The Impact of Natural Selection on an *ABCC11* SNP Determining Earwax Type

Jun Ohashi*,[1] Izumi Naka,[1] and Naoyuki Tsuchiya[1]

[1]Doctoral Program in Life System Medical Sciences, Graduate School of Comprehensive Human Sciences, University of Tsukuba, Tsukuba, Ibaraki, Japan

***Corresponding author:**E-mail: juno-tky@umin.ac.jp.

Associate editor: Anne Stone

Abstract

A nonsynonymous single nucleotide polymorphism (SNP), rs17822931-G/A (538G>A; Gly180Arg), in the *ABCC11* gene determines human earwax type (i.e., wet or dry) and is one of most differentiated nonsynonymous SNPs between East Asian and African populations. A recent genome-wide scan for positive selection revealed that a genomic region spanning *ABCC11*, *LONP2*, and *SIAH1* genes has been subjected to a selective sweep in East Asians. Considering the potential functional significance as well as the population differentiation of SNPs located in that region, rs17822931 is the most plausible candidate polymorphism to have undergone geographically restricted positive selection. In this study, we estimated the selection intensity or selection coefficient of rs17822931-A in East Asians by analyzing two microsatellite loci flanking rs17822931 in the African (HapMap-YRI) and East Asian (HapMap-JPT and HapMap-CHB) populations. Assuming a recessive selection model, a coalescent-based simulation approach suggested that the selection coefficient of rs17822931-A had been approximately 0.01 in the East Asian population, and a simulation experiment using a pseudo-sampling variable revealed that the mutation of rs17822931-A occurred 2006 generations (95% credible interval, 1,023–3,901 generations) ago. In addition, we show that absolute latitude is significantly associated with the allele frequency of rs17822931-A in Asian, Native American, and European populations, implying that the selective advantage of rs17822931-A is related to an adaptation to a cold climate. Our results provide a striking example of how local adaptation has played a significant role in the diversification of human traits.

Key words: *ABCC11*, adaptive evolution, Asians, HapMap, microsatellite marker, natural selection.

Introduction

The type of human earwax is determined by a nonsynonymous single nucleotide polymorphism (SNP), rs17822931-G/A (538G>A; Gly180Arg), in the *ABCC11* gene (MIM *607040) (Yoshiura et al. 2006). The GG and GA genotypes correspond to the wet type of earwax and the AA genotype to the dry type. The rs17822931-A allele leading to dry earwax in a recessive manner is nearly absent in African populations, whereas it is found in European populations and is very common in East Asian populations (supplementary fig. S1, Supplementary Material online) (Yoshiura et al. 2006).

A recent genome-wide scan for positive selection revealed that a genomic region spanning *ABCC11*, *LONP2*, and *SIAH1* genes has been subjected to a selective sweep in East Asians (EAS) (Kimura et al. 2007). Figure 1 shows the averaged heterozygosity in the genomic region around rs17822931 of the *ABCC11* gene, calculated based on the HapMap data (The International HapMap Consortium 2003, 2005). A remarkable reduction in heterozygosity around rs17822931 is found in EAS (i.e., Japanese in Tokyo, Japan [JPT] + Han Chinese in Beijing, China [CHB]), whereas such a reduction is not observed in Yoruba in Ibadan, Nigeria (YRI) (fig. 1). Furthermore, compared with YRI and CEPH Utah residents with ancestry from northern and western Europe (CEU), linkage disequilibrium (LD) around rs17822931 is more extended in EAS (supplementary fig. S2, Supplementary Material online).

Although the genomic region containing rs17822931 appears to have been subjected to positive selection (Kimura et al. 2007), it is hard to readily conclude that rs17822931 is a direct target of positive selection because rs17822931 is located in a large LD block spanning ∼370 kb (from rs520151 to rs7206867) in EAS (supplementary fig. S2, Supplementary Material online). Of 11 SNPs in this LD block, two SNPs, rs17822931 in *ABCC11* and rs6500380 in *LONP2*, are highly differentiated between EAS and YRI and also between EAS and CEU (supplementary fig. S3, Supplementary Material online). Although it is difficult to statistically examine if natural selection has acted against rs17822931 rather than against rs6500380 owing to the strong LD between them ($r^2 = 0.91$), rs6500380, which is located in intron 12 of *LOMP2*, seems to have less functional significance compared with rs17822931, which leads to the Gly180Arg amino acid change in *ABCC11*. In addition, rs17822931 has been reported to be one of most differentiated nonsynonymous SNPs between East Asian and African populations (The International HapMap Consortium 2005; Wang et al. 2007). Taken together, rs17822931 is the most plausible candidate polymorphism to have undergone geographically restricted positive selection in East Asia.

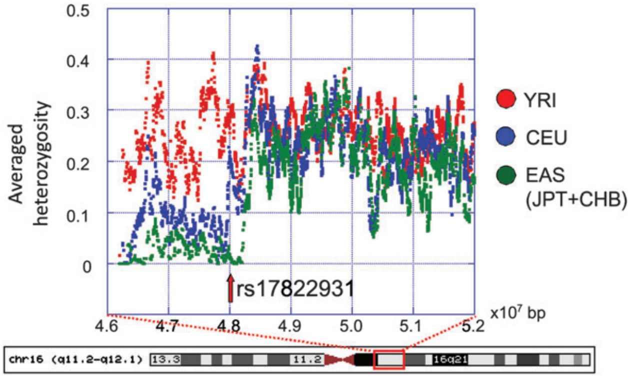

Fig. 1 The averaged heterozygosity around rs17822931 of ABCC11 in HapMap populations. The *x* axis represents the genomic position of the focal SNP on chromosome 16. The *y* axis represents the averaged heterozygosity calculated for SNPs within 25 kb on either side of the focal SNP. The allele frequency data of SNPs on chromosome 16 were obtained from the HapMap database, and only SNPs that were analyzed in three HapMap populations (i.e., YRI, CEU, and EAS [JPT + CHB]) were selected for the calculation. The red arrow indicates the position of rs17822931.

