Gene Flow and Natural Selection in Oceanic Human Populations Inferred from Genome-Wide SNP Typing

Ryosuke Kimura,* Jun Ohashi,† Yasuhiro Matsumura,‡ Minato Nakazawa,§ Tsukasa Inaoka,∥ Ryutaro Ohtsuka,¶ Motoki Osawa,* and Katsushi Tokunaga†

*Department of Forensic Medicine, Tokai University School of Medicine, Kanagawa, Japan; †Department of Human Genetics, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan; ‡Division of Health Informatics and Education, National Institute of Health and Nutrition, Tokyo, Japan; §Socio-Environmental Health Sciences, Graduate School of Medicine, Gunma University, Gunma, Japan; ∥Department of Environmental Sociology, Faculty of Agriculture, Saga University, Saga, Japan; and ¶National Institute for Environmental Studies, Ibaraki, Japan

It is suggested that the major prehistoric human colonizations of Oceania occurred twice, namely, about 50,000 and 4,000 years ago. The first settlers are considered as ancestors of indigenous people in New Guinea and Australia. The second settlers are Austronesian-speaking people who dispersed by voyaging in the Pacific Ocean. In this study, we performed genome-wide single-nucleotide polymorphism (SNP) typing on an indigenous Melanesian (Papuan) population, Gidra, and a Polynesian population, Tongans, by using the Affymetrix 500K assay. The SNP data were analyzed together with the data of the HapMap samples provided by Affymetrix. In agreement with previous studies, our phylogenetic analysis indicated that indigenous Melanesians are genetically closer to Asians than to Africans and European Americans. Population structure analyses revealed that the Tongan population is genetically originated from Asians at 70% and indigenous Melanesians at 30%, which thus supports the so-called Slow train model. We also applied the SNP data to genome-wide scans for positive selection by examining haplotypic variation and identified many candidates of locally selected genes. Providing a clue to understand human adaptation to environments, our approach based on evolutionary genetics must contribute to revealing unknown gene functions as well as functional differences between alleles. Conversely, this approach can also shed some light onto the invisible phenotypic differences between populations.

Introduction

The peopling of Oceania has intrigued anthropologists because it is one of the most mysterious adventures in human history. The first colonization of New Guinea and Australia by modern humans is thought to have occurred by about 50 thousand years ago (KYA) when these lands formed a continent called Sahul (White and O’Connell 1979; Roberts et al. 1990). These first settlers are considered as ancestors of indigenous Melanesians (Papuans) and Australians, who are anthropologically classified into the Australoid. The second major migration to Oceania was made about 4 KYA by Austronesian-speaking people (Bellwood 1989, 1991). They voyaged to Polynesia during 1-3 KYA. As for the origin of Polynesians, essentially 2 opposite models have been proposed. One is called the “Express train” model, which supposes that Austronesian-speaking people rapidly dispersed to Polynesia with negligible admixture with indigenous Melanesians (Diamond 1988). The other is the “Entangled bank” model hypothesizing that Polynesians were derived from Melanesian populations affected by social and trade networks with people from Southeast Asia (Terrell 1988). Between these polar opposites, there are models such as the “Slow train” that suggests significant genetic contributions from both original Austronesian-speaking people and indigenous Melanesians.

To elucidate the origin of these populations, genetic evidence is most direct and informative. It has been well known that indigenous Melanesians and Australians, who have the phenotypes similar to Africans in visible traits such as skin color and hair shape, are genetically closer to Asians than to Africans and Europeans (Nei and Roychoudhury 1993; Zhivotovsky et al. 2004). This fact indicates that the first settlers of Oceania have shared a common ancestry with Asians after the divergence from Europeans. The genetic origin of Polynesians is still controversial although a number of studies have focused on this point. The analyses of mitochondrial DNA (mtDNA) have supported an Asian origin of Polynesians without admixture with indigenous Melanesians as expected in the Express train model (Lum et al. 1994, 1998; Melton et al. 1995; Redd et al. 1995), whereas the analyses of the Y chromosome have revealed that indigenous Melanesians predominantly contributed to the genetic components of Polynesians as described in the Entangled bank model (Kayser et al. 2000; Su et al. 2000; Capelli et al. 2001). However, both mtDNA and the Y chromosome are haploids that transmit without recombination. Classical studies analyzing autosomal markers have shown different results depending on the marker used (Hill et al. 1985; Serjeantson 1985; O’Shaughnessy et al. 1990; Cavalli-Sforza et al. 1994; Martinson 1996; Serjeantson and Gao 1996). Therefore, analyses using a large number of autosomal loci are required for further elucidation.

The first settlers of Oceania, that is, indigenous Melanesians and Australians, must have been exposed to various selective pressures due to environmental differences during the long migration from Africa and due to the uniqueness of environments in Oceania after the settlement. They were isolated until recent time and developed their own lifestyle, which might have thus resulted in new selective pressures. Especially, the slow growth, short stature, and lightweight characteristics of New Guineans are generally assumed to reflect an adaptation to the low energy and nutrient densities of diets in which tubers and root crops predominate (Norgan 1995). In contrast, a distinctive characteristic of Polynesians is their large body size. From this phenotype, there has been implied the presence of a “thifty...
genotype” that is associated with saved energy expenditure and efficient fat storage (Neel 1962; Bindon and Baker 1997). An alternative explanation for Polynesians’s large body size is based on the “Bergman’s rule,” a principle that correlates body mass with environmental temperature (Houghton 1990; Bindon and Baker 1997). However, the validity of the thrifty genotype and Bergman’s rule hypotheses in Polynesians is still open to debate.

