of $\hat{\pi}_{RSP}$ and $\hat{\pi}_{SSCP}$ considerably overlap, the SSCP effort appears to have well-characterized the reference PCR products.

More generally, SSCP appears to be an effective tool for the discovery of substitutions. This same conclusion has been reached by Aguadé et al. (1994) as a result of stratified nucleotide sequencing of 36 SSCP allelic classes in the *su(s)* and *su(w^o)* genes in *Drosophila melanogaster*. This analysis has revealed just one nucleotide substitution cryptic to SSCP.

Nucleotide diversity in *ALDH2* non-coding regions, at 0.0008 ± 0.0005 , is similar to that of several nuclear genes. For example, direct sequencing has revealed $\hat{\pi} = 0.0009 \pm 0.0005$ in β -globin introns (Fullerton et al. 1994) and 0.0011 ± 0.0004 in four-fold degenerate coding positions (Li and Sadler 1991). Direct sequencing of 8700 bp of the non-coding sequence in 71 people at the *LPL* gene has revealed a nucleotide diversity of 0.0020 ± 0.0010 (Nickerson et al. 1998). This level of variation is not unexpected given the large evolutionary variance. It appears, then, that the pattern of polymorphism in *ALDH2* may offer a preview of what can be expected in many human nuclear genes.

One critical issue is that for haplotype construction; it is desirable to have several variable sites with appreciable levels of polymorphism. Here, it is possible to estimate the expected number of such polymorphisms harbored in the search sample. The finding that three of the nucleotide substitutions in the 2250 bp of the reference PCR products reach polymorphic frequencies predicts, across the 44 kb of *ALDH2*, the existence of 59 such polymorphic sites. To the extent that this typifies nuclear genes, we expect that sufficient polymorphism exists for haplotype construction in many genes.

Another issue, with regards to evolutionary studies, is the adequacy of the search sample. This can be evaluated, because the use of SSCP in In8 to genotype variable sites in the worldwide survey fortuitously provides an expanded search sample, albeit for a shorter DNA segment: 547 bp in 756 people from five continents, as opposed to 2250 bp in 60 people from three continents. The worldwide survey has revealed the same In8 SSCPs (sites 5 and 6) that have been discovered in the search sample. In addition, it has revealed a population-specific variant that barely attained a 5% frequency (site 4) and five rare variants (sites 7, 8, 9, 13, and 20). Thus, screening more people in more populations essentially only yields additional rare variants. This result can be attributed to the finding that sequences from different human populations share a considerable portion of their ancestral histories (Pluzhnikov and Donnelly 1996).

Pluzhnikov and Donnelly's (1996) elegant cost-benefit analysis of the trade-off between searching nucleotides (L) and chromosomes (n) suggests that our search strategy was appropriate. These authors have sought to minimize the normalized sampling variance of nucleotide diversity by computer simulating the effects of increasing L or decreasing n for a given nL . In a panmictic population, normalized sampling variance is minimized when only 5–10 chromosomes are sampled for long L . However, for subdivided populations, these authors recommend searching fewer nucleotides in more chromosomes and the sampling of chro-

mosomes from each population. For *ALDH2*, the strategy has been to sample chromosomes from populations representing continental divisions. The inclusion of more Africans in the search sample may have enhanced the search efforts with respect to diversity (Tishkoff et al. 1996, 1998).

The finding that the search sample was appropriate provides the basis for further evolutionary studies of *ALDH2*. In particular, it is interesting to determine whether natural selection has maintained the deficiency allele in Asia (Ikuta et al. 1986; Goldman and Enoch 1990). One way to approach this question is to investigate whether the neutral apparent age of the deficiency allele, estimated by using coalescence methods, is older than its restricted geographic distribution would suggest (R. J. Peterson et al., in preparation).

With respect to the genotyping methods used in the worldwide survey, these have proved to be adequate. In the future, emerging DNA technologies are likely to automate the mass discovery and genotyping of nucleotide substitutions greatly. These technologies include the miniaturization of PCR and electrophoretic equipment (Burke et al. 1997), DNA chips (Pease et al. 1994; Chee et al. 1996), and mass spectrometry (Ch'ang et al. 1995).