To date, a number of genomic regions and polymorphisms have been reported to be subjected to strong positive selection in human populations (The International HapMap Consortium 2005; Voight et al. 2006; Kimura et al. 2007; Grossman et al. 2010). However, the selection intensities for most of the selected alleles remain unclear. Although molecular variation data are required to investigate the selection intensity, a shortage of SNPs in the genomic region containing *ABCC11* in Japanese has been reported (Yoshiura et al. 2006). Microsatellite markers are also suitable for studying the selection intensity because of the high degree of polymorphism caused by a high mutation rate (Tishkoff et al. 2001). Thus, in this study, we infer the selection intensity of rs17822931-A in EAS by using a mathematical approach based on the data from two microsatellite loci flanking the *ABCC11* gene and we also estimate the age of rs17822931-A. Furthermore, we discuss the possible factor causing local adaptation of rs17822931-A.

Materials and Methods

Ethics Statement

This study was approved by the Research Ethics Committee of the Graduate School of Comprehensive Human Sciences of the University of Tsukuba.

DNA Samples

DNA samples from HapMap panel subjects, including 58 subjects from YRI, 43 JPT, and 45 CHB, were analyzed in this study.

Genotyping

For genotyping two microsatellites, D16S0513i and D16S0218i, polymerase chain reaction (PCR) was performed using the following sets of primers: 5′-CAAGTCTTAAAGTTCTCAGTGGC-3′ (forward primer) and 5′-GAACAATCAGGATATAAGGAACC-3′ (reverse primer) for D16S0513 and 5′-TGTAATCCTAACTACTCAGGACAC-3′ (forward primer) and 5′-GGACTCTTTATTCTCCAGTGAG-3′ (reverse primer) for D16S0218i. PCR was performed with an initial denaturation at 96 °C for 10 min, followed by 35 cycles of denaturation at 96 °C for 30 s, annealing at 58 °C for 30 s, and extension at 72 °C for 1 min using a thermal cycler (GeneAmp PCR system 9700; Perkin-Elmer Applied Biosystems). The D16S0513i genotype was analyzed by direct sequencing. According to the Gene Diversity DataBase System, D16S0218i consists of AAAAC units; however, the PCR-direct sequencing of D16S0218i detected microsatellite alleles with an AATAC unit. Therefore, to distinguish these alleles from alleles consisting only of AAAAC repeats, we devised a PCR-single-strand conformation polymorphism (PCR-SSCP) method followed by PCR-direct sequencing. Heterozygous genomic DNAs were subjected to PCR-SSCP analysis. Subsequently, each of the separated alleles was recovered from the gel and served as the template for direct sequencing. For PCR-SSCP, 2 µl of solution containing the PCR product was mixed with 6 µl of denaturing solution (95% formamide, 20 mM ethylenediaminetetraacetic acid [EDTA], 0.05% bromophenol blue, 0.05% xylene cyanol FF), and the mixture was denatured at 96 °C for 5 min and immediately cooled on ice. One microliter of each mixture was applied to 10% polyacrylamide gel (acrylamide:bisacrylamide = 49:1) containing 5.0% glycerol. Electrophoresis was carried out for 100 min in 0.5× TBE (45 mM Tris-borate [pH 8.0], 1 mM EDTA) at 20 °C at a constant current of 20 mA/gel using a minigel electrophoresis apparatus with a constant temperature control system (90 × 80 × 1 mm, AE 6410 and AE 6370; ATTO, Tokyo, Japan). Single-stranded DNA fragments in the gel were visualized by SYBR Gold staining. After electrophoresis of DNA fragments through an acrylamide gel, multiple separate DNA bands were excised. The gel pieces were then immersed in 20 µl of TE buffer in a 0.5-ml Eppendorf tube and heated at 95 °C for 10 min. The extract was briefly vortexed and centrifuged. A 1-µl sample of supernatant was then subjected to the second PCR, which was performed under the same conditions as the first PCR but with 25 cycles. Direct sequencing was then performed to complete the genotyping.

Statistical Analysis

The allele frequency and genotype data of SNPs around rs17822931 of *ABCC11* on chromosome 16 were obtained from the HapMap database. Only SNPs analyzed in three HapMap populations (i.e., YRI, EAS, and CEU) were further selected for the calculation of the heterozygosity. The averaged heterozygosity was calculated for SNPs within 25 kb on either side of the focal SNP. To evaluate the structure of LD around rs17822931, the absolute D' values, $|D'|$, and r^2 for all pairwise combinations of SNPs with minor allele frequency of more than 0.05 in each population were estimated using Haploview software (Barrett et al. 2005), and pairwise $|D'|$ values were further visualized as a heat map by use of the GOLD program (Abecasis and Cookson 2000). Deviation from Hardy–Weinberg equilibrium at each microsatellite locus was assessed by Monte Carlo simulation using SNPAlyze version 7.0 (Dynacom, Yokohama, Japan). The frequencies of D16S0513i-rs17822931-D16S0218i haplotypes were estimated using the expectation–maximization algorithm (Excoffier and Slatkin1995) implemented in SNPAlyze version 7.0. The numbers of synonymous (d_S) and nonsynonymous (d_N) substitutions per site between the *ABCC11* coding sequences of human (NM_033151), chimpanzee (XM_001163586), and macaque (XM_001114216) were estimated using DnaSP version 5 (Librado and Rozas 2009) based on the method of Nei and Gojobori (1986). A regression analysis was performed to examine the association between latitude and allele frequency in Asian, Native American, and European populations. The data of latitude for each population and allele frequencies of 47 SNPs (i.e., rs17822931 and 46 nonlinked SNPs) were obtained from the ALFRED database (Osier et al. 2001; Rajeevan et al. 2003; Sanchez et al. 2006; Yoshiura et al. 2006). *P* values less than 0.05 were considered statistically significant.