Recent advances in DNA technologies have now enabled us to perform genome-wide single-nucleotide polymorphism (SNP) typing. The Affymetrix GeneChip Human 500K arrays used in this study are commercially provided DNA chips that can genotype about 500,000 SNPs for each individual. The preponderant number of typing data would assure us of accurate estimation of the admixture rate between Asian and Melanesian lineages in Polynesians. Moreover, genome-wide SNP data are applicable to genome-wide scans for genetic regions under positive selection. Several researchers have recently developed methods to identify signatures of positive selection from SNP data based on hitchhiking events and selective sweeps and have conducted genome-wide scans using SNP databases from the HapMap project and Perlegen Sciences (Kim and Stephan 2002; Sabeti et al. 2002; Nielsen et al. 2005; Voight et al. 2006; Wang et al. 2006; Kimura et al. 2007; Tang et al. 2007; Williamson et al. 2007). The strategy based on evolutionary genetics has provided cues to reveal genotype–phenotype association (Fujimoto et al. 2008; Kayser, Liu, et al. 2008).

The present study investigates the peopling of Oceania with a special focus on the admixture rate between Asian and indigenous Melanesian lineages in Polynesians. We also performed genome-wide scans for positive selection on Oceanic populations. For these purposes, we subjected an indigenous Melanesian (Papuan) population, Gidra, and a Polynesian population, Tongans, for genome-wide SNP typing with the Affymetrix GeneChip Human 500K array set.

Materials and Methods

Samples

Individuals from 2 Oceanic populations, Gidra in Papua New Guinea (GDP samples, n = 24) and Tongans from Nukualofa, Kingdom of Tonga (TGN samples, n = 24), were subjected to our study. The Gidra are Papuan-speaking people that inhabit the lowlands of Western Province, Papua, New Guinea. This population has been reported to have a small size and to be isolated (Ohtsuka 1986). Tongans are Austronesian language–speaking people in Polynesia. Informed consent for participation was obtained from all the subjects in their own language. This study was approved by the Research Ethics Committee at The University of Tokyo. In addition to these populations, we referred to genotyping data of the HapMap samples that were publicly provided by Affymetrix (http://www.affymetrix.com). The HapMap samples included 30 trios (90 individuals) each from Yoruba from Ibadan in Nigeria (YRI) and US residents with ancestry from northern and western Europe (CEU) and 45 unrelated individuals each from Han Chinese from Beijing (CHB) and Japanese from Tokyo (JPT) (The International HapMap Consortium 2005). Here, East Asians (EAS) denote JPT and CHB together. We selected only unrelated individuals (parents in trios) from YRI and CEU and also removed a JPT individual that showed high inbreeding. In total, 209 HapMap individuals (60 YRI; 60 CEU; 45 CHB; and 44 JPT) were thus used for our analyses.

Genotyping and Data Quality Control

SNP genotyping was performed with the Affymetrix GeneChip GeneChip Human 500K array set. In brief, genomic DNA (250 ng) was digested with a restriction enzyme (NspI or Styl) and ligated to adaptors that recognize the cohesive 4-bp overhangs. These fragments were amplified with polymerase chain reaction using a generic primer that recognizes the adaptor sequence. The amplified DNA was then fragmentated, labeled, and hybridized to a microarray chip. The chip was scanned with Affimetric GeneChip Scanner 3000. The genotypes were determined with GeneChip Genotyping Analysis Software based on the Dynamic Model algorithm, in which a strict confidence threshold of $P = 0.26$ was selected. Only the autosomal SNPs (490,031 SNPs) were analyzed in this study. The SNPs were filtered with a criterion of missing rate <0.25 in every population (supplementary table S1, Supplementary Material online). According to our typing, the missing rates for GDP and TGN were slightly high probably due to DNA quality. We excluded SNPs with $P < 0.01$ in chi-square test for the Hardy–Weinberg equilibrium, which accounted for 0.024–0.033 of the polymorphic loci (2.4–3.3 times higher than expected), because it was highly possible for these SNPs to be mistyped or to be located on copy number variations. We also removed those SNPs that were monomorphic in all the populations. Finally, 393,971 autosomal SNPs remained. Because all the SNPs covered 2.7 Gbp of the genome, the average SNP interval was 6.8 kbp/SNP.