Considerable variation among populations was evident in the allele and haplotype frequency distributions. In particular, the frequency of the variant at site 1 ranged from 10% in the African Biaka to 94% in the South American R. Surui. Partly because of this, haplotype H1 ranged in frequency from 1% in the Swedes to 61% in the Biaka, and haplotype H2 ranged in frequency from 10% in the Biaka to 94% in the R. Surui. These differences are interesting and are the subject of a subsequent paper on the effects of population subdivision on worldwide patterns of *ALDH2* linkage disequilibrium (R. J. Peterson et al., in preparation).

The haplotype states and frequencies used in this study were estimated statistically. Eleven of the 13 haplotypes, including all four of the widely distributed haplotypes, were directly observed in multi-site homozygotes or single-site heterozygotes. The two inferred haplotypes accounted for just 3 of the 1512 chromosomes. Thus, the haplotype estimates appear to be robust. For this study, transmission could not be directly observed for all haplotypes, because families were unavailable for most samples. Further, the large number of chromosomes made single molecule dilution impractical (Ruano et al. 1990), and the distance between sites is too great for allele-specific amplification and long-range PCR (Fullerton et al. 1994). Although 13 unique haplotypes have been estimated, only four have a high frequency in the worldwide data set. The small number of estimated *ALDH2* haplotypes indicates strong linkage disequilibrium. Indeed, the D' and related analysis suggests that recombination was nearly absent. The genotyping of more African samples, especially in conjunction with genotyping the In8 C252T site and the In3 *RsaI* site, may reveal less linkage disequilibrium in Africans (Tishkoff et al. 1996, 1998). Even if this occurs, the high levels of *ALDH2* linkage disequilibrium suggest that analysis of extended *ALDH2* haplotypes should be particularly informative.

The low number of observed recombinant chromosomes seems unusual in that a much larger proportion of chromosomes is usually recombinant (Harding et al. 1997). However, complete linkage disequilibrium across 40 kb is not unprecedented. In the *NF1* gene, complete association has been observed between alleles at sites separated by 80 kb (Jorde et al. 1993).

Furthermore, complete linkage disequilibrium across 40 kb is not mandatory in order for haplotype construction to be generally useful as demonstrated by several analyses of genes with less than complete linkage disequilibrium. These genes include the β -globin gene cluster (Wainscoat et al. 1986; Long et al. 1990), *CD4* (Tishkoff et al. 1996), *DRD2* (Castiglione et al. 1995), *NF1* (Purandare et al. 1996), and *APC* (Jorde et al. 1994). Importantly, linkage disequilibrium analyses of these gene regions have been profitably applied to a diverse set of questions, including those of human evolution and demographic history. Together, this predicts that construction of molecular haplotypes will be a generally useful strategy for nuclear genes.

Acknowledgements We should like to thank an anonymous reviewer for comments that improved this manuscript. We thank Longina Akhtar for maintaining the Laboratory of Neurogenetics cell lines. Ken Kidd, Su-Jen Tsai, Dr. Chandanayingyong, and Dr. Park kindly provided DNA samples, and Mark Stewart provided details on the variant at site 1. Margrit Urbanek, Andrew Bergen, and Jaakko Lappalainen contributed invaluable advice on laboratory techniques and useful discussions. Ken Weiss, Andy Clark, Mark Stoneking, and Henry Harpending provide helpful comments on earlier drafts. This research complies with current USA laws and was supported by a National Institute of Alcohol Abuse and Alcoholism pre-doctoral Intramural Research Training Award to R.J.P.