Coalescent Simulation

Using SelSim program (Spencer and Coop 2004), the variance in the repeat number, S^2, among the simulated chromosomes was computed for various values of $4Nu$ in a single stepwise mutation model without mutational bias. Because the 57 YRI individuals and the 114 YRI chromosomes were successfully genotyped for D16S0513i, we sampled 114 chromosomes in each simulation run and calculated the average S^2 from 1,000 runs. For D16S0218i, a more complicated approach was conducted because the stepwise mutation rate of alleles consisting of only the simple repeat units AAAAC may be different from that of alleles with an AATAC unit. Because the population frequency of a microsatellite allele with an AATAC unit was low in EAS populations (i.e., 2/174), alleles with an AATAC unit were not considered in the estimation of selection intensity. Thus, the mutation parameter for microsatellite alleles with only AAAAC units was estimated as follows. First, we assumed that a point mutation from A to T occurred only once and that the total population frequency of alleles with an AATAC unit has been increased to 0.2 in the YRI population. In the SelSim simulation, two loci, a selectively neutral SNP and microsatellite, with complete linkage (i.e., no recombination between the two loci) were assumed, and microsatellite alleles on the chromosome with a derived allele at the SNP were regarded as ones with an AATAC repeat unit. The SelSim simulation was conducted under the conditions in which the final allele frequency of a derived allele at the SNP was 0.2, which reflects the fact that in YRI the present total frequency of alleles with an AATAC unit is 0.2. After each simulation run, 88 chromosomes with an ancestral allele at the SNP (i.e., microsatellite alleles with only AAAAC units) were sampled, and the average S^2 after 1,000 runs was calculated. The regression line for each microsatellite was obtained from the average S^2.

The SelSim program was also used to estimate the selection intensity. In the simulation, two linked loci were assumed: an SNP with A and G alleles (i.e., rs17822931-A/G) and a microsatellite locus (i.e., D16S0513i or D16S0218i). The relative fitnesses of the AA, AG, and GG genotypes at the SNP were given by 1+*s*, 1, and 1, respectively, where $s > 0$. The present allele frequency of A was set at 0.93 in a population with the size of 3,100. The mutation rate estimated for EAS was applied for each microsatellite locus. The genetic distances between rs17822931 and D16S0513i (0.071 cM) and between rs17822931 and D16S0218i (0.194 cM) were obtained from the HapMap database. For each *s*, 1,000 simulation runs were performed, and the variance of repeat number S^2 for a microsatellite locus among the sampled chromosomes was calculated after each run. The rejection method was used to accept only simulation runs that showed S^2 within 20% of 0.128 for D16S0513i and of 0.190 for D16S0218i.

Estimation of Age

A conventional forward-time computer simulation is helpful for estimating the age of a mutation (Ohashi et al. 2004), but it requires substantial computation time because at least $2N$ random numbers must be drawn in each generation to produce a diploid population with size N. To avoid such a time-consuming process, Monte Carlo experiments using the "pseudo-sampling variable" (PSV), proposed by Kimura (Kimura 1980), were performed, which allows us to shorten the simulation process. Instead of drawing $2N$ random numbers, a single uniform random number is generated with a suitable mean and a variance to produce the allele frequency in the next generation. Assuming recessive selection, the allele frequency of rs17822931-A in the next generation, p', is given by the recursion:

$$p' = \frac{p^2(1+s) + p(1-p)}{p^2(1+s) + 2p(1-p) + (1-p)^2} + \zeta_{\mathrm{PSV}},$$

where p is the allele frequency of rs17822931-A in the present generation, s a selection coefficient as described above, and ξ_{PSV} represents a uniform random variable with mean 0 and variance $\frac{p(1-p)}{2N}$. The first term on the right in the above recursion equation indicates the expected allele frequency of rs17822931-A in the next generation. When a random number, R, uniformly distributed in the range between 0 and 1 (i.e., $E(R) = 1/2$ and $\mathrm{Var}(R) = 1/12$) is used in the computer simulation, R must be transformed to $aR + b$ to attain the mean of 0 and the variance of $\frac{p(1-p)}{2N}$. Because a new random number, $aR + b$, must satisfy the following equations: $E(aR + b) = aE(R) + b = a/2 + b = 0$ and $\mathrm{Var}(aR + b) = a^2\mathrm{Var}(R) = a^2/12 = \frac{p(1-p)}{2N}$, a and b should be $2\sqrt{\frac{3p(1-p)}{2N}}$ and $-\sqrt{\frac{3p(1-p)}{2N}}$, respectively. Accordingly, ξ_{PSV} is given by $(2R-1)\sqrt{\frac{3p(1-p)}{2N}}$ (Hartl and Clark 2007). Because Monte Carlo experiments using the PSV may not provide a good approximation for p,

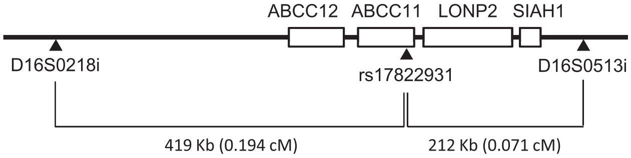

FIG. 2 Location of polymorphisms analyzed in this study. The positions of polymorphisms rs17822931, D16S0513i, and D16S0218i are indicated by arrows. The four genes including *ABCC11* located in this region are indicated by open boxes. The genetic distances between rs17822931 and D16S0513i (0.071 cM) and between rs17822931 and D16S0218i (0.194 cM) were obtained from the HapMap database.

which is close to the absorbing states (i.e., 0 and 1), we evaluated the mean fixation time of a neutral mutation. The obtained time was very close to 4*N* generations, which is expected under neutrality. Thus, the present Monte Carlo experiments using the PSV are suitable for estimation of the age of a mutant allele. In the experiments, we considered the first arrival time at allele frequency of 0.93 (i.e., the present allele frequency of rs17822931-A in EAS) under recessive selection with *s* of 0.01 in a population with size *N* of 3,100 as the age of rs17822931-A. In each simulation run, when the allele frequency of rs17822931-A with an initial frequency of 1/2*N* exceeded 0.93, the number of generations was recorded. The mean and 95% credible interval were calculated for 1,000 successful runs.