$F_{ST}$ between Populations and Phylogenetic Tree among Individuals

For each SNP, we calculated $F_{ST}$ between pairs of populations. The genetic distance between each pair of individuals was calculated simply from the average nucleotide difference of 2 chromosomes drawn at random from different individuals. For locus $l$, the nucleotide difference between individuals $x$ and $y$ is defined as $h_{xy,l} = (d_{l1} + d_{l2} + d_{l3} + d_{l4})/4$, where indicator $d_{ab}$ is 1 when chromosome $a$ in individual $x$ is different from chromosome $b$ in individual $y$ and zero when otherwise. For biallelic loci, $h_{xy,l}$ can only take 3 values: 0 (e.g., AA:AA), 1/2 (e.g., AA:AB or AB:AB), and 1 (e.g., AA:BB). The average nucleotide difference between 2 individuals $x$ and $y$ ($H_{xy}$) can be obtained by averaging $h_{xy,l}$ over $L$ analyzed loci. Suppose $a_{x}$ and $b_{x}$ are the frequency of the allele $A$ and $B$, respectively, at locus $l$ in the population $x$, $E(H_{xy}) = 2a_{x}b_{x}d_{AB2}/L = D_{xy}$ when individuals $x$ and $y$ are randomly extracted from the same
population \( X \). On the other hand, \( E(H_{x,y}) = \Sigma (1 - a_x a_y - b_x b_y)/L \) \((=D_{xy})\) when individuals \( x \) and \( y \) are derived from different populations \( X \) and \( Y \). Therefore, under a large number of loci analyzed, every \( H_{xy} \) value becomes nearly equal to \( D_X \) or \( D_{XY} \). From a distant matrix obtained, we constructed phylogenetic trees of individuals using the Neighbor-Joining method (Saitou and Nei 1987) with Molecular Evolutionary Genetics Analysis version 3.1 (Kumar et al. 2004). The length of the outer branch for an individual in the phylogenetic trees (fig. 1A and B) is nearly equal to \( D_X/2 \), whereas the length of the inner branch between 2 populations is nearly equal to the Nei’s (1973) minimum genetic distance, \( D_m = D_{XY} - (D_X + D_Y)/2 \). We also performed multidimensional scaling (MDS) analyses using the distance matrix for individuals to observe the homogeneity of the populations (Kruskal and Wish 1978).

**Population Structure Analysis**

A cluster analysis for population structure was performed using the STRUCTURE version 2.0 software program (Pritchard et al. 2000). Because a number of SNPs were typed and neighboring SNPs were in strong linkage disequilibrium (LD) with each other, 1/100 of the “typed SNPs” (3,940 SNPs) were selected for a set, and 4 different sets were used for this analysis. In all runs of the STRUCTURE algorithm, we used 10,000 Markov chain Monte Carlo replications after a burn-in of length 10,000, with a model of correlated allele frequencies. The number of groups \( k \) was assumed to be 2–5. \( (D \) and \( E) \) LD coefficients and physical distance. To calculate the LD coefficients, 48 chromosomes were used in each population.

**Estimation of Haplotypes and Missing Genotypes**

The estimation of the haplotypes and missing genotypes was performed with fastPHASE version 1.2 (Scheet and Stephens 2006). We used 5 random starts of the expectation-maximization algorithm with population label information. An allele frequency spectrum for each population was drawn after estimating the missing genotypes. The LD coefficients, \( D’ \) and \( r^2 \), for each population were also calculated using 48 chromosomes when the physical distance between 2 SNPs was less than 250 kb. Although haplotype estimation may be inaccurate, especially for rare haplotypes in the presence of low LD, the accuracy in the frequency of major haplotypes would be retained to some extent. Therefore, the inaccuracy in haplotype estimation is thought to have only a slight effect on the following analyses for scanning positive selection.
The extended haplotype homozygosity (EHH) statistic is defined as the probability that any 2 chromosomes of a particular core allele have the same extended haplotype. The unbiased estimate of this statistic is calculated as:

\[ \text{EHH}_A = \frac{\sum_{i=1}^{m} \left( e_{Ai} \right)^2}{\sum_{i=1}^{m} e_{Ai}} \]

where \( e_{Ai} \) denotes the number of chromosomes with a particular allele \( A \), \( e_{Ai} \) denotes the number of chromosomes with \( i \)th extended haplotype, and \( m \) denotes the number of extended haplotypes.

To detect an unusual increase in EHH, we can compare the target allele and the other allele at the core SNP. In this study, every SNP with a minor allele frequency \( \geq 10\% \) was subjected to EHH computation. The EHH value for the target allele (EHH\(_T\)) was calculated in the range from the core SNP to the position just before EHH\(_T\) drops below 0.4, where we do not need to use the physical (bp) or genetic (cM) distance to decide the range for calculation (supplementary fig. S1, Supplementary Material online). In comparison to the integrated EHH (iEHH) reported previously (Voight et al. 2006) in which EHH are integrated until EHH reaches 0.05, our definition of EHH calculation hardly changes the detection power but dramatically reduces the computational time (data not shown). The EHH value for the other reference allele (EHH\(_R\)) was calculated also in the same range. Here, EHH\(_R/EHH_T\), which is reciprocal to the relative EHH defined generally, was used as a statistic because EHH\(_R\) is zero when all the extended haplotypes for the reference allele are unique. The EHH\(_R/EHH_T\) values for both teromeric and centromeric sides of the target SNP were averaged. We thereafter further calculated the average of the relative EHH (AREHH) for \( \sum_{i=1}^{m} \left( e_{Ai} \right)^2/\sum_{i=1}^{m} e_{Ai} \) values for both teromeric and centromeric sides of the target SNP were averaged. We picked 15 SNPs windows for generating the AREHH values. A definition of the window by fixed physical size (such as 200 kbp) can generate windows with a small number of SNPs because of low SNP density in the DNA chip, which are prone to yield low AREHH values by chance. Although the physical size of the windows can be very large depending on the SNP density in our definition (supplementary fig. S1, Supplementary Material online), its effect is conservative in statistical testing. The windows across the genome were decided without overlap in each population. In the previous genome-wide scans based on EHH-related tests, values of the original statistic (such as unbiased iHS) for each bin of the allele frequency were standardized according to their empirical distribution (Voigh et al. 2006; Sabeti et al. 2007; Tang et al. 2007). However, the standardization consumes time and, in addition, conceals the distribution of values of the original statistic, which disturbs fair comparison in the distribution between real data and simulation results. This can be a problem if the demographic model assumed in the simulation is not adequately imitated to the real situation. In our method, there is no necessity for standardizing the original statistic.