References

- Aguadé M, Meyers W, Long AD, Langley CH (1994) Single-strand conformation polymorphism analysis coupled with stratified DNA sequencing reveals reduced sequence variation in the *su(s)* and *su(w^d)* regions of the *Drosophila melanogaster* X chromosome. *Proc Natl Acad Sci USA* 91:4658–4662
- Barr C, Kidd KK (1993) Population frequencies of the A1 allele at the dopamine D2 receptor locus. *Biol Psychiatry* 34:204–209
- Bowcock AM, Bucci C, Hebert JM, Kidd JR, Kidd KK, Friedlander JS, Cavalli-Sforza LL (1987) Study of 47 DNA markers in five populations from four continents. *Gene Geogr* 1:47–64
- Burke DT, Burns MA, Mastrangelo C (1997) Microfabrication technologies for integrated nucleic acid analysis. *Genome Res* 7:189–197
- Ch'ang L-Y, Tang K, Schell M, Ringelberg C, Matteson KJ, Allman SL, Chen CH (1995) Detection of $\Delta F508$ mutation of the cystic fibrosis gene by matrix-assisted laser desorption/ionization mass spectrometry. *Rapid Commun Mass Spectrometry* 9:772–774
- Cann RL, Stoneking M, Wilson AC (1987) Mitochondrial DNA and human evolution. *Nature* 325:31–36
- Castiglione CM, Deinard AS, Speed WC, Sirugo G, Rosenbaum HC, Zhang Y, Grandy DK, Grigorenko EL, Bonne-Tamir B, Pakstis AJ, Kidd JR, Kidd KK (1995) Evolution of haplotypes at the *DRD2* locus. *Am J Hum Genet* 57:1445–1456
- Chandanayingyong D, Stephens HA, Fan L, Sirikong M, Longta P, Vangserathana R, Lekmak S, Longta K, Bejrachandra S, Rungruang E (1994) HLA-DPB1 polymorphism in the Thais of Southeast Asia. *Hum Immunol* 40:20–4