Results

Allelic Diversity at Microsatellite Loci

Two microsatellite loci, D16S0513i and D16S0218i, flanking the *ABCC11* gene (fig. 2) and located outside an LD block (supplementary fig. S2, Supplementary Material online) were selected from the Gene Diversity DataBase System and genotyped by PCR-direct sequencing. For a complicated microsatellite, D16S0218i, we devised a PCR-SSCP method followed by PCR-direct sequencing (fig. 3). Accordingly, we successfully determined the genotypes for D16S0513i and D16S0218i in the YRI and EAS populations. The allele frequencies of these two microsatellites in YRI and EAS are presented in Table 1. In the YRI population, several microsatellite alleles with an AATAC unit were observed. Among these, $(AAAAC)_6AATAC(AAAAC)_2$ was the most frequent, and the other three alleles with an AATAC could have been generated by a single slippage mutation of $(AAAAC)_6AATAC(AAAAC)_2$. Thus, it is speculated that a point mutation from A to T may have occurred in the $(AAAAC)_9$ allele.

In the Hardy–Weinberg equilibrium test, only D16S0513i in EAS, wherein more heterozygotes were observed than expected, showed a *P* value below 0.05 ($P = 0.017$). Although the excess of heterozygotes was mainly due to the observed $(TC)_{10}/(TC)_{12}$ heterozygotes (the individual genotype data are provided in supplementary table S1, Supplementary Material online), it was unlikely to have been caused by a genotyping error. In addition, this *P* value might not necessarily be considered significant when multiple testing was taken into account. Thus, the EAS data were used for further analyses.

The estimated frequencies of D16S0513i-rs17822931-D16S0218i haplotypes were different between the YRI and EAS populations (supplementary tables S2 and S3, Supplementary Material online). The haplotype diversity was lower in EAS than in YRI. Among haplotypes consisting of rs17822931-A, $(TC)_{11}$-A-$(AAAAC)_7$ was most frequent in EAS (supplementary table S3, Supplementary Material online), suggesting that the rs17822931-A mutation occurred in $(TC)_{11}$-G-$(AAAAC)_7$, which is also the most frequent haplotype in YRI (supplementary table S2, Supplementary Material online).

To evaluate the allelic diversities at microsatellite loci in YRI and EAS, we calculated the variance of repeat number, S^2. The variance of repeat number, S^2, is defined

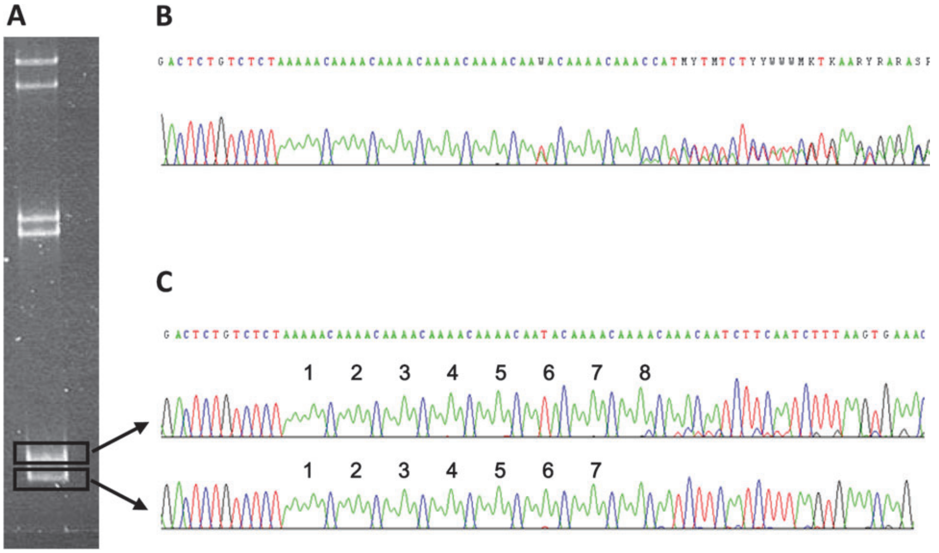

FIG. 3 Genotyping for D16S0218i. (*A*) Single-strand DNA fragments separated by electrophoresis in PCR-SSCP analysis. (*B*) The sequence obtained from PCR-direct sequencing for a heterozygote sample. (*C*) The sequence obtained from PCR-direct sequencing for each band. Accordingly, this sample was determined to be a heterozygote of $(AAAAC)_7$ and $(AAAAC)_5(AATAC)(AAAAC)_2$.

Table 1. Allele Frequencies at rs17822931, D16S0513i, and D16S0218i.

		Population[a]	
Polymorphisms	**Alleles**	**YRI (2*N* = 116)**	**EAS (JPT + CHB) (2*N* = 176)**
rs1782293	A	0 (0%)	164 (93%)
	G	116 (100%)	12 (7%)
	Total	116 (100%)	176 (100%)
D16S0513i	$(TC)_9$	27 (24%)	0 (0%)
	$(TC)_{10}$	4 (3%)	4 (2%)
	$(TC)_{11}$	63 (55%)	147 (87%)
	$(TC)_{12}$	20 (18%)	19 (11%)
	Total	114 (100%)	170 (100%)
D16S0218i	$(AAAAC)_4$	0 (0%)	1 (1%)
	$(AAAAC)_6$	0 (0%)	6 (3%)
	$(AAAAC)_7$	67 (61%)	147 (84%)
	$(AAAAC)_8$	7 (6%)	18 (10%)
	$(AAAAC)_9$	13 (12%)	0 (0%)
	$(AAAAC)_{10}$	1 (1%)	0 (0%)
	$(AAAAC)_5(AATAC)(AAAAC)_2$	7 (6%)	2 (1%)
	$(AAAAC)_6(AATAC)(AAAAC)_1$	1 (1%)	0 (0%)
	$(AAAAC)_6(AATAC)(AAAAC)_2$	12 (11%)	0 (0%)
	$(AAAAC)_6(AATAC)(AAAAC)_2$	2 (2%)	0 (0%)
	Total	110 (100%)	174 (100%)

[a] Only HapMap samples with rs17822931 genotypes available were analyzed.

as $S^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \bar{X})^2$, where m is the number of sampled chromosomes, X_i is the repeat number of ith chromosome, and $\bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_i$. The observed S^2 values at D16S0513i for the 114 YRI and 170 EAS chromosomes were 1.056 and 0.128, respectively. When alleles consisting only of AAAAC repeats were considered, the S^2 values at D16S0218i for YRI and EAS were 0.612 and 0.190, respectively. As expected from the degrees of heterozygosity (fig. 1) and LD (supplementary fig. S2, Supplementary Material online) based on SNPs in a genomic region studied, the variance in repeat number or the allelic diversity at the microsatellite locus was lower in EAS than in YRI owing to the selective sweep in the former population.

Mutation Parameter (4*Nu*)

In this study, we used S^2 as the measurement in the computer simulation to infer the selection intensity of rs17822931-A. First, the stepwise mutation rates at the flanking microsatellite loci in EAS had to be estimated; however, it is difficult to estimate the mutation parameters $4Nu$ (where N is the population size and u is the single stepwise mutation rate) at D16S0513i and D16S0218i in EAS by a coalescent approach because these loci have been strongly influenced by positive selection operating at rs17822931. Unlike in EAS, the two microsatellite loci appear to have been free from positive selection in YRI, wherein the rs17822931-A allele was not observed and there was no reduction in heterozygosity around the *ABCC11* gene (fig. 1). Thus, the YRI data can be used to estimate the mutation parameter $4Nu$ under the assumption of selective neutrality. In brief, using a coalescent simulation program, SelSim, the variance in the repeat number, S^2, among the simulated chromosomes was computed for various values of $4Nu$ under neutrality, and then the linear regression equation for the simulated data was calculated for each microsatellite (see Materials and Methods for details). For D16S0513i, the linear regression equation for the simulated data was $S^2 = -0.037 + 4Nu \times 0.513$, and the S^2 observed at D16S0513i in YRI was 1.056 (fig. 4). Thus, $4Nu$ at D16S0513i in YRI was estimated to be 2.131.

For D16S0218i, a more complicated approach was required because a point mutation from A to T (i.e., from AAAAC to AATAC) was found within a repeat in the YRI population (table 1). The mutation rate of alleles consisting of only simple repeat units AAAAC may be different from that of alleles with an AATAC unit. For simplicity, we did not consider alleles with an AATAC unit in the estimation of selection intensity because the population frequency of a microsatellite allele with an AATAC unit was low (i.e., 2/174) in EAS (table 1). The regression equation for simulated chromosomes with microsatellite alleles consisting only of AAAAC was $S^2 = 0.037 + 4Nu \times 0.420$ (fig. 4). Substituting the observed S^2 of 0.612 in this equation for 88 YRI chromosomes with only AAAAC units, $4Nu$ for AAAAC repeats at D16S0218i in YRI was estimated to be 1.368.

Estimation of Selection Coefficient

Recently, it was reported that the effective population size for CEU, JPT, and CHB is approximately 3,100, whereas for YRI, the approximate size is 7,500 (Tenesa et al. 2007). Thus, assuming that there is no difference in mutation rates at D16S0513i and D16S0218i between YRI and EAS, $4Nu$ for D16S0513i and D16S0218i in EAS were estimated to be 0.881 ($= 2.131 \times 3{,}100/7{,}500$) and 0.565 ($= 1.368 \times 3{,}100/7{,}500$), respectively, under neutrality. Here, we further assumed that the effective population size for EAS (JPT + CHB) was close to that for JPT or CHB

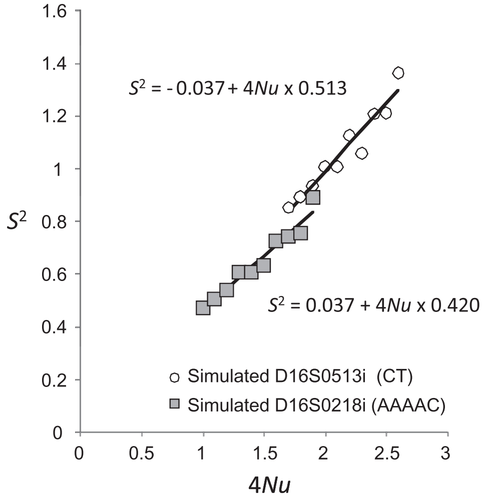

Fig. 4 Relationship between 4*Nu* and S^2. Using a coalescent simulation, the variance of repeat number S^2 among the simulated chromosomes was computed for various values of 4*Nu*. For each 4*Nu*, 1,000 runs were performed and the open circle and shaded square indicate the average values of S^2 for D16S0513i and AAAAC repeats of D16S0218i, respectively. The linear regression equations for the simulated data are also shown.

because these East Asian populations share recent common ancestors.