Computer Simulations for the Modified LRH Test

To elucidate the behavior of AREHH, computer simulations were performed. Because the SNPs typed in the DNA chips were chosen according to the allele frequency in populations analyzed in large-scale projects, not in local populations analyzed in this study, it is not easy to reflect such a process in a general coalescent simulation. An important point is that DNA chips are expected to contain SNPs with decreased heterozygosity under selective sweeps in our studied populations but not to include SNPs specific to them. To control such bias, the simulations were divided into 2 phases: a neutral ancestral phase and a selection phase (supplementary fig. S2, Supplementary Material online). The neutral ancestral phase was operated with a coalescent simulation for choosing typed SNPs and creating a founder state of the selection phase. The selection phase was carried out with forward-time simulation. Another strong point of this strategy is that we can extract the results at any point of generations in the forward-time simulation. However, because of the computational load, the forward-time simulation restricts the population size and the number of sites. Therefore, we assumed a small population size, \( N = 1,000 \), and instead a high recombination rate, \( r = 10^{-2} \) (per base pairs per generation), so that \( 4Nr = 4 \times 10^{-4} \).

In the forward-time simulation of the selection phase, we simulated 81 loci (including the selected locus under positive selection at the center) with constant 6-kb intervals.
without assuming new mutation. Here, the locus of the chromosome of the founder generation was denoted by \((i, j)\), not by allelic state, and thus the identical-by-descent state at each generation could be obtained. We examined the strength of selection at \(s = 0.15\) or 0.075 \((2N_s = 300\) or 150\) for codominant selective conditions or at \(s = 0\) for neutral conditions, where \(s\) and \(s/2\) are the selection coefficients for homozygotes and heterozygotes, respectively. In addition to the constant population model \((N = 1,000)\), a population decline \((N = 500)\) model in the selection phase was tested (supplementary fig. S2, Supplementary Material online). Under the selection condition, we assumed that an advantageous mutation generated in a single chromosome increases by positive selection. The simulation results were extracted at those generations where the advantageous allele frequencies become 15%, 25%, 35%, 45%, 55%, 65%, 75%, and 85% (supplementary fig. S2, Supplementary Material online), which are near to 100 and 200 generations in the case of \(s = 0.15\) and 0.075, respectively. Under neutral conditions, the simulation results were extracted at 100 or 200 generations. For each parameter setting, the simulation runs were replicated 500 times.

The coalescent simulation of the neutral ancestral phase was operated with cosi program (Schaffner et al. 2005), which is a modification of Hudson’s ms program. In the simulation, we assumed a 500-kbp region, a constant population size of \(N = 1,000\), and a mutation rate of \(\mu = 1.5 \times 10^{-7}\) (per base pairs per generation) in which we have \(4N\mu = 6 \times 10^{-4}\) and sampled all the chromosomes in the population \((2n = 2N = 2,000)\). To choose typed SNPs, we set 2-kb windows with 4-kb intervals between adjoining windows (total 80 windows). Thereafter, the SNPs with the highest minor allele frequency in every window were chosen. These SNPs were relocated to have constant 6-kbp intervals, which were used as the founder state of the selection phase. The coalescent simulation was then repeated to create 500 founder states.

The results of the selection phase that denoted by \((i, j)\) were connected to the results of the neutral ancestral phase one by one, and the denotations were replaced by allelic state. We calculated EHH_E/EEH_{TF} values for major alleles with the frequency \(\leq 90\%\) as described above. In a few cases that the EHH_{TF} value did not decay below 0.4 at the end SNP (1st or 81st), the EHH_E/EEH_{TF} value was calculated at the end SNP. To compute AREHH, the selected locus was excluded, and 15 SNPs around the selected locus were used. Although our simulation models may lack rigorously to imitate the actual demographic history of populations, they are useful to estimate roughly the behavior of the statistic.

To determine the null distribution of the EHH statistic across the genome under neutrality, we also performed a genome-size neutral simulation as previously reported (Kimura et al. 2007). In brief, a neutral coalescent simulation using cosi program was performed for African, European, and East Asian populations with a flexible recombination rate and a fitting demographic model proposed previously (Schaffner et al. 2005). To correct the ascertainment bias of the selected SNPs on the Affymetrix 500K chips, we extracted the typed SNPs from the simulation data using a rejection method based on the allele frequency spectrum of the simulation and real data.

Computer Simulations for \(R_M\) and \(R_H\) Test

In the same manner as that described in a previous study (Kimura et al. 2007), we simulated the detection powers of \(R_M\) and \(R_H\) to see the effect of the SNP density, sample size, and the initial number of the advantageous alleles. We therefore designed 2 constant-size populations \((N = 1,000)\) that diverged for 200 generations and assumed \(s = 0.15\) \((2N_s = 300)\) for a model of complete selective sweeps. The frequency of the selected allele was set at a single chromosome or 20% when positive selection began to take effect. In addition, we examined a model of partial selective sweeps in which the advantageous mutation reaches an 80% frequency under the positive selection of \(s = 0.085\) \((2N_s = 170)\) for 200 generations.