- Chee M, Yang R, Hubbell E, Berno A, Huang XC, Stern D, Winkler J, Lockhart DJ, Morris MS, Fodor SP (1996) Accessing genetic information with high-density DNA arrays. *Science* 274: 610–614
- Dausset J, Cann H, Cohen D, Lathrop M, Lalouel J-M, White R (1990) Centre d'Etude du Polymorphisme Humain (CEPH): collaborative genetic mapping of the human genome. *Genomics* 6:575–577
- Dempster AP, Laird NM, Rubin DB (1977) Maximum-likelihood from incomplete data via the EM algorithm. *J R Stat Soc B* 39:1–38
- Excoffier L, Slatkin M (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol* 12:921–927
- Fullerton SM, Harding RM, Boyce AJ, Clegg JB (1994) Molecular and population genetic analysis of allelic sequence diversity at the human β -globin locus. *Proc Natl Acad Sci USA* 91:1805–1809
- Goedde HW, Harada S, Agarwal DP (1979) Racial differences in alcohol sensitivity: a new hypothesis. *Hum Genet* 51:331–4
- Goldman D, Enoch M-A (1990) Genetic epidemiology of ethanol metabolic enzymes: a role for selection. *World Rev Nutr Diet* 63:143–160
- Hammer MF (1995) A recent common ancestry for human Y chromosomes. *Nature* 378:376–378
- Harada S, Agarwal DP, Goedde HW, Tagaki S, Ishikawa B (1982) Possible protective role against alcoholism for aldehyde dehydrogenase isozyme deficiency in Japan. *Lancet* II:827
- Harding RM, Fullerton SM, Griffiths RC, Bond J, Cox MJ, Schneider JA, Moulin DS, Clegg JB (1997) Archaic African and Asian lineages in the genetic ancestry of modern humans. *Am J Hum Genet* 60:772–789
- Hästbacka J, Chapelle A de la, Kaitila I, Sistonen P, Weaver A, Lander E (1992) Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. *Nat Genet* 2:204–211
- Hästbacka J, Chapelle A de la, Mahtani MM, Clines G, Reeve-Daly MP, Daly M, Hamilton BA, Kusumi K, Trivedi B, Weaver A, Coloma A, Lovett M, Buckler A, Kaitila I, Lander E (1994) The diastrophic dysplasia gene encodes a novel sulfate transporter: positional cloning by fine-structure linkage disequilibrium mapping. *Cell* 78:1073–1087
- Hawley ME, Kidd KK (1995) HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. *J Hered* 86:409–411
- Hsu LC, Tani K, Fujiyoshi T, Kurachi K, Yoshida A (1985) Cloning of cDNAs for human aldehyde dehydrogenase 1 and 2. *Proc Natl Acad Sci USA* 82:3771–3775
- Hsu LC, Bendel RE, Yoshida A (1988) Genomic structure of the human mitochondrial aldehyde dehydrogenase gene. *Genomics* 2:57–65
- Huber CG, Oefner PJ, Bonn GK (1992) High-performance liquid chromatographic separation of detritylated oligonucleotides on highly cross-linked poly-(styrene-divinylbenzene) particles. *J Chromatogr* 599:113–118
- Hudson RR (1982) Estimating genetic variability with restriction endonucleases. *Genetics* 100:711–719
- Ikuta T, Szeto S, Yoshida A (1986) Three human alcohol dehydrogenase subunits: cDNA structure and molecular and evolutionary divergence. *Proc Natl Acad Sci, USA* 83:634–638
- Jorde LB, Watkins WS, Viskochil D, O'Connell P, Ward K (1993) Linkage disequilibrium in the neurofibromatosis 1 (NF1) region: implications for gene mapping. *Am J Hum Genet* 53: 1038–1050
- Jorde LB, Watkins WS, Carlson M, Groden J, Albertsen H, Thliveris A, Leppert M (1994) Linkage disequilibrium predicts physical distance in the adenomatous polyposis coli region. *Am J Hum Genet* 54:884–898
- Kerem B-S, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, Buchwald M, Tsui L-C (1989) Identification of the cystic fibrosis gene: genetic analysis. *Science* 245:1073–1080
- Kidd JR, Black FL, Weiss KM, Balazs I, Kidd KK (1991) Studies of three Amerindian populations using nuclear DNA polymorphisms. *Hum Biol* 63:775–794
- Kwiatkowski DJ, Henske EP, Weimer K, Ozelius L, Gusella JF, Haines J (1992) Construction of a GT polymorphism map of human 9q. *Genomics* 12:229–240
- Lerman LS, Fischer SG, Hurley I, Silverstein K, Lumelsky N (1984) Sequence-determined DNA separations. *Annu Rev Biophys Bioeng* 13:399–423
- Lewontin RC (1964) The interaction of selection and linkage. I. General considerations: heterotic models. *Genetics* 49:49–67
- Li W-H, Sadler LA (1991) Low nucleotide diversity in man. *Genetics* 129:513–523
- Liu Q, Sommer SS (1994) Parameters affecting the sensitivities of dideoxy fingerprinting and SSCP. *PCR Methods Appl* 4:97–108
- Long JC, Chakravarti A, Boehm CD, Antonarakis S, Kazazian HH (1990) Phylogeny of human β -globin haplotypes and its implications for recent human evolution. *Am J Phys Anth* 81:113–130
- Long JC, Williams RC, Urbanek M (1995) An E-M algorithm and testing strategy for multiple-locus haplotypes. *Am J Hum Genet* 56:799–810
- Neel JV, Gershowitz H, Mohrenweiser HW, Amos B, Kostyu DD, Salzano FM, Mestriner MA, Lawrence D, Simoes AL, Smouse PE, Oliver WJ, Spielman RS, Neel JV Jr (1980) Genetic studies on the Ticuna, an enigmatic tribe of Central Amazonas. *Ann Hum Genet* 44:37–54
- Nei M (1987) *Molecular evolutionary genetics*. Columbia University Press, New York
- Nei M, Tajima F (1981) DNA polymorphism detectable by restriction endonucleases. *Genetics* 97:145–163
- Nickerson DA, Taylor SL, Weiss KM, Clark AG, Hutchinson RG, Stengård J, Salomaa V, Vartiainen E, Boerwinkle E, Sing CF (1998) DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. *Nat Genet* 19:233–40
- Novoradovsky A, Tsai S-J, Goldfarb L, Peterson RJ, Long JC, 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
- Orita M, Suzuki Y, Sekiya T, Hayashi K (1989) Rapid and sensitive detection of point mutations and DNA polymorphisms using the polymerase chain reaction. *Genomics* 5:874–879
- Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SP (1994) Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc Natl Acad Sci USA* 91:5022–5026
- Peterson RJ (1998) Haplotype analysis: applications to *ALDH2* evolution as evidenced by coalescence and linkage disequilibrium analyses. Doctoral dissertation, Pennsylvania State University
- Pluzhnikov A, Donnelly P (1996) Optimal sequencing strategies for surveying molecular genetic diversity. *Genetics* 144:1247–1262
- Purandare SM, Cawthon R, Nelson LM, Sawada S, Watkins WS, Ward K, Jorde LB, Viskochil DH (1996) Genotyping of PCR-based polymorphisms and linkage-disequilibrium analysis at the *NFI* locus. *Am J Hum Genet* 59:159–166
- Raghunathan L, Hsu LC, Klisak I, Sparkes RS, Yoshida A, Mohandas T (1988) Regional localization of the human genes for aldehyde dehydrogenase-1 and aldehyde dehydrogenase-2. *Genomics* 2:267–269
- Ravnik-Glavac M, Glavic D, Dean M (1994) Sensitivity of single-strand conformation polymorphism and heteroduplex method for mutation detection in the cystic fibrosis gene. *Hum Mol Genet* 3:801–807
- Ruano G, Kidd KK, Stephens JC (1990) Haplotype of multiple polymorphisms resolved by enzymatic amplification of single DNA molecules. *Proc Natl Acad Sci USA* 87:6296–6300