To estimate the selection intensity, we again used SelSim. In the simulation, we assumed two linked loci: a SNP (i.e., rs17822931-A/G) and a microsatellite locus (i.e., D16S0513i or D16S0218i). Because the rs17822931-A allele leads to dry earwax in a recessive manner, the relative fitnesses of the AA, AG, and GG genotypes at rs17822931 were given by 1 + *s*, 1, and 1, respectively (i.e., recessive selection model). The present population frequency of rs17822931-A was set as 0.93 as observed in EAS, and the population size, *N*, was 3,100 in the simulation. The mutation rate estimated above for EAS was applied for each microsatellite locus. The genetic distances between rs17822931 and D16S0513i and between rs17822931 and D16S0218i were assumed to be 0.071 cM and 0.194 cM (fig. 2). For each *s*, we conducted 1,000 simulation runs, and S^2 for a microsatellite locus among the sampled chromosomes was calculated after each run. We used the rejection method to accept only simulation runs that resembled the observed value of S^2 in EAS when the run was terminated (i.e., S^2 of accepted runs were within 20% of 0.128 for D16S0513i and 0.190 for D16S0218i, respectively). Finally, the number of accepted runs in 1,000 trails was counted (fig. 5A). The peaks of D16S0513i and D16S0218i were observed at *s* of 0.02 and 0.002, respectively. When both results were combined, the peak was found at *s* of 0.01. Thus, we conclude that the selection coefficient of rs17822931-A has been approximately 0.01 (i.e., scaled selection coefficient 2*Ns* = 62) in East Asians. Because microsatellite loci can be polymorphic owing to the high mutation rate even after a strong selective sweep, the present approach using microsatellite markers near a polymorphism subjected to positive selection should be useful for estimating the selection intensity.

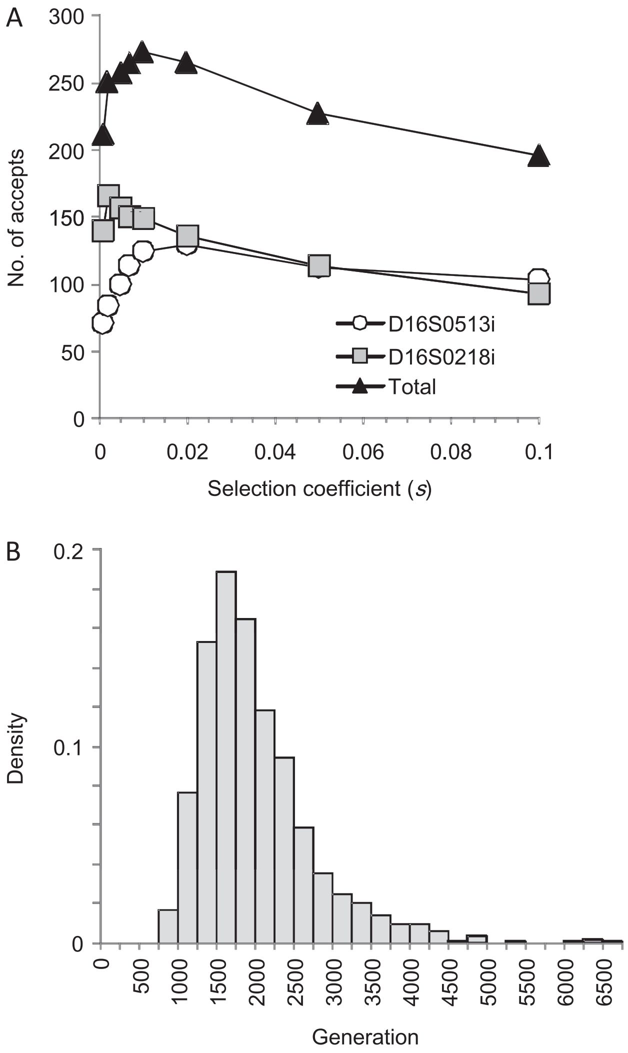

Fig. 5 Selection intensity and age of rs17822931-A. (*A*) Numbers of accepts in 1,000 simulation runs for D16S0513i (open circle) and D16S0218i (shaded square). The total (closed triangle) indicates the sum of the numbers of accepts for D16S0513i and D16S0218i. The maximum peak was observed at *s* = 0.01 for the total. (*B*) Frequency distribution of the age of rs17822931-A obtained from the simulation experiments using a PSV under a recessive selection model assuming *N* = 3100 and *s* = 0.01. The mean age of rs17822931-A was 2,006 generations, and the 95% credible interval was 1,023–3,901 generations.

Age of rs17822931-A

Next, we estimated the age of rs17822931-A by using Monte Carlo experiments. We considered the first arrival time at an allele frequency of 0.93 (i.e., the present allele frequency of rs17822931-A in EAS) under recessive selection with *s* of 0.01 in a population with size *N* of 3,100 as the age of rs17822931-A. It should be noted that we did not consider cases where rs17822931-A was lost from a population without an excess of the allele frequency of 0.93 in the calculation of the mean age. Figure 5*B* shows the frequency distribution of age in 1,000 successful runs. The

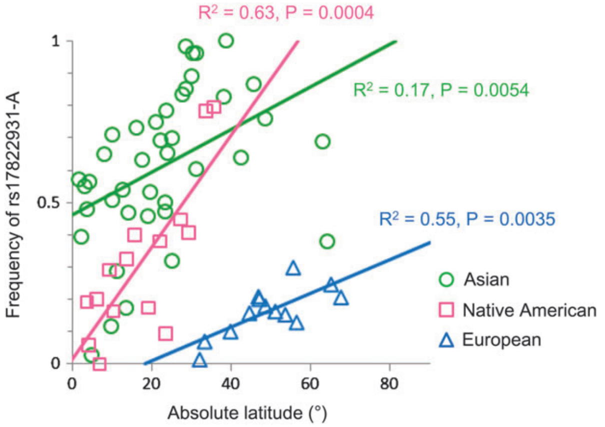

Fig. 6 Significant association of absolute latitude with the frequency of rs17822931-A. Linear regression lines are shown for plots of Asian, Native American, and European populations. The standardized regression coefficients are 0.43 ($P = 0.0054$), 0.79 ($P = 0.0004$), and 0.74 ($P = 0.0035$), respectively.

mean age of rs17822931-A was 2,006 generations, and the 95% credible interval was 1,023–3,901 generations.