Results and Discussion

Genetic Differentiation and Admixture between Populations

\(F_{ST}\) values exhibited a genetic differentiation between GDP and another non-African population which was relatively high in comparison to that between any other non-African pairs (supplementary fig. S3, Supplementary Material online). A Neighbor-Joining tree among individuals demonstrated the GDP individuals to have a small diversity within the population (fig. 1A). Taken together, these results are consistent with the fact that this population has been isolated and also possessed a small population size (Ohtsuka 1986). A close relationship between TGN and EAS was inferred from the MDS analyses (supplementary fig. S4, Supplementary Material online) as well as low \(F_{ST}\) values, whereas TGN individuals lay between EAS and GDP individuals in the phylogenetic tree (fig. 1A). Because the population admixture can distort the shape of the phylogenetic tree, we reconstructed another tree removing the TGN individuals. In the reconstructed tree (fig. 1B), the branch length changed but the topology was still retained. This means that indigenous Melanesians have a stronger genetic affinity with Asians than with Africans and European Americans as previously reported (Nei and Roychoudhury 1993; Zhivotovsky et al. 2004).

The results of the STRUCTURE analyses clearly suggested Tongans to originate from an admixture population between Asians and indigenous Melanesians (fig. 1C). When the number of groups assumed \((k)\) was 4 in the STRUCTURE analyses, then individuals in YRI, CEU, EAS, and GDP were assigned to 4 respective groups, which are thought to correspond to classical human races, that is, Negroid, Caucasoid, Mongoloid, and Australoid. These analyses suggested that the Tongan population is genetically derived from Mongoloid at 70.1%, from Australoid at 27.7%, and from the others at 2.2%, which are proportions that are similar to those estimated in some of the previous small-scale studies (Sercel et al. 1985; Martinson 1996). Most recent studies analyzing a large number of
autosomal microsatellites have also showed almost same genetic contributions of Asians and indigenous Melanesians to Polynesians (Friedlaender et al. 2008; Kayser, Lao, et al. 2008). Only a few individuals showed a small genetic contribution from Europeans, thus indicating relatively recent immigration. On the other hand, because the proportion of genetic contribution from Asians and Melanesians in Tongan individuals was homogeneous, it is suggested that the admixture occurred long ago and people have only randomly mated after that. This is also inferred from a tight cluster of TGN individuals in the MDS analysis for the 3 populations (TGN, GDP, and EAS) (supplementary fig. S4, Supplementary Material online). Our results support the Slow train model, obviously ruling out the Express train and Entangled bank models. In addition, the proportions observed in this study were compatible with the sex-biased contribution inferred from previous mtDNA and Y-chromosome data, that is, a nearly 100% Asian origin for maternal lineage and 35% Asian and 65% indigenous Melanesian origins for paternal lineage (Kayser et al. 2006).

Linkage Disequilibrium

The allele frequency spectra after the estimation of the haplotype phase and missing genotype with the fastPHASE algorithm are shown in supplementary figure S5 (Supplementary Material online). We calculated LD coefficients, $D^*$ and $r^2$, in each population using 48 chromosomes (fig. 1D and E). Both of the coefficients were high in GDP and TGN, low in YRI, and intermediate in EAS and CEU, which is thought to reflect their past population sizes. As for TGN, the high LD coefficients can also be attributed partly to the population admixture.

Scans for Selective Sweeps with an LRH Test

To scan for partial selective sweeps in the genome, we employed a modified LRH test. Figure 2 represents the manner that the pattern of EHH$_{R}$/EHH$_{T}$ for SNPs around the selected locus becomes bipolar as the frequency of the advantageous allele increases. This indicates that a hitchhiking allele, which generally has a higher frequency than the selected allele, showed a low EHH$_{R}$/EHH$_{T}$ value and the other allele showed a high EHH$_{R}$/EHH$_{T}$ value. When the frequency of the advantageous allele is still low, the distribution of the EHH$_{R}$/EHH$_{T}$ values is similar to the neutral case, suggesting difficulty of detecting positive selection in such a case. However, after the selected allele becomes the major allele (over 50%), the major allele of neighboring SNPs also showed a very low EHH$_{R}$/EHH$_{T}$ value. Therefore, we can detect such a signature of strong positive selection even without typing the locus under selection using the EHH$_{R}$/EHH$_{T}$ values for the major allele of neighboring SNPs. In this study, we calculated AREHH, that is, the average of the EHH$_{R}$/EHH$_{T}$ values for alleles having a 50–90% frequency over 15 continuous SNPs. Before we applied this method to real data, its performance was examined with a computer simulation. Figure 3 exhibits the results of simulations for estimating the power of our method, which was affected by the frequency of the selected allele (fig. 3A). Under neutrality, the first percentile of the value of AREHH was 0.819. When the threshold was set at this value, then the advantageous mutation ($2N_s = 300$) that reached frequencies of 55%, 65%, 75%, and 85% was detected at a probability of 72.8%, 87.2%, 92.8%, and 86.2%, respectively. Because the number of sampled chromosomes hardly altered the detection power (fig. 3B), we thought that the 24 individuals sampled in this study were therefore adequate. The strength of selection ($2N_s = 300$ or 150) had a substantial effect on the detection power (fig. 3C), which may reflect the opportunity for recombination events that depend on the time needed to reach the examined frequency. In addition, a population decline caused a decrease in the detection power and an increase in the false positive rate at a certain threshold (fig. 3C), thus suggesting the limits of an approach based on a comparison between alleles.

The real distributions of AREHH across genomic windows were considerably different among the populations (fig. 3D). To determine the null distribution of this statistic under the neutrality, we also carried out a genome-size coalescent simulation for East Asian, European, and African populations according to a validated demographic model reported previously (Schaffner et al. 2005) (fig. 3D). The results of the simulation demonstrated the demographic history of populations to affect the AREHH values, which can partly explain the interpopulation difference in the distribution. However, the real and null distributions were substantially deviated in CEU and EAS, whereas those matched well in YRI. This deviation may tell us how often humans have been exposed to selective pressures in the out-of-Africa duration. Alternatively, because the deviation is too large, it may be attributed to population substructures and more strong bottlenecks that were not considered in the simulation model used.

To identify candidates, we set the threshold at the second percentile of the empirical distribution (AREHH below 0.484 in TGN, 0.542 in GDP, 0.512 in EAS, 0.579 in CEU, and 0.842 in YRI) although these thresholds may hold only a low power to detect positive selection especially in GDP. Nonetheless, our modified method could detect the selected genes that have been previously reported such as LCT in CEU and ALDH1A2 in EAS (Bersaglieri et al. 2004; Oota et al. 2004). The AREHH values are plotted on their chromosomal positions in supplementary figures S6 and S7 (Supplementary Material online). We also exhibit the rank of AREHH values in supplementary data S1–S5 (Supplementary Material online).

Scans for Population-Specific Selective Sweeps

The approach based on interallelic comparison in EHH is not applicable to scanning for loci fixed already, and it has only a low power when the population has experienced severe bottlenecks as described above. Although the composite likelihood test that is based on the allele frequency spectrum can detect complete selective sweeps (Kim and Stephan 2002; Nielsen et al. 2005; Williamson et al.
Similarly to the aforementioned approach, this test captures loci under positive selection even if it has operated in the common ancestral population. However, we are now most interested in the selective sweeps occurring locally in Oceanic populations. To detect population-specific selective sweeps, therefore, we calculated the interpopulation ratio of haplotype homozygosity, $R_H$, and the interpopulation ratio of homozygosity for the test population’s most frequent haplotype, $R_M$ (Kimura et al. 2007). The $R_H$ value can be an indicator of nucleotide diversity and past recombination events, whereas the $R_M$ value can be an indicator of genetic differentiation like $F_{ST}$. Previous reports (Sabeti et al. 2007; Tang et al. 2007) have proposed similar approaches based on comparison between populations, which require calculation of the interpopulation ratio of EHH values for every allele. In the $R_H$ and $R_M$ test, we can avoid redundant tests for neighboring SNPs in strong LD with each other. In addition, the $R_M$ value measuring haplotypic
FIG. 3.—The results of computer simulations for estimating the power of statistics. (A–D) Cumulative distributions of AREHH. Various values of the selected allele frequency ($P = 55\%$, 65\%, 75\%, or 85\%) (A), the chromosome sample size ($2n = 120$ or 48) (B), the strength of selection ($2Ns = 300$ or 150) and the population size (constant or decline) (C) were simulated to examine the detection power. The 15 SNPs window for the AREHH calculation spanned but excluded the selected locus. In the panels (B) and (C), the cases of $P = 75\%$ are shown for the selection condition. In the default setting, the population size ($N$) is 1,000, the number of chromosomes sampled ($2n$) is 120, and the strength of selection ($2Ns$) is 300 in the selection (Sel) condition or 0 in the neutral (Neu) condition. Under the population decline models, $N$ is 500 in the selection phase. The number of generations after the population decline ($G$) is dependent on the condition of the simulations. Cumulative distributions for the real data and the data from the genome-size neutral simulation using a demographic model are also shown (D). (E–H) Estimation of the detection powers of $R_M$ and $R_H$. Cumulative distributions of $R_M$ (E) and $R_H$ (F) in method 1 (block definition: $H_M \geq 0.9$) and those (G and H) in method 2 (block definition: $H_M \geq 0.5$) are shown. Different initial frequencies (IFs) of the selected allele at the start of the selection phase (single chromosome or 20%) and different final frequencies of the selected mutation (fixed or 80%) were tested using methods 1 and 2, respectively. Different values of the SNP density (2 or 6 kb/SNP) and chromosome sample size ($2n = 120$ or 48) were also simulated.
differentiation enables us to capture the differentiation of untyped polymorphisms more powerfully than the $F_{ST}$ value for each SNP (Kimura et al. 2007).

The block definition of $H_M \geq 0.9$, which we call method 1 here, is appropriate to detect complete selective sweeps in which advantageous alleles have reached (near) fixation. When we performed a simulation assuming 2 diverged populations with a constant size and 6 kb of SNP intervals that is the similar density as mounted on the Affymetrix 500K chips, thresholds of $R_M < 0.05$ and $R_H < 0.3$ realized approximately 80% power (fig. 3E and F). The block definition of $H_M \geq 0.5$, or method 2, is potentialized to detect the alleles under selection that have reached a frequency of over approximately 70%. Figure 3G and H shows the detection power for the cases in which a single advantageous mutation increased to 80% frequency, which is lower than the power for a complete selective sweep. For a selected allele with 80% frequency, the thresholds of $R_M < 0.1$ and $R_H < 0.5$ had approximately an 80% power. As previously reported (Kimura et al. 2007), the distributions of $R_M$ and $R_H$ shift depend on the demographic history of the populations. Especially, it should be noted that decline of the test population’s size results in downshift of the distributions in both cases of selection and neutrality. Therefore, if we test a population that has experienced a decline in size, then our simulations assuming a model with constant-size populations are thought to give conservative estimation of the detection power.

Our first interest is to elucidate whether there are any mutations that were generated after the divergence from Asians and then reached fixation in Polynesians. For this purpose, we applied method 1 ($H_M \geq 0.9$) to the test for TGN using EAS as the reference population (TGN vs. EAS). As a result, we did not observe any block satisfying the thresholds of $R_M < 0.05$ and $R_H < 0.3$ (table 1). Taking into account the powerfulness of these thresholds (fig. 3E and F), the result suggests that there was no (or few, if any) mutation newly generated and fixed in Polynesians. Because the dispersal of Austronesian-speaking people is thought to be dated at 6 KYA at the most, then the divergence time would be too short for a new mutation on autosomes to reach fixation in Tongans. Although the near fixation of an Austronesian-specific type of mtDNA has previously been observed in Polynesians (Redd et al. 1995), this would be due to the small population size of mtDNA that is one-fourth of the autosomal population size. As for Polynesia-specific complete selective sweeps, it remains an alternative possibility that old-standing alleles originated from Asians and/or from indigenous Melanesians have been fixed by positive selection in Polynesians. A looser threshold of the statistics may detect such loci, yet only a low power can be expected if the frequency of the selected allele was relatively high at the time when the selective pressure started to operate (fig. 3E and F). When we chose the threshold of $R_M < 0.25$ in TGN versus EAS, most of such blocks showed high $R_M$ values in TGN versus GDP instead ( Supplementary Material online). This indicates that when a haplotype with a low frequency in Asians reached either complete or near fixation in Polynesians, then the same haplotype from indigenous Melanesians may have also contributed to the fixation in most cases. Although these loci may be potential candidates for complete selective sweeps on standing alleles, careful interpretation is needed because most of these may be false positives generated by genetic drift. When we applied method 2 ($H_M \geq 0.5$) to TGN using the threshold of $R_M < 0.1$ and $R_H < 0.5$, numerous blocks were detected as candidates for loci where a single mutation gained a high frequency but did not reach fixation (table 1). As the reference population (EAS, CEU, or YRI) becomes genetically closer to TGN, the number of blocks with $R_M < 0.1$ and $R_H < 0.5$ becomes smaller (table 1). Giving an attention to the overlap of the results of TGN versus EAS, TGN versus CEU, and TGN versus YRI, we could then further narrow the candidates down to 54 regions (0.11% of the total blocks) ( Supplementary Material online).

In a scan for complete selective sweeps on GDP using method 1, blocks showing very low $R_M$ and $R_H$ values were abundant even when EAS was used as the reference population (table 1). This downshift in the distribution is thought to reflect the small population size as well as the long divergence time from the other populations (Kimura et al. 2007). The number of the blocks that satisfy $R_M < 0.05$ and $R_H < 0.3$ against the 3 reference populations (EAS, CEU, and YRI) was 81 regions across the genome (0.22% of the total blocks) (table 1 and supplementary data S8 [Supplementary Material online]). In method 2, the blocks with $R_M < 0.1$ and $R_H < 0.5$ were so many that we could not effectively reduce the candidates (1,738 regions) (table 1). Because these thresholds are thought to be too conservative for testing a small-size population (Kimura et al. 2007), we would be able to use more strict thresholds while keeping a high detection power. Therefore, we used $R_M < 0.01$ and $R_H < 0.25$, which correspond to approximately 50% power in our simulation.

<table>
<thead>
<tr>
<th>Method 1 ($H_M \geq 0.9$)</th>
<th>Total Blocks</th>
<th>versus EAS</th>
<th>versus CEU</th>
<th>versus YRI</th>
<th>All Filters</th>
</tr>
</thead>
<tbody>
<tr>
<td>TGN 30,822</td>
<td>$R_M &lt; 0.05$</td>
<td>$R_H &lt; 0.3$</td>
<td>$R_M &lt; 0.05$</td>
<td>$R_H &lt; 0.3$</td>
<td>$R_M &lt; 0.05$</td>
</tr>
<tr>
<td>GDP 37,010</td>
<td>0</td>
<td>7</td>
<td>73</td>
<td>291</td>
<td>952</td>
</tr>
<tr>
<td>Method 2 ($H_M \geq 0.5$)</td>
<td>$R_M &lt; 0.1$</td>
<td>$R_H &lt; 0.5$</td>
<td>$R_M &lt; 0.1$</td>
<td>$R_H &lt; 0.5$</td>
<td>$R_M &lt; 0.1$</td>
</tr>
<tr>
<td>TGN 48,037</td>
<td>286</td>
<td>2,075</td>
<td>3,141</td>
<td>7,294</td>
<td>8,592</td>
</tr>
<tr>
<td>GDP 42,510</td>
<td>5,181</td>
<td>8,640</td>
<td>6,693</td>
<td>11,761</td>
<td>11,362</td>
</tr>
<tr>
<td>Method 2 ($H_M \geq 0.5$)</td>
<td>$R_M &lt; 0.01$</td>
<td>$R_H &lt; 0.25$</td>
<td>$R_M &lt; 0.01$</td>
<td>$R_H &lt; 0.25$</td>
<td>$R_M &lt; 0.01$</td>
</tr>
<tr>
<td>GDP 42,510</td>
<td>1,146</td>
<td>1,663</td>
<td>1,712</td>
<td>2,494</td>
<td>4,342</td>
</tr>
</tbody>
</table>
assuming constant-size populations (fig. 3G and H). These conditions could reduce the candidates down to 202 regions (0.48% of the total blocks) (table 1 and supplementary data S9 [Supplementary Material online]).

Candidate Regions under Selective Sweeps

The methods used to scan for selective sweeps in this study have their own characteristics. The test using AREHH is potentialized to detect selective sweeps where the selected allele has gained a greater than 50% frequency, but it has not yet reached fixation. This test can detect selective sweeps occurring in the common ancestry of different populations as well as in a local population. In the method 1 of the $R_M$ and $R_H$ test ($H_M \geq 0.9$), the thresholds of $R_M < 0.05$ and $R_H < 0.3$ detect only loci fixed or nearly fixed by population-specific positive selection. If a looser threshold such as $R_M < 0.25$ is used in the same test, we may identify positive selections that have acted on old-standing alleles, but only a low detection power and high false positive rate can be expected. Method 2 of the $R_M$ and $R_H$ test ($H_M \geq 0.5$) is applicable to a scan for those regions where the locally selected allele reached over approximately 70% frequency including fixation. The chromosomal positions of the candidate regions detected by the respective methods are exhibited in supplementary figure S8 (Supplementary Material online). In some regions, the signatures detected by different methods overlapped. Such regions are considered to have a higher possibility to be true positives. Other regions show the signature unique to one method, which may be attributed to the uniqueness of the characteristics of the methods or to type I and type II errors.

Our scans suggested no private mutation to exist on the Tongan autosomes that had reached fixation. However, there remain alternative possibilities that old-standing alleles have reached fixation by local selective pressures and that newly generated advantageous mutations have gained a high frequency but have not yet reached fixation. The block showing the lowest $R_M$ value (0.076) in the test of TGN versus EAS using method 1 was located at 92788024–92838919 on chromosome 12 (supplementary data S6, Supplementary Material online), which is at 41-kb distance from the CRADD gene (fig. 4). It is worth noting that an approximately 500-kb deletion around this gene in mouse has been reported to cause a “high growth” mutant that shows a proportional increase in tissue and organ size without obesity (Horvat and Medrano 1998). Another candidate for the selected region in which an old-standing allele reached fixation was VLDLR (supplementary data S6, Supplementary Material online), which is involved in triglyceride and fatty acid metabolism (Tacken et al. 2001). In addition, overlapping signatures in both methods 1 and 2 (supplementary data S6 and S7, Supplementary Material online) were observed in the gene region of EXT2, which is a causal gene of the type II form of multiple exostoses, and it plays a crucial role in bone formation (Stickens and Evans 1997). These genes can be candidates that are associated with the large fat, muscle, and bone masses of Polynesians. A recent paper examining the interpopulation differentiation of the type II diabetes–associated genes has suggested that a susceptible allele of PPARGC1A may play a role in the large difference in the prevalence of the disease between Polynesians and neighboring populations (Myles et al. 2007). However, our scans did not identify any signature of positive selection on the gene region of PPARGC1A.

One of the strongest signatures of selective sweeps in GDP was located at the region including the LHX4 and ACBD6 genes on chromosome 1 (supplementary data S8, Supplementary Material online). LHX4 encodes a transcriptional regulator involved in the control of the development of the pituitary gland, and mutations in this gene are associated with syndromic short stature and pituitary defects (Machinis et al. 2001; Castro-Feijoo et al. 2005). ACBD6 is a binding protein of acyl-coenzyme A that has a role in fatty acid metabolism (Entrez Gene). The gene region of IGF1R, which is the receptor for insulin-like growth factor 1 (Castro-Feijoo et al. 2005), was also considered to be a candidate of the selected genes (supplementary data S8, Supplementary Material online). These genes may be involved in the slow growth, short stature, and lightweight characteristics of New Guineans. Future association studies between genotypes and phenotypes are indispensable.
Other candidates of selective sweeps in Oceanic populations included several interesting genes such as DDX58, SIAT4A (supplementary data S7, Supplementary Material online), and IVNS1ABP (supplementary data S8, Supplementary Material online), which code molecules related with infection of the influenza A viruses (Wolff et al. 1998; Shinya et al. 2006; Mibayashi et al. 2007; Nicholls et al. 2007). If we could identify a protective effect of the selected allele against the influenza, these kinds of signatures may therefore suggest evidence for the epidemic history of the virus in Oceania and human conquest of the disease by genetic adaptation.

We observed the candidates of selective sweeps that include no gene or genes whose functions have not been known yet. The selected loci should have some phenotypic functions because natural selection acts on phenotypes. Therefore, the scans for signatures of selective sweeps can be a trigger to identify genes or DNA sequences with some important function as well as to determine the functional difference between alleles. Such an approach based on evolutionary genetics, which thus provide clues to understand how humans have adapted to our environments, are therefore also expected to help elucidate the genomic functions if further functional and association studies on the candidates are carried out. Conversely, this approach may also shed some light on the invisible phenotypic difference between populations. Our study demonstrated that genome-wide SNP typing systems, which have exerted their power for identifying disease-associated polymorphisms (The Wellcome Trust Case Control Consortium 2007), are also useful for evolutionary study on human populations.

Supplementary Material

Supplementary table S1, figures S1–S8, and data S1–S9 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We are grateful to the Gidra people in Papua New Guinea and the Tongan people for their kind cooperation in providing blood samples. We thank the staff of the Department of Health, Western Province of Papua New Guinea, Dr Tetsuro Hongo at Yamanashi Institute of Environmental Sciences, Dr Taniela Palu at Ministry of Health, Kingdom of Tonga, Dr Viliami Tangi at Diabetes Clinic, Kingdom of Tonga, Dr Kazumichi Katayama at Kyoto University for help in sample collection, and 2 anonymous reviewers for helpful comments. This study was partly supported by a Grant-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science, and Technology of Japan. This research was done mainly at the Department of Human Genetics, Graduate School of Medicine, The University of Tokyo.

Literature Cited


The Wellcome Trust Case Control Consortium. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 447:661–678.


Yoko Satta, Associate Editor

Accepted May 26, 2008