- Sheffield VC, Beck JS, Kwitek AE, Sandstrom DW, Stone EM (1993) The sensitivity of single-strand conformation polymorphism analysis for the detection of single base substitutions. *Genomics* 16:325–332
- Taveré S, Balding DJ, Griffiths RC, Donnelly P (1997) Inferring coalescence times from DNA sequence data. *Genetics* 145:505–518
- Teng Y-S (1981) Human liver aldehyde dehydrogenase in Chinese and Asiatic Indians: gene deletion and its possible implications in alcohol metabolism. *Biochem Genet* 19:107–114
- Thompson EA, Deeb S, Walker D, Motulsky AG (1988) The detection of linkage disequilibrium between closely linked markers: RFLPs at the AI-CIII apolipoprotein genes. *Am J Hum Genet* 42:113–124
- Tishkoff SA, Dietzsch E, Speed W, Pakstis AJ, Kidd JR, Cheung K, Bonnè-Tamir B, Santachiara-Benerecetti AS, Moral P, Krings M, Pääbo S, Watson E, Risch N, Jenkins T, Kidd KK (1996) Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science* 271:1380–1387
- Tishkoff SA, Goldman A, Calafell F, Speed WC, Deinard AS, Bonne-Tamir B, Kidd JR, Pakstis AJ, Jenkins T, Kidd KK (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
- Urbanek M, Goldman D, Long JC (1996) The apportionment of dinucleotide repeat diversity in Native Americans and Europeans: a new approach to measuring gene identity reveals asymmetric patterns of divergence. *Mol Biol Evol* 13:943–953
- Vigilant L, Stoneking M, Harpending H, Hawkes K, Wilson AC (1991) African populations and the evolution of human mitochondrial DNA. *Science* 253:1503–1507
- Wainscoat JS, Hill AVS, Boyce AL, Flint J, Hernandez M, Thein SL, Old JM, Lynch JR, Falusi AG, Weatherall DJ, Clegg JB (1986) Evolutionary relationships of human populations from an analysis of nuclear DNA polymorphisms. *Nature* 319:491–493
- Watkins WS, Zenger R, O'Brien E, Nyman D, Eriksson AW, Renlund M, Jorde LB (1994) Linkage disequilibrium patterns vary with chromosomal location: a case study from the von Willebrand factor region. *Am J Hum Genet* 55:348–355
- Weir BS (1996) *Genetic data analysis II: methods for discrete population genetic data*. Sinauer, Sunderland, Mass.
- Wolff PH (1972) Ethnic differences in alcohol sensitivity. *Science* 175:449–50
- Yoshida A, Huang I-Y, Ikawa M (1984) Molecular abnormality of an inactive aldehyde dehydrogenase variant commonly found in Orientals. *Proc Natl Acad Sci USA* 81:258–261