The d_N/d_S Ratio

To assess the selective constraint on *ABCC11*, we estimated the numbers of synonymous (d_S) and nonsynonymous (d_N) substitutions per site between the *ABCC11* coding sequences of human (NM_033151) and chimpanzee (XM_001163586) (table 2). The obtained d_N/d_S ratio 0.24 was less than 1. Furthermore, the d_N/d_S ratio between human and macaque (XM_00111421) sequences (0.40) was lower than that between chimpanzee and macaque (0.45).

Association between Latitude and Allele Frequency of rs17822931-A

Finally, we examined the association between latitude and the rs17822931-A allele frequency in worldwide populations in the ALFRED database (Osier et al. 2001; Rajeevan et al. 2003) (fig. 6). Interestingly, absolute latitude was significantly associated with the population frequency of rs17822931-A across regions (i.e., Asians, Native Americans, and Europeans), even though the populations have different histories of migration and expansion. Here, African populations were not included in the analysis because of the absence of rs17822931-A. Also, Oceanic populations were not evaluated because of the small sample size. To evaluate if absolute latitude is commonly associated with allele frequency in human populations, 46 nonlinked SNPs studied by Sanchez et al. (2006) were further analyzed as controls (supplementary fig. S4 and table S4, Supplementary Material online). Significant associations (P value $<$ 0.05) were detected for several SNPs. However, unlike rs17822931, no control SNP showed significant association in all three regions.

Discussion

To identify the signature of recent positive selection in human populations, a number of LD-based methods have been developed (Sabeti et al. 2002; Voight et al. 2006; Kimura et al. 2007; Grossman et al. 2010). However, such methods generally do not allow us to estimate the selection intensity of an advantageous allele. This study suggests that the selection coefficient of rs17822931-A has been 0.01 in East Asians (fig. 5*A*). The present approach using microsatellite markers near a polymorphism subjected to positive selection should be useful for estimating the selection intensity of an advantageous allele even after a strong selective sweep.

The estimated age of rs17822931-A was 2,006 generations. Although the 95% credible interval was large (1,023–3,901 generations; fig. 5*B*), the date of the rs17822931-A mutation is in agreement with the generally accepted "Out of Africa" scenario; the date for the divergence of African and non-African populations is 3,500 generations ago and that for the divergence of European and Asian populations is 2,000 generations ago (Schaffner et al. 2005). When rs17822931-A was assumed to be selectively neutral (i.e., $s = 0$), the mean age of rs17822931-A was 10,038 generations (95% credible interval: 3,742–23,796 generations). The geographical distribution of rs17822931-A (supplementary fig. S1, Supplementary Material online) suggests that rs17822931-A appeared after an "Out of Africa" event. Thus, we may conclude that, without the assumption of positive selection, it is difficult to explain the rapid increase in the allele frequency of rs17822931-A in East Asian populations.

Polymorphisms that are directly subjected to natural selection must have functional significance. The functional significance of rs17822931-G/A (G180R) has been recently described (Toyoda et al. 2009). Immunofluorescence staining of human ABCC11 protein in tissue specimens containing ceruminous apocrine glands from human subjects with the rs17822931-G/A (180G/R) or rs17822931-A/A (180R/R) genotype revealed that the ABCC11 G180 protein is localized in the intracellular granules and large vacuoles, whereas neither granular nor vacuolar localization was detected for R180. Furthermore, the level of R180 protein was lower than that of G180 due to its proteasomal degradation

Table 2. The Rates of Nonsynonymous Substitutions (d_N) and Synonymous Substitutions (d_S).

		Nonsynonymous			Synonymous			
Species 1	**Species 2**	**No. of Differences**	**No. of Sites**	d_N	**No. of Differences**	**No. of Sites**	d_S	d_N/d_S **Ratio**
Human	Chimpanzee	17	3152.17	0.0054	22	993.83	0.0225	0.24
Human	Macaque	115.5	3149.67	0.0376	87.5	996.33	0.0934	0.40
Chimpanzee	Macaque	119.5	3150.5	0.0389	80.5	995.5	0.0856	0.45

Note.—The *ABCC11* mRNA sequences of human, chimpanzee, and macaque are NM_033151, XM_001163586, and XM_001114216, respectively. The numbers of differences and sites were calculated based on the method of Nei and Gojobori (1986).

(Toyoda et al. 2009). Thus, the rs17822931-A mutation seems to cause a loss of function of the ABCC11 protein. To examine if the ABCC11 protein has no selective constraint, the d_N/d_S ratios of human, chimpanzee, and macaque were calculated (table 2). The results suggest that purifying selection has operated against the *ABCC11* gene in the human lineage after divergence from chimpanzee. Nevertheless, the frequency of rs17822931-A, a loss of function mutation, has increased in non-African populations. Thus, an "Out of Africa" event may have removed a long-standing selective constraint on the *ABCC11* gene.

What is the cause of the selective advantage of rs17822931-A? Although the physiological function of earwax is poorly understood (Matsunaga 1962), dry earwax itself is unlikely to have provided a substantial advantage. The rs17822931-GG and GA genotypes (wet earwax) are also strongly associated with axillary osmidrosis, suggesting that the ABCC11 protein has an excretory function in the axillary apocrine gland (Nakano et al. 2009). The ancestral environment of East Asians is thought to have been much colder than that of Africans. As Yoshiura et al. (2006) suggested, the ancestors of East Asians with the rs17822931-AA genotype might have had some selective advantages, such as less sweating, as an adaptation to the cold climate. Because average daily and annual temperatures, on the whole, correlate with absolute latitude (New et al. 2002), in order to examine the "cold adaptation hypothesis," the association between absolute latitude and allele frequency was evaluated for 47 SNPs including rs17822931 in worldwide populations. Absolute latitude was found to be associated with rs17822931-A allele frequency across three regions (i.e., Asian, Native American, and European populations), whereas significant associations in all three regions were not observed for any control SNPs (supplementary fig. S4 and table S4, Supplementary Material online). Thus, we conclude that absolute latitude is not always associated with allele frequency across human populations and that the geographical distribution of rs17822931-A is rather an exception. Therefore, the observed latitudinal cline in the rs17822931-A allele frequency strongly supports the cold adaptation hypothesis, although the possibility that other conditions associated with latitude (e.g., sunlight and microbial environment) might play a role cannot be excluded.

Despite great progress in understanding the associations between genetic variants and traits in humans owing to genome-wide association study using a number of SNP markers, the effect of local adaptation on the difference in phenotype among human populations remains to be studied. The present study provides a striking example of how local adaptation has played a significant role in the phenotypic diversification of human traits.

Supplementary Material

Supplementary figures S1, S2, S3, and S4 and tables S1, S2, S3, and S4 are available at *Molecular Biology and Evolution online* (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank two anonymous reviewers for their constructive comments and valuable suggestions on the manuscript. We thank Dr Flaminia Miyamasu for proofreading of the revised manuscript. This work was partly supported by a KAKENHI (22370084) Grant-in-Aid for Scientific Research (B) from the Ministry of Education, Culture, Sports, Science, and Technology of Japan.

References

Abecasis GR, Cookson WO. 2000. GOLD–graphical overview of linkage disequilibrium. *Bioinformatics* 16:182–183.

Barrett JC, Fry B, Maller J, Daly MJ. 2005. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21:263–265.

Excoffier L, Slatkin M. 1995. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol.* 12:921–927.

Grossman SR, Shylakhter I, Karlsson EK, et al. (13 co-authors). 2010. A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science* 327:883–886.

Hartl DL, Clark AG. 2007. Principles of population genetics. Sunderland (MA): Sinauer Associates.

Kimura M. 1980. Average time until fixation of a mutant allele in a finite population under continued mutation pressure: studies by analytical, numerical, and pseudo-sampling methods. *Proc Natl Acad Sci U S A.* 77:522–526.

Kimura R, Fujimoto A, Tokunaga K, Ohashi J. 2007. A practical genome scan for population-specific strong selective sweeps that have reached fixation. *PLoS One.* 2:e286.

Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25:1451–1452.

Matsunaga E. 1962. The dimorphism in human normal cerumen. *Ann Hum Genet.* 25:273–286.

Nakano M, Miwa N, Hirano A, Yoshiura K, Niikawa N. 2009. A strong association of axillary osmidrosis with the wet earwax type determined by genotyping of the ABCC11 gene. *BMC Genet.* 10:42.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol.* 3:418–426.

New M, Lister D, Hulme M, Makin I. 2002. A high-resolution data set of surface climate over global land areas. *Climate Res.* 21:1–25.

Ohashi J, Naka I, Patarapotikul J, Hananantachai H, Brittenham G, Looareesuwan S, Clark AG, Tokunaga K. 2004. Extended linkage disequilibrium surrounding the hemoglobin E variant due to malarial selection. *Am J Hum Genet.* 74:1198–1208.

Osier MV, Cheung KH, Kidd JR, Pakstis AJ, Miller PL, Kidd KK. 2001. ALFRED: an allele frequency database for diverse populations and DNA polymorphisms–an update. *Nucleic Acids Res.* 29:317–319.

Rajeevan H, Osier MV, Cheung KH, et al. (13 co-authors). 2003. ALFRED: the ALelle FREquency Database. *Update Nucleic Acids Res.* 31:270–271.

Sabeti PC, Reich DE, Higgins JM, et al. (17 co-authors). 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419:832–837.

Sanchez JJ, Phillips C, Borsting C, et al. (13 co-authors). 2006. A multiplex assay with 52 single nucleotide polymorphisms for human identification. *Electrophoresis* 27:1713–1724.

Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. 2005. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* 15:1576–1583.

Spencer CC, Coop G. 2004. SelSim: a program to simulate population genetic data with natural selection and recombination. *Bioinformatics* 20:3673–3675.

Tenesa A, Navarro P, Hayes BJ, Duffy DL, Clarke GM, Goddard ME, Visscher PM. 2007. Recent human effective population size estimated from linkage disequilibrium. *Genome Res.* 17:520–526.

The International HapMap Consortium. 2003. The International HapMap Project. *Nature* 426:789–796.

The International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* 437:1299–1320.

Tishkoff SA, Varkonyi R, Cahinhinan N, et al. (17 co-authors). 2001. Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. *Science* 293:455–462.

Toyoda Y, Sakurai A, Mitani Y, et al. (12 co-authors). 2009. Earwax, osmidrosis, and breast cancer: why does one SNP (538G>A) in the human ABC transporter ABCC11 gene determine earwax type? *Faseb J.* 23:2001–2013.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. *PLoS Biol.* 4:e72.

Wang Z, Wang J, Tantoso E, Wang B, Tai AY, Ooi LL, Chong SS, Lee CG. 2007. Signatures of recent positive selection at the ATP-binding cassette drug transporter superfamily gene loci. *Hum Mol Genet.* 16:1367–1380.

Yoshiura K, Kinoshita A, Ishida T, et al. (39 co-authors). 2006. A SNP in the ABCC11 gene is the determinant of human earwax type. *Nat Genet.* 38:324–330.