Two-dimensional site frequency spectrum for detecting, classifying and dating incomplete selective sweeps

Yoko Satta1*, Wanjing Zheng1, Kumiko V. Nishiyama1, Risa L. Iwasaki1, Toshiyuki Hayakawa2,3, Naoko T. Fujito4 and Naoyuki Takahata1

1School of Advanced Sciences, SOKENDAI (The Graduate University for Advanced Studies), Hayama, Kanagawa 240-0193, Japan
2Graduate School of Systems Life Sciences and 3Faculty of Arts and Science, Kyushu University, Fukuoka 819-0395, Japan
4Institute for Human Genetics and Department of Epidemiology and Biostatistics, University of California, San Francisco, CA 94143, USA

(Received 25 April 2019, accepted 8 September 2019; J-STAGE Advance published date: 11 December 2019)

The two-dimensional site frequency spectrum (2D SFS) was investigated to describe the intra-allelic variability (IAV) maintained within a derived allele (D) group that has undergone an incomplete selective sweep against an ancestral allele group. We observed that recombination certainly muddles the ancestral relationships of allelic lineages between the two allele groups; however, the 2D SFS reveals intriguing signatures of recombination as well as the genealogical structure of the D group, particularly the size of a mutation and the time to the most recent common ancestor (TMRCA). Coalescent simulations were performed to achieve powerful and robust 2D SFS-based statistics with special reference to accurate evaluation of IAV, significance of recombination effects, and distinction between hard and soft selective sweeps. These studies were extended to a case wherein an incomplete selective sweep is no longer in progress and ceased in the recent past. The 2D SFS-based method was applied to 100 intronic linkage disequilibrium regions randomly chosen from the East Asian population of modern humans to examine the P value distributions of the summary statistics under the null hypothesis of neutrality in a nonequilibrium demographic model. We argue that about 96% of intronic variants are non-adaptive with a 10% false discovery rate. Furthermore, this method was applied to six genomic regions in Eurasian populations that were claimed to have experienced recent selective sweeps. We found that two of these genomic regions did not have significant signals of selective sweeps, but the remaining four had undergone hard and soft sweeps and were dated, in terms of TMRCA, after the major out-of-Africa dispersal of modern humans.

Key words: intra-allelic variability, linkage disequilibrium, recombination, SNP, soft and hard sweeps

INTRODUCTION

Based on the site frequency spectrum (SFS) of single nucleotide polymorphisms (SNPs) in the linkage disequilibrium (LD) region of a focal (core) site under positive Darwinian selection, we previously proposed an inference method that represents SFS and LD information as hierarchical barcodes and measures sweep signals by a newly defined statistic, \( F_c \) (Fujito et al., 2018a). If a selective sweep is incomplete (Sabeti et al., 2002; Pickrell et al., 2009; Peter et al., 2012; Ferrer-Admetlla et al., 2014; Field et al., 2016), derived and ancestral alleles coexist and segregate at a core site in the sample, and these core alleles define two allele (or haplotype) groups in the extended LD region, similar to that in the long-range haplotype (LRH) test (Vitti et al., 2013; Fan et al., 2016). Polymorphism within a derived allele (D) group under positive selection was termed intra-allelic variability (IAV) by Slatkin and Rannala (1997, 2000), and \( F_c \) evaluates it as the proportion of derived alleles in the D group relative to those in the entire sample. It was shown that \( F_c \) is robust not only to the temporally changing demography by measuring relative IAV, but also to recombination by excluding high-frequency derived alleles. However,
while the procedure of exclusion is simple, it is entirely heuristic. Thus, it is desirable to further develop relevant theories and methods. More importantly, although recombination certainly obscures the allelic genealogy in a sample, the outcome is intimately related to the genealogical structure of IAV, especially the family size or the size of a mutation that occurs in a branch with a specified number of descendant lineages (Fu, 1995).

A selective sweep can be either hard (Maynard Smith and Haigh, 1974; Kaplan et al., 1989; Stephan et al., 1992; Braverman et al., 1995) or soft (Orr and Betancourt, 2001; Innan and Kim, 2004; Herisson and Pennings, 2005; Przeworski et al., 2005). A hard sweep is driven by a single beneficial haplotype, whereas a soft sweep is driven by multiple beneficial haplotypes that may form because of recurrent mutation/migration or as standing variations (Pennings and Herisson, 2006; Teshima et al., 2006; Akey, 2009; Scheinfeldt et al., 2009; Peter et al., 2012). Fujito et al. (2018a) considered a soft sweep for generality, but they could not develop a method that can distinguish between the two modes of adaptation. Using 11 summary statistics of selection tests such as those in Tajima (1989), Fay and Wu (2000) and Sabeti et al. (2002), Pybus et al. (2015) developed a machine learning framework to detect and classify hard sweeps in human populations. Also, using a machine learning method trained for the estimated population mutation rate (Watterson, 1975; Nei and Li, 1979), haplotype homozygosity (Garud et al., 2015; Harris et al., 2018, for unphased genotype data) and LD (Kelly, 1997), Schrider and Kern (2017) found that soft sweeps prevail in recent human adaptation (Ronen et al., 2013; Sheehan and Song, 2016). Schrider and Kern’s method could infer the mode of selection with robustness to nonequilibrium demography and was applied to the case of complete selective sweeps. However, to identify incomplete selective sweeps, different summaries and/or theoretical approaches are feasible or even preferable, as discussed by Grossman et al. (2010) and many others (Vitti et al., 2013; Ferrