

Evidence for Recent Positive Selection at the Human *AIM1* Locus in a European Population

Mikiko Soejima,* Hidenori Tachida,† Takafumi Ishida,‡ Akinori Sano,* and Yoshiro Koda*

*Department of Forensic Medicine and Human Genetics, Kurume University School of Medicine, Kurume, Japan; †Department of Biology, Faculty of Sciences, Kyushu University, Fukuoka, Japan; and ‡Department of Biological Sciences, Graduate School of Science, University of Tokyo, Tokyo, Japan

Two missense polymorphisms (E272K and L374F) of the *AIM1* locus, encoding a melanocyte differentiation antigen, were shown to have a clear association with human ethnicities. These two nonpathogenic single nucleotide polymorphisms (SNPs) may be associated with human pigmentation variation. In this study, we investigated sequence variation in the coding region and exon-flanking sequence and found low genetic variation only in subjects of European descent. All four statistical tests applied to the 7.55-kb region surrounding the L374F polymorphism detected statistically significant deviations from selective neutrality in Europeans. In addition, haplotype analysis revealed that one haplotype carrying 374F was overrepresented in this population, and the low rate of variation, with some features of selective sweeps, was shown to be statistically significant. These results suggest that positive selection recently has been acting or has acted on at least this region of the melanogenic gene and that an advantageous haplotype spread rapidly in Europe.

Introduction

Human (antigen in melanoma-1 (AIM-1) was first identified as a melanoma antigen (Harada et al. 2001). The gene is located on chromosome 5p and consists of seven exons spanning a region of approximately 40 kb. The encoded protein is a 530-amino acid polypeptide and is predicted to span the lipid bilayer 12 times (Newton et al. 2001). This gene is expressed in human melanoma and shows homology to plant sucrose-proton symporters. Its homologue in medaka fish is responsible for the pigment phenotype of *b* mutants with different pigmentation phenotypes and encodes a transporter that mediates melanin synthesis (Fukamachi, Shimada, and Shima 2001). This medaka protein consists of 576 amino acids and is 55% identical to human AIM-1. The phenotypes of spliced or nonsense mutants exhibit fewer melanized melanophores than do the missense mutants. Newton et al. (2001) identified the human *AIM1* locus, which they designated *MATP* (membrane-associated transporter protein), and a mouse gene located on chromosome 15 in a region containing the pigmentation locus, *underwhite* (*uw*). These proteins share 82% sequence identity. A series of mouse mutants have been described for this locus, and all were characterized by various degrees of pigment reduction in the eyes and fur. Recently, a study of mouse primary melanocytes carrying the *underwhite* (*uw*) mutation suggested that *AIM1* is important for proper tyrosine processing and intracellular trafficking (Costin et al. 2003).

Albinism represents a group of genetically heterogeneous hereditary abnormalities of melanin pigment synthesis that result in a deficiency or complete absence of melanin in affected individuals. Four different types of oculocutaneous albinism (OCA) have been reported to date (OCA1, OCA2, OCA3, and OCA4). A homozygous G to A transition of the exon 2 splice-acceptor site of the *AIM1* gene was identified in a Turkish patient with OCA as the fourth pathological OCA gene (Newton et al. 2001). In addition to this pathogenic mutation, Newton

et al. identified two polymorphisms in this gene, L374F and T329T, in diverse populations from North America, Asia, Europe, and Africa. Reports about mutations in *AIM1* of OCA4 patients in various populations have been accumulating (Inagaki et al. 2004; Rundshagen et al. 2004; Suzuki et al. 2005). Recently, we investigated sequence variation of the coding region of *AIM1* in randomly selected individuals of European descent in South Africa, Ghanaians, Japanese, and New Guinea islanders (Nakayama et al. 2002) and identified two nonsynonymous polymorphisms. They are G814A and G1122C, resulting in E272K and L374F, respectively. The 272K allele was found in Japanese and New Guinea islander populations, and the 374F allele was found only in the South African Europeans with an allele frequency of 0.89. In addition to our results, Yuasa et al. (2004) showed that the 374F allele was present at a frequency as high as 0.96 in a German population, while it was completely absent in a Japanese population. These results suggest that the 374F allele is specific to Europeans. More recently, Graf, Hodgson, and van Daal (2005) reported that these two nonpathogenic nonsynonymous single nucleotide polymorphisms (SNPs) in the *AIM1* gene were associated with normal human pigmentation variation, i.e., 374L is significantly associated with dark hair, skin, and eye color in Europeans. These results raise the possibility of a directional selection favoring the 374F allele in some environmental conditions, such as areas of low sunlight. To determine whether a selective force is acting on this locus, we analyzed the sequence variation in the coding region, exon-flanking region, and the 7.55-kb region surrounding the L374F polymorphism of the *AIM1* gene.

Materials and Methods

DNA Samples

This study protocol was approved by the ethical committee of Kurume University School of Medicine. Randomly selected individuals from populations sampled were 80 Chinese from Guangzhou (south China), 121 Ghanaians from Accra, 102 Xhosa (Africans) and 101 European-Africans from Cape Town, and 54 Sinhalese and 58 Tamils from Sri Lanka. In addition, genomic DNA of one chimpanzee and one orangutan were also sampled.

Key words: *AIM1*, directional selection, pigmentation.

E-mail: ykoda@med.kurume-u.ac.jp.

Mol. Biol. Evol. 23(1):179–188. 2006

doi:10.1093/molbev/msj018

Advance Access publication September 14, 2005

Direct Sequencing Analysis

The primers used and the temperature profile for amplification of the coding region, exon-flanking sequence, and flanking sequence of L374F polymorphism of *AIM1* are available at <http://www.med.kurume-u.ac.jp/med/foren/AIM1supp1.htm>. All amplifications were performed in 20 μ l of 1 \times buffer containing 1 U of Takara *Ex Taq* polymerase (Takara, Tokyo, Japan). Resultant polymerase chain reaction (PCR) products were purified using Microcon PCR (Millipore, Bedford, Mass.), and the DNA sequence was determined as described previously (Koda et al. 2004). The PCR primers and several sequence-specific internal primers (not shown) were used for DNA sequencing.

Denaturing High Performance Liquid Chromatography Analysis

To identify the L374F polymorphism in various populations, denaturing high performance liquid chromatography (DHPLC) analysis was performed. Amplification was done in 25 μ l of a mixture of 1 \times buffer and 0.75 U of Opti-mase polymerase (Transgenomic, Courtaboeuf, France). The primers used, temperature profile for PCR, and analysis condition are available at <http://www.med.kurume-u.ac.jp/med/foren/AIM1supp2.htm>. Resultant PCR products were denatured for 5 min at 95°C, then gradually reannealed over 40 min by decreasing the sample temperature from 95°C to 25°C, and 5 μ l of the annealed sample was injected into a WAVE-MD system (Transgenomic). Two microliters of the samples that had an unexpected peak profile was treated with 1 U of Exonuclease I (New England Biolabs, Beverly, Mass.) and 1 U of shrimp alkaline phosphatase (Roche Diagnostics, Mannheim, Germany) for 1 h at 37°C and then heat denatured for 15 min at 72°C.

PCR–Restriction Fragment Length Polymorphism Analysis to Identify E272K Polymorphism

The region containing E272K was amplified in 20 μ l of 1 \times buffer containing 1 U of Takara *Ex Taq* polymerase. The temperature profile was denaturation at 96°C for 3 min, followed by 35 cycles of denaturation at 96°C for 15 s, annealing at 55°C for 30 s, and extension at 72°C for 30 s. The resultant PCR products were digested with *TaqI* (Toyobo, Osaka, Japan), which digested the restriction site in alleles containing 272K, and then were electrophoresed.

Data Analysis

To measure the diversity within a population, the number of segregating sites S for sample size n [$\theta_w = S / \left(\sum_{i=1}^{n-1} 1/i \right)$] and the average pairwise sequence difference (k) were calculated. Under the equilibrium neutral model, θ_w and k provide alternative estimators of $\theta = 4N_e\mu$, where N_e is the effective population size and μ is the mutation rate per sequence per generation (Watterson 1975; Nei and Tajima 1981). From the two estimates of θ , the test statistic D proposed by Tajima (1989) was computed. In addition, test statistics D and F proposed by Fu and Li (1993) were computed with the number of

mutations in external branches η_e , using a chimpanzee sequence as the outgroup. In these tests, departures from neutrality are detected as differences between estimates of θ derived from the number of mutations in external branches of the genealogy and from θ_w (giving the index D) or k (giving the index F). Fay and Wu's (2000) H statistic, which is based on the differences between two estimators of θ : π (or k) and θ_H , an estimator based on the frequency of the derived variants, was computed. For Fu and Li's D and F tests and Fay and Wu's H test, a 1,785-bp segment in intron 4, which is thought to have been inserted into the human sequence after divergence from the human-chimpanzee common ancestor, was excluded. Nucleotide diversity (π) was estimated by dividing k by the number of sites. We used the McDonald-Kreitman test (1991) to determine if the ratio of intra- and inter-specific variation is homogenous between synonymous and nonsynonymous substitutions. All analyses were carried out using DnaSP ver. 4.10 (<http://www.ub.es/dnasp/>) (J. Rozas and R. Rozas 1999).

Haplotype frequencies were inferred using PHASE (version 2.02) (Stephens, Smith, and Donnelly 2001; Stephens and Donnelly 2003). To simplify inference, we ignored less frequent polymorphisms, those less than 5% in the total population samples. The Network 4.1 package (<http://www.fluxus-technology.com/sharenet.htm>) was used to construct the median-joining network using the method of Bandelt, Forster, and Rohl (1999), which reflects the mutational relationships among the inferred haplotypes and the evolutionary history of genetic changes at the *AIM1* locus with the common chimpanzee as an outgroup species. The insertion/deletion polymorphisms were not included in any of the data analyses.

The haplotype test of Hudson et al. (1994) was used on the Europeans and Chinese. Data sets expected under the neutrality assumption and conditioned on the observed number of segregating sites were generated by a coalescent simulation program, ms (Hudson 2002). We assumed a model with no recombination.

To detect a selective sweep, we used the composite-likelihood ratio (CLR) test proposed by Kim and Stephan (2002) and also used a modified method of it to allow incomplete sweeps like those obtained by Meiklejohn et al. (2004). In this test, the different hypotheses are tested using the likelihood ratios $LR_1 = \log(L1/L0)$ and $LR_2 = \log(L2/L1)$ of the maximum composite likelihoods under the neutral model ($L0$), the complete sweep model ($L1$), and incomplete sweep model ($L2$). In addition, the CLR test that does not distinguish ancestral/derived alleles (Kim and Stephan 2002) and a goodness-of-fit (GOF) test, which aims to distinguish between selective sweep and demography (Jensen et al. 2005), were also performed. All these CLR and GOF test statistics and the estimates of the strength of selection $\alpha = 2N_e s$, where N_e is the effective population size and s is the selection coefficient, were calculated using programs kindly provided by Y. Kim. The time since the fixation and the strength of selection at the *AIM1* locus in a European population were estimated by using the maximum-likelihood method proposed by Meiklejohn et al. (2004).

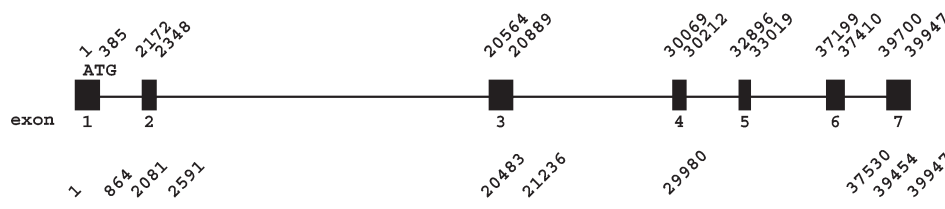


FIG. 1.—Gene structure of *AIM1*. Shaded boxes represent the seven exons. Nucleotide positions start with the adenine of the first ATG. The analyzed regions are underlined.

Results

Sequence Variation in Coding Region, Intronic Region, 3' Untranslated Region, and 7,551-bp Region

All seven exons, the exon-flanking sequences, and the continuous 7,551 bp surrounding the L374F polymorphism of the *AIM1* locus (fig. 1) were sequenced from 17 randomly selected Europeans in South Africa and 10 each of Chinese, Sinhalese in Sri Lanka, and Ghanaians and Xhosans in South Africa. In total, we identified five SNPs in a 1,593-bp coding region, two of them synonymous and three nonsynonymous, and 55 polymorphisms in an 8,558-bp intron sequence, seven indel polymorphisms, and 48 SNPs (table 1).

One chimpanzee sequence and one orangutan partial sequence were also obtained to infer the ancestral and derived states of each human polymorphism and haplotype. Because comparison of the human sequence with the chimpanzee and the orangutan sequences revealed that 1,785 bp in intron 4 (from 30,437 to 32,222 bp) had been inserted into the human sequence after divergence from the human-chimpanzee common ancestor, we could not estimate the ancestral states of 13 intronic SNPs. There were five synonymous and six nonsynonymous fixed differences in the coding region. We applied the McDonald-Kreitman test to the *AIM1* coding sequence, but there was no significant excess of nonsynonymous substitutions between human and chimpanzee sequences as compared to nonsynonymous polymorphisms in humans ($P = 1$). In the intronic sequence, 79 fixed differences were encountered: they were 69 dinucleotide and 10 indel differences. The chimpanzee sequence was monomorphic at all the positions that were polymorphic in humans. Human-chimpanzee divergence was 0.95% (80 sites), that is, 0.69% for the coding and 1.0% for the noncoding regions.

Nucleotide diversities were calculated for the total sequenced region, nonsynonymous sites, noncoding region + synonymous sites, and the chosen 7,551-bp region encompassing a part of intron 3, exon 4, intron 4, exon 5, intron 5, exon 6, and a part of intron 6 (table 2) because this partition could reveal a contrasting pattern of population-specific variation. In addition, summary statistics for the 7,551-bp region are also presented. The π values for each population except Europeans and for the pooled sample at the total sequenced region, noncoding region + synonymous sites, and 7,551-bp region were similar to the average π (0.0011) for a fourfold degenerate site in humans (Li and Sadler 1991), but the π values for Europeans were quite low (table 2). Because of low-frequency SNPs and high frequency-derived SNPs (five sites), all four statistical tests in

the 7,551-bp region detected statistically significant deviations from selective neutrality for Europeans (table 2).

As reported previously (Nakayama et al. 2002; Yuasa et al. 2004; Graf, Hodgson, and van Daal 2005), two non-synonymous polymorphisms (E272K and L374F) show a population-specific pattern. However, because our sample size for the sequence analysis was small, we then checked the frequencies of E272K and L374F polymorphisms in a larger sample of the same populations (101 Europeans, 80 Chinese, 54 Sinhalese, 121 Ghanaians, 102 Xhosans) and 58 Tamils in Sri Lanka by PCR–restriction fragment length polymorphism and DHPLC. These sites show distinctive frequency distributions among different population groups; that is, 272K is the common allele of Asian populations, such as Chinese (43.4%), Sinhalese (20.4%), and Tamils (12.1%), but is rare in Europeans (2.5%), Xhosans (3.4%), and Ghanaians (4.1%). The 374F is exclusively found in Europeans (91.6%) but not in the other five populations (0%–1.9%). The fact that this derived SNP, 374F, is dominant in only a single population also raises the possibility that a kind of population-specific selection has been acting on it or its surrounding region.

Haplotype Variation

To elucidate the history of the generation of variations and recombinations of this locus, we next inferred haplotypes on the basis of genotype data from 23 of 53 SNPs, whose frequencies were more than 5% in the total sample of the five populations by using PHASE software version 2.1.1. As shown in table 3, 31 haplotypes were estimated in the five populations. A total of 7, 6, 3, 16, and 10 haplotypes were observed in the Chinese, Sinhalese, European, Xhosan, and Ghanaian populations, respectively (table 4). Only one haplotype in the Europeans contained 374F (haplotype 12).

To determine the relationships among inferred haplotypes, we constructed a median-joining network by using Network software v.4.1 (fig. 2). It is clear that haplotype 12 is European specific and has no derived haplotypes.

Signature of Directional Selection

Positive selection can play an important role by sweeping out nucleotide variation around selected sites (Maynard Smith and Haigh 1974). The haplotype test can address whether it is likely that a kind of haplotype structure will occur under neutrality. In Europeans, PHASE analysis inferred that 32 of the 34 chromosomes share the same haplotype (haplotype 12) when we considered

Table 1
Polymorphic Sites and Their Frequencies in Chinese, Sinhalese, European, Xhosan, Ghanaian, and Pooled Samples

Location	Site	Polymorphism	Ancestral	Frequencies of Derived or Minor Alleles in					
				CN (<i>n</i> = 20)	SH (<i>n</i> = 20)	EU (<i>n</i> = 34)	XA (<i>n</i> = 20)	GH (<i>n</i> = 20)	POOL (<i>n</i> = 114)
Int 1	508	T/A	T	0	0	0	0	0.05	0.009
Int 1	849	G/A	G	0	0	0	0.05	0	0.009
Int 1	2121*	C/T	C	0.1	0.15	0.03	0	0	0.053
Int 2	2450	C/T	C	0	0.05	0	0	0	0.009
Ex 3	20781	G/T,R (L260F)	G	0	0.05	0	0	0	0.009
Ex 3	20815*	G/A,R (E272K)	G	0.5	0.3	0.03	0	0.05	0.158
Int 3	20948*	TTG/-	TTG	0.8	0.55	0.97	0.75	0.95	0.825
Int 3	21110*	GT/-	—	0.2	0.45	0.03	0.2	0.05	0.167
Ex 4	30167*	G/A, SY (T329T)	G	0.2	0.45	0.03	0.1	0.05	0.149
Int 4	30285	-/TGGGCT	—	0	0	0.03	0	0	0.009
Int 4	30341	G/A	G	0	0	0	0	0.05	0.009
Int 4	30441	A/G	Minor:G	0	0	0	0.2	0	0.035
Int 4	30510	C/T	Minor:T	0	0	0	0.05	0	0.009
Int 4	30541*	G/A	Minor:A	0.15	0.45	0	0.2	0.05	0.149
Int 4	30752	G/A	Minor:A	0	0	0	0.05	0	0.009
Int 4	31150	T/G	Minor:G	0	0	0	0.05	0	0.009
Int 4	31307	G/A	Minor:A	0.05	0	0	0	0	0.009
Int 4	31308*	C/T	Minor:T	0.8	0.7	0.06	0.25	0.2	0.360
Int 4	31421	T/G	Minor:G	0	0	0	0.05	0	0.009
Int 4	31480*	A/G	Minor:G	0	0	0	0.05	0.35	0.070
Int 4	31551*	G/A	Minor:A	0	0	0	0.05	0.35	0.070
Int 4	31611	G/A	Minor:A	0	0	0	0.2	0	0.035
Int 4	31634	G/A	Minor:A	0	0	0	0	0.05	0.009
Int 4	31866*	A/G	Minor:G	0.05	0.3	0.94	0.4	0.15	0.439
Int 4	32289	G/T	G	0	0	0	0.05	0	0.009
Int 4	32300*	C/G	C	0.2	0.3	0.94	0.6	0.8	0.614
Int 4	32325	A/G	A	0	0	0	0.15	0	0.026
Int 4	32408	G/A	G	0	0	0	0	0.05	0.009
Int 4	32572*	G/A	G	0.05	0.2	0.94	0.3	0.15	0.404
Int 4	32586*	T/C	T	0.6	0.25	0.03	0	0.10	0.175
Int 4	32617*	G/T	G	0.2	0.45	0.03	0.2	0.05	0.167
Int 4	32747	G/A	G	0	0	0	0	0.05	0.009
Int 4	32852*	C/A	C	0.6	0.3	0.03	0	0.05	0.175
Ex 5	32985*	G/C, R (L374F)	G	0	0	0.94	0	0	0.281
Int 5	33127*	T/C	T	0.15	0	0	0.15	0.15	0.079
Int 5	33163	T/C	T	0	0	0	0.05	0	0.009
Int 5	33197	A/G	A	0	0	0	0.15	0.05	0.035
Int 5	33339	T/C	T	0	0	0.03	0	0	0.009
Int 5	33433	TCAG/-	TCAG	0	0	0	0.05	0	0.009
Int 5	33562*	C/A	C	0	0.2	0.94	0.1	0	0.333
Int 5	33890	A/G	A	0	0	0.03	0	0	0.009
Int 5	33975*	T/C	T	0.6	0.25	0.03	0	0.1	0.175
Int 5	34227*	T/A	T	0.2	0.05	0	0.25	0.05	0.096
Int 5	34381*	T/G	T	0.6	0.25	0.03	0	0.1	0.175
Int 5	34560	G/T	G	0	0	0	0	0.05	0.009
Int 5	34699*	CAAT/-	CAAT	0	0	0	0	0.35	0.061
Int 5	34780	G/-	G	0	0.15	0	0	0	0.026
Int 5	35263	A/G	A	0.05	0	0	0	0	0.009
Int 5	35268*	A/G	A	0	0	0	0.05	0.35	0.070
Int 5	35680	C/T	C	0	0.1	0	0	0	0.018
Int 5	36002	CA/-	CA	1.0	1.0	1.0	0.9	1.0	0.982
Int 5	36016	T/A	T	0.05	0	0	0	0	0.009
Int 5	36089*	A/G	A	0.05	0.2	0.94	0.2	0.2	0.395
Int 5	36359*	A/G	A	0.6	0.25	0.03	0.05	0.1	0.184
Int 5	36749	G/C	G	0	0	0	0.05	0	0.009
Int 5	36832	A/G	A	0	0	0	0.05	0	0.009
Int 5	37110	G/C	G	0	0	0	0.05	0	0.009
Int 5	37134*	G/T	G	0.6	0.25	0.03	0.1	0.1	0.193
Ex 6	37293	C/G, SY (L417L)	C	0.1	0	0	0	0	0.018
Int 6	39610	A/G	A	0	0	0	0.05	0	0.009

NOTE.—Nucleotide positions start with the adenine of the first ATG. Asterisks (*) mark the position of polymorphisms whose frequency is >0.05 in the pooled sample. Int, intron; Ex, exon; CN, Chinese; SH, Sinhalese; EU, Europeans; XA, Xhosans; GH, Ghanaians; POOL, pooled (combined) sample.

polymorphic sites whose frequencies were more than 5% in the total sample of the five populations, whereas at least 28 chromosomes share the same haplotype carrying 374F when we considered all polymorphic sites; of the 17 indi-

viduals, 13 and 2 were homozygous and heterozygous for only one site, respectively. The estimated *P* value from 10,000 random coalescent simulations of observing 28 or more of the same haplotype is 0.0009. Coalescent

Table 2
Nucleotide Variation and Population Genetic Statistics at *AIMI* in Various Populations

Population	Chinese	Sinhalese	Europeans	Xhosans	Ghanaians	Pooled
<i>n</i>	20	20	34	20	20	114
Total sequenced region (10,174 bp)						
<i>S</i>	22	22	19	31	27	53
$\theta_w/\text{sites} (\times 10^{-4})$	6.1	6.1	4.6	8.6	7.5	9.8
$\pi (\times 10^{-4})$	6.6	7.5	1.4	6.4	5.4	7.5
Nonsynonymous sites (1,195 bp)						
<i>S</i>	1	2	2	0	1	3
$\theta_w/\text{sites} (\times 10^{-4})$	2.4	4.8	4.1	0	2.4	5.9
$\pi (\times 10^{-4})$	4.4	4.5	1.5	0	0.8	6.1
Noncoding region + synonymous sites (8,979 bp)						
<i>S</i>	21	20	17	31	26	50
$\theta_w/\text{sites} (\times 10^{-4})$	6.6	6.3	4.6	9.7	8.2	10.5
$\pi (\times 10^{-4})$	6.9	7.9	1.4	7.1	6.0	7.7
L374F-flanking region (7,551 bp)						
<i>S</i>	20	17	17	29	25	46
$\theta_w/\text{sites} (\times 10^{-4})$	7.5	6.3	5.5	10.8	9.3	11.5
$\pi (\times 10^{-4})$	8.0	8.7	1.8	8.7	7.0	9.6
η_s	6	1	10	14	11	19
Tajima's <i>D</i>	0.27279 (NS)	1.41565 (NS)	-2.22568 ($P < 0.01$)	-0.75320 (NS)	-0.95938 (NS)	-0.52030 (NS)
Fu and Li's <i>F</i>	0.38992 (NS)	1.47184 (NS)	-3.22069 ($P < 0.02$)	-0.83432 (NS)	-1.49835 (NS)	-1.81585 (NS)
Fu and Li's <i>D</i>	0.19466 (NS)	1.24705 (NS)	-2.89964 ($P < 0.02$)	-0.67014 (NS)	-1.21422 (NS)	-2.31393 (NS)
Fay and Wu's <i>H</i>	0.26316 (NS)	3.03158 ($P < 0.05$)	-7.98574 ($P < 0.025$)	2.60000 (NS)	1.60000 (NS)	3.16659 (NS)

NOTE.—*S*, number of segregating sites (excluding insertion/deletion); *n*, number of chromosomes; $\theta_w/\text{sites} = S / (\sum_{i=1}^{n-1} 1/i)$ /sites; NS, not significant.

simulations were conditioned on the number of segregating sites observed. Also, the three haplotypes that bear the K allele at the E272K polymorphic site (i.e., haplotypes 24, 25, and 26, but particularly 24) are overrepresented in the samples, especially in Chinese. This raises the possibility that the 272K allele also might be under selection in some populations. We used the haplotype test to address this possibility as well. There are 10 with the same haplotype in 20 chromosomes in Chinese, and the test did not yield a significant result.

The CLR test (Kim and Stephan 2002; Meiklejohn et al. 2004) was also conducted to detect positive selection in this region by examining a local reduction of variation and a skew of the frequency spectrum caused by a hitchhiking event. For both simulation and composite-likelihood analyses, the sequence was divided into noncoding and coding regions, and the per-nucleotide mutation rate for each region was given as θ , where θ is Watterson's estimator of the population mutation parameter calculated from the data (Watterson 1975). The scaled per-nucleotide recombination parameter ($R_n = 4N_e r$) was estimated to be 4×10^{-4} by the assumption of effective population size (N_e) as 10,000 (Takahata 1993) and recombination rates (r) as 1×10^{-8} that was calculated to be the average for humans (Kong et al. 2002). Because we could not determine the ancestral status of the 1,785-bp region, we tentatively treated high-frequency alleles in the 1,785-bp region as ancestral alleles (0) and low-frequency alleles as derived alleles (1). Then, the CLRs were obtained for both situations, one for distinguishing ancestral/derived alleles (LR_{option1}) and the other for not distinguishing them (LR_{option2}). We compared the LR₁ values obtained by polymorphism data of Europeans with those obtained by 1,000 simulated data sets for a neutral model for Europeans. As shown in table 5, the *AIMI* data sets in Europeans yielded significantly large ratios (LRs) regardless of the states of

alleles, and the strength of selection ($\alpha = 2N_e s$) is 220–240. However, this test is not robust enough to reveal an undetected population structure or a recent bottleneck.

To distinguish between selective sweep and demography, we also used the GOF test (Jensen et al. 2005). We conducted 1,000 simulations to obtain data sets for the complete sweep model for Europeans. We compared GOF statistics obtained by polymorphism data with those obtained by simulation data for the complete sweep model in Europeans. The GOF values with both option1 and option2 in Europeans were not significant. These results suggested that the original rejection of neutrality in Europeans by the CLR test is more likely to be due to a selective sweep than demography alone.

In addition to the average recombination rate in the human genome, we also perform CRL and GOF tests under lower ($R_n = 4 \times 10^{-5}$ and 4×10^{-6}) recombination rates. Although the strength of selection decreased to less than 1/10 of the original estimate, we cannot still reject complete sweep model by these tests.

To determine whether the sweep is complete or not, we then compared LR₂ values obtained from polymorphism data in Europeans with those obtained by simulation data under the complete sweep model (Meiklejohn et al. 2004). The LR₂ values from polymorphism data in Europeans were not significant. Based on this, we cannot reject a complete sweep model in Europeans. In addition, the selection coefficient s was estimated to be about 0.01 when N_e is assumed to be 10,000 (Takahata 1993) because $\alpha = 2N_e s$ was estimated to be 200 by the CLR test (table 5).

The age of a recently derived haplotype can be estimated on the basis of the number of new mutations that have occurred among the alleles since they last shared a common ancestor. Thus, the age of alleles containing 374F (haplotype 12) of the *AIMI* gene was estimated by following the method of Meiklejohn et al. (2004).

Table 3
Estimated Haplotypes of *AIM1*

Haplotype	2121	‡20815	30167	30541	31308	31480	31551	31866	32300	32572	32586	32617	32852	‡32985	33127	33562	33975	34227	34381	35268	36089	36359	37134
CHIMP	C	G	G	-	-	-	-	-	C	G	T	G	C	G	T	C	T	T	T	A	A	A	G
1	*	*	*	G	T	A	G	A	*	*	C	*	A	*	*	*	C	*	G	*	*	G	T
2	*	*	*	G	T	G	A	A	G	*	*	*	*	*	*	*	*	*	*	G	*	*	*
3	*	*	*	G	C	A	G	A	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
4	*	*	*	G	C	A	G	A	G	*	*	*	A	*	C	*	*	*	*	*	*	*	*
5	*	*	*	G	C	A	G	A	G	*	*	*	*	*	*	*	*	*	*	*	*	*	*
6	*	*	*	G	C	A	G	A	G	*	*	*	*	*	*	*	*	A	*	*	*	*	*
7	*	*	*	G	C	A	G	A	G	*	*	*	*	*	C	*	*	*	*	*	*	*	*
8	*	*	*	G	C	A	G	G	G	A	*	*	*	*	*	*	*	*	*	*	*	*	*
9	*	*	*	G	C	A	G	G	G	A	*	*	*	*	*	*	*	*	*	*	G	*	*
10	*	*	*	G	C	A	G	G	G	A	*	*	*	*	*	*	*	A	*	*	*	*	*
11	*	*	*	G	C	A	G	G	G	A	*	*	*	*	*	A	*	*	*	*	G	*	*
12	*	*	*	G	C	A	G	A	G	A	*	*	*	C	*	*	*	*	*	*	G	*	*
13	*	*	*	G	C	G	A	A	G	*	*	*	A	*	*	*	*	*	*	G	*	*	*
14	*	*	*	G	C	G	A	A	G	*	*	*	*	*	*	*	*	*	*	G	*	*	*
15	*	*	*	A	T	A	G	G	*	*	*	T	*	*	*	*	*	A	*	*	*	*	*
16	*	*	*	G	C	A	G	A	*	*	*	*	*	*	*	*	*	*	*	*	*	*	T
17	*	*	*	A	T	A	G	A	*	*	*	T	*	*	*	*	*	*	*	*	*	*	*
18	*	*	*	A	T	A	G	A	*	*	*	T	*	*	*	*	*	A	*	*	*	*	*
19	*	*	A	G	T	A	G	A	*	*	*	*	*	*	*	*	*	*	*	*	G	*	*
20	*	*	A	G	T	A	G	A	*	*	*	*	*	*	*	*	*	*	*	*	*	G	T
21	*	*	A	G	C	A	G	G	G	A	*	*	*	*	*	A	*	*	*	*	G	*	*
22	*	*	A	A	T	A	G	A	*	*	*	T	*	*	*	*	*	*	*	*	*	*	*
23	*	*	A	A	T	A	G	A	*	*	*	T	*	*	*	*	*	A	*	*	*	*	*
24	*	A	*	G	T	A	G	A	*	*	C	*	A	*	*	*	C	*	G	*	*	G	T
25	*	A	*	G	T	A	G	A	*	*	C	*	*	*	*	*	C	*	G	*	*	G	T
26	*	A	*	G	C	A	G	G	G	A	*	*	A	*	*	A	*	*	*	*	G	*	*
27	T	*	*	G	C	A	G	A	G	*	*	*	*	*	C	*	*	*	*	*	*	*	*
28	T	*	A	G	T	A	G	A	*	*	*	T	*	*	*	*	*	A	*	*	*	*	*
29	T	*	A	A	T	A	G	A	*	*	*	T	*	*	*	*	*	*	*	*	*	*	*
30	T	*	A	A	T	A	G	G	*	*	*	T	*	*	*	*	*	*	*	*	*	*	*
31	T	*	A	A	T	A	G	A	*	*	*	T	*	*	*	*	*	*	*	*	*	*	*

NOTE.—Bases identical to those in the chimpanzee sequence (CHIMP) are marked with a dot. Asterisks (‡) mark the position of E272K (20815) and L374F (32985).

Table 4
Haplotype Frequency of AIM1

Haplotype	Populations					
	CN	SH	EU	XA	GH	POOL
1	2				1	3
2					2	2
3				2		2
4				1		1
5				1	3	4
6				1		1
7	2			2	3	7
8				1		1
9	1			2	3	6
10				1		1
11		3		1		4
12			32			32
13				1		1
14					5	5
15				2		2
16				1		1
17				1		1
18				1	1	2
19					1	1
20				1		1
21				1		1
22		7				7
23	3				1	4
24	10	5	1			16
25					1	1
26		1				1
27	1					1
28		2				2
29			1			1
30	1					1
31		1				1
total	20	20	34	20	20	114

NOTE.—CN, Chinese; SH, Sinhalese; EU, Europeans; XA, Xhosans; GH, Ghanaians; POOL, pooled (combined) sample.

There are 32 sequences containing 374F with two derived mutations (positions 33339 and 33890) in the 5,766-bp region (7,551–1,785 bp); thus, the expected number of mutations in the European sample is $32 t\mu$, where μ is the mutation rate per sequence per year and t is the time since the common ancestor. On the basis of the observed silent site divergence, comparing 57 sites (synonymous and non-coding sites) in *AIM1* between human and chimpanzee, and assuming a divergence time of 5 Myr for these two species, μ is estimated to be 5.7×10^{-6} . The maximum-likelihood estimate of t is 10,965 years. Ninety-five percent confidence intervals for this age can be calculated by finding t_{\max} and t_{\min} such that $P(S \leq 2|t_{\max}) = 0.025$ and $P(S \geq 2|t_{\min}) = 0.025$. This produces the values $t_{\min} = 1,328$ and $t_{\max} = 39,609$ years.

Discussion

Sunshine exerts both beneficial and harmful effects on human health. In the presence of substantial ambient UV radiation, those with pale skin are more at risk of major skin cancers and more likely to suffer from one of a range of porphyrias. However, pale skin is advantageous for vitamin D synthesis at higher latitudes (Rees 2004). Short-wavelength UV (UVB) converts 7-dehydrocholesterol into an essential precursor of cholecalciferol (vitamin D₃);

a deficiency of vitamin D causes rickets, a characteristic pattern of growth abnormalities and bone deformities. Most variance in skin color is between continents rather than within continents, which is compatible with the view that there has been strong and recent selection on skin pigmentation. Pigmentation results from the production and deposition of melanin, which is synthesized from tyrosine as either black/brown eumelanin or yellow/red pheomelanin. Melanin biosynthesis takes place in melanosomes within melanocytes. Melanosomes differ in their shape, size, number, and distribution, depending on the type of melanin they contain, and this determines the pigmentation phenotype of hair and skin. Despite a large number of murine coat-color mutations, only one gene in humans, the *melanocortin 1 receptor (MC1R)*, is known to account for substantial variation in skin and hair color and skin cancer incidence (Rees 2004). Although the *MC1R* gene has been studied extensively and can explain some of the normal pigmentation variation in humans, little is understood about other key pigmentation genes and their effects on normal pigmentation. However, one SNP in the 3' untranslated region of another pigmentation gene, the *agouti signaling protein (ASIP)* gene, is recently reported to be associated with human pigmentation (Kanetsky et al. 2002; Bonilla et al. 2005). In addition, both 272K and 374L of the *AIM1* locus have been reported to have a strong association with dark hair, skin, and eye color in Europeans in Queensland, Australia (Graf, Hodgson, and van Daal 2005). However, the 374L allele is an ancestral allele and common in other populations, including Africans, whereas the 272K allele is derived and rare in Africans. In addition, 374L and 272K are in complete linkage disequilibrium in Europeans ($|D'| = 1$). Thus, it is likely that 374L rather than the 272K allele plays an important role in dark hair, skin, and eye pigmentation in Europeans. In other words, individuals having the *AIM1* allele 374F might exhibit pale hair, skin, and eye pigmentation and have advantages for vitamin D synthesis at higher latitudes.

Our present study suggests that the genetic variation is low in the nonsynonymous sites but higher in the silent sites in the African population, while the variation is low in both synonymous and silent sites in Europeans. On the other hand, the diversity is the highest in both nonsynonymous and silent sites in Asians. As in *AIM1*, a low genetic diversity of nonsynonymous sites in *MC1R* was observed in Africans (Rees 2004). These observations suggest that these two genes are probably under strong functional constraints in Africa where any deviation from eumelanin production appears to be evolutionarily deleterious. In contrast to African pattern, European polymorphic pattern of *AIM1* is quite different from that of *MC1R*. We found skewed allele frequency in *AIM1* while diversity is increased without an overrepresented derived allele at *MC1R* (Rana et al. 1999; Harding et al. 2000). Thus, it is likely that there was a recent selective sweep possibly by a mutation (374F) at *AIM1* but no such event occurred at *MC1R* in Europeans.

In this study, all four statistical tests and the haplotype test detected statistically significant deviations from selective neutrality for Europeans because of the skewed allele frequency (table 2). However, most of the tests we applied

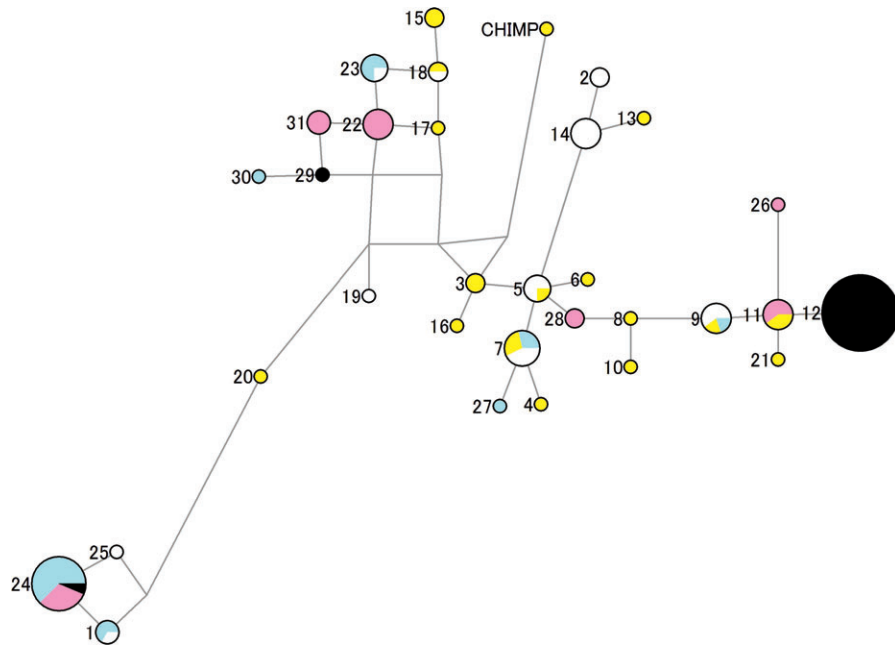


FIG. 2.—Median-joining networks for predicted *AIM1* haplotypes, based on 23 SNPs and the chimpanzee sequence. The haplotype numbers are the same as in tables 3 and 4. The size of each node is proportional to the haplotype frequency in the pooled sample. Each node is colored to indicate the number of times it was found in each population. Blue, pink, black, yellow, and white circles represent Chinese, Sinhalese, Europeans, Xhosans, and Ghanaians, respectively.

are sensitive to demographic history. For example, Tajima's *D* can become strongly negative when the population has gone through a prolonged bottleneck and expanded thereafter (Sano and Tachida 2005). However, our previous studies suggest that such a severe population bottleneck is unlikely to be the cause (Koda et al. 2004; Soejima et al. 2005). In addition, the GOF test proposed by Jensen et al. (2005) suggested that the original rejection of neutrality in Europeans by the CLR test is more likely to be due to a selective sweep than demography alone. To examine the possibility of an incomplete sweep on the putative target, 374F, we performed the CLR test on the European population, but the result of this test cannot reject the complete selective sweep model. This may be because the 374F allele is almost fixed in Europeans. In fact, the CLR values obtained by the polymorphic data in Europeans were quite similar to the average values from 1,000 simulation data of the incomplete sweep model using the assumption that 374F is the

beneficial allele (target site is 32985, beneficial allele frequency is 0.94, and α is 4000). Thus, the polymorphic pattern of *AIM1* in Europeans could be explained by the incomplete sweep acting on 374F, and the strength of selection might be greater than $\alpha = 200$ obtained from the complete sweep assumption (table 5).

To date, several examples of incomplete selective sweeps have been reported. Malaria is one of the selective agents that has influenced genetic diversity in humans. Sabeti et al. (2002) used extended haplotype homozygosity (EHH) analysis to investigate recent positive selection at two genes carrying common variants implicated in resistance to malaria, *G6PD* and the CD40 ligand gene (*TNFSF5*). One haplotype that carries the protection-associated alleles in both genes was common in Africa (18% for *G6PD* and 34% for *TNFSF5*), where malaria is endemic, but was absent outside Africa and had a much higher relative EHH of 413 kb for *G6PD* and 506 kb for *TNFSF5* than other haplotypes of comparable frequency. The estimated ages of these haplotypes were about 2,500 years for *G6PD* and 6,600 years for *TNFSF5* (and none of the test statistics showed significant values for these genes). We can speculate that the origin of the selective force acting on *AIM1* is older than that of malaria and that the selective sweep is more complete at *AIM1* in Europeans than at these malaria-resistance genes in Africans.

Recently, Graf, Hodgson, and van Daal (2005) proposed that the 374F allele results in a reduction of function that alters the intracellular trafficking of melanosomal proteins, creating an environment for decreased melanin production. Such a functional difference allows those in high latitudes to absorb the UVB necessary for the synthesis of vitamin D₃. Conversely, this newly arisen allele does not

Table 5
Results of CLR and GOF Tests Applied to *AIM1* in Europeans

	Option1	Option2
LR ₁ ^a	14.174 (<i>P</i> = 0.006)	8.554 (<i>P</i> = 0.004)
GOF ^b	560.226 (<i>P</i> = 0.311)	2166.311 (<i>P</i> = 0.308)
α (LR ₁)	220.97	242.07
LR ₂ ^b	0.2659 (<i>P</i> = 0.253)	
α (LR ₂)	3956.98	

NOTE.—Option1, distinguishing ancestral/derived allele; Option2, not distinguishing ancestral/derived allele.

^a *P* values are based on 1,000 replicates of simulations under the neutral model.

^b *P* values are based on 1,000 replicates of simulations under the complete sweep model.

confer a benefit in the equatorial areas but rather enhances the harmful effects of UV radiation. Functional analysis to determine the differences in melanogenesis of both alleles is necessary to understand not only the selective force but also the function of this molecule in normal pigmentation.

Acknowledgments

This paper is dedicated to the memory of Osamu Takenaka. We thank Yuseob Kim for kindly giving us the computer program. We thank Hiroshi Kimura (Chiba Institute of Science, Choshi, Japan) for helpful advice and support and also thank Ernette D. du Toit (Department of Immunology, Medical School, Cape Town, South Africa) for providing DNA samples. This work was supported by grants-in-aid for Scientific Research from the Ministry of Education, Science, Culture and Sports of Japan and Mitsui Life Social Welfare Foundation. We thank Katherine Ono for the English editing of this manuscript.

Literature Cited

- Bandelt, H., P. Forster, and A. Rohlf. 1999. Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* **16**:37–48.
- Bonilla, C., L.-A. Boxill, S. A. M. Donald, T. Williams, N. Sylvester, E. J. Parra, S. Dios, H. L. Norton, M. D. Shriver, and R. A. Kittles. 2005. The 8818G allele of the agouti signaling protein *ASIP* gene is ancestral and is associated with darker skin color in African Americans. *Hum. Genet.* **116**:402–406.
- Costin, G.-E., J. C. Valencia, W. D. Vieira, M. L. Lamoreux, and V. J. Hearing. 2003. Tyrosinase processing and intracellular trafficking is disrupted in mouse primary melanocytes carrying the underwhite (uw) mutation. A model for oculocutaneous albinism (OCA) type 4. *J. Cell Sci.* **116**:3203–3212.
- Fay, J. C., and C.-I. Wu. 2000. Hitchhiking under positive Darwinian selection. *Genetics* **155**:1405–1413.
- Fu, Y. X., and W. H. Li. 1993. Statistical tests of neutrality of mutations. *Genetics* **133**:693–709.
- Fukamachi, S., A. Shimada, and A. Shima. 2001. Mutations in the gene encoding B, a novel transporter protein, reduce melanin content in medaka. *Nat. Genet.* **28**:381–385.
- Graf, J., R. Hodgson, and A. van Daal. 2005. Single nucleotide polymorphisms in the MATP gene are associated with normal human pigmentation variation. *Hum. Mutat.* **25**:278–284.
- Harada, M., Y. F. Li, M. El-Gamil, S. A. Rosenberg, and P. F. Robbins. 2001. Use of an in vitro immunoselected tumor line to identify shared melanoma antigens recognized by HLA-A*0201-restricted T cells. *Cancer Res.* **61**:1089–1094.
- Harding, R. M., E. Healy, A. J. Ray et al. (11 co-authors). 2000. Evidence for variable selective pressures at MC1R. *Am. J. Hum. Genet.* **66**:1351–1361.
- Hudson, R. R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**:337–338.
- Hudson, R. R., K. Bailey, D. Skarecky, J. Kwiatowski, and F. J. Ayala. 1994. Evidence for positive selection in the superoxide dismutase (Sod) region of *Drosophila melanogaster*. *Genetics* **136**:1329–1340.
- Inagaki, K., T. Suzuki, H. Shimizu et al. (14 co-authors). 2004. Oculocutaneous albinism type 4 is one of the most common types of albinism in Japan. *Am. J. Hum. Genet.* **74**:466–471.
- Jensen, J. D., Y. Kim, V. Bauer DuMont, C. F. Aquadro, and C. D. Bustamante. 2005. Distinguishing between selective sweeps and demography using DNA polymorphism data. *Genetics* **170**:1401–1410.
- Kanetsky, P. A., J. Swoyer, S. Panossian, R. Holmes, D. Guerry, and T. R. Rebbeck. 2002. A polymorphism in the agouti signaling protein gene is associated with human pigmentation. *Am. J. Hum. Genet.* **70**:770–775.
- Kim, Y., and W. Stephan. 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* **160**:765–777.
- Koda, Y., H. Tachida, M. Soejima, O. Takenaka, and H. Kimura. 2004. Population differences in DNA sequence variation and linkage disequilibrium at the *PON1* gene. *Ann. Hum. Genet.* **68**:110–119.
- Kong, A., D. F. Gudbjartsson, S. Jesus et al. (16 co-authors). 2002. A high-resolution recombination map of the human genome. *Nat. Genet.* **31**:241–247.
- Li, W. H., and L. A. Sadler. 1991. Low nucleotide diversity in man. *Genetics* **129**:513–523.
- Maynard Smith, J., and J. Haigh. 1974. The hitch-hiking effect of a favourable gene. *Genet. Res.* **23**:23–35.
- McDonald, J. H., and M. Kreitman. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351**:652–654.
- Meiklejohn, C. D., Y. Kim, D. L. Hartl, and J. Parsch. 2004. Identification of a locus under complex positive selection in *Drosophila simulans* by haplotype mapping and composite-likelihood estimation. *Genetics* **168**:265–279.
- Nakayama, K., S. Fukamachi, H. Kimura, Y. Koda, A. Soemantri, and T. Ishida. 2002. Distinctive distribution of AIM1 polymorphism among major human populations with different skin color. *J. Hum. Genet.* **47**:92–94.
- Nei, M., and F. Tajima. 1981. DNA polymorphism detectable by restriction endonucleases. *Genetics* **97**:145–163.
- Newton, J. M., O. Cohen-Barak, N. Hagiwara, J. M. Gardner, M. T. Davisson, R. A. King, and M. H. Brilliant. 2001. Mutations in the human orthologue of the mouse underwhite gene (uw) underlie a new form of oculocutaneous albinism, OCA4. *Am. J. Hum. Genet.* **69**:981–988.
- Rana, B. K., D. Hewett-Emmett, L. Jin et al. (12 co-authors). 1999. High polymorphism at the human melanocortin 1 receptor locus. *Genetics* **151**:1547–1557.
- Rees, J. L. 2004. The genetics of sun sensitivity in humans. *Am. J. Hum. Genet.* **75**:739–751.
- Rozas, J., and R. Rozas. 1999. DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics* **15**:174–175.
- Rundshagen, U., C. Zuhlke, S. Opitz, E. Schwinger, and B. Kasmann-Kellner. 2004. Mutations in the MATP gene in five German patients affected by oculocutaneous albinism type 4. *Hum. Mutat.* **23**:106–110.
- Sabeti, P. C., D. E. Reich, J. M. Higgins et al. (17 co-authors). 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**:832–837.
- Sano, A., and H. Tachida. 2005. Gene genealogy and properties of test statistics of neutrality under population growth. *Genetics* **169**:1687–1697.
- Soejima, M., H. Tachida, M. Tsuneoka, O. Takenaka, H. Kimura, and Y. Koda. 2005. Nucleotide sequence analyses of human complement 6 (C6) gene suggest balancing selection. *Ann. Hum. Genet.* **69**:239–252.
- Stephens, M., and P. Donnelly. 2003. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet.* **73**:1162–1169.
- Stephens, M., N. J. Smith, and P. Donnelly. 2001. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* **68**:978–989.
- Suzuki, T., K. Inagaki, K. Fukui, A. Obana, S. T. Lee, and Y. Tomita. 2005. A Korean case of oculocutaneous albinism

- type IV caused by a D157N mutation in the MATP gene. *Br. J. Dermatol.* **152**:174–175.
- Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**:585–595.
- Takahata, N. 1993. Allelic genealogy and human evolution. *Mol. Biol. Evol.* **10**:2–22.
- Watterson, G. A. 1975. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**:256–276.
- Yuasa, I., K. Umetsu, G. Watanabe, H. Nakamura, M. Endoh, and Y. Irizawa. 2004. MATP polymorphisms in Germans and Japanese: the L374F mutation as a population marker for Caucasoids. *Int. J. Leg. Med.* **118**:364–366.

Jianzhi Zhang, Associate Editor

Accepted September 8, 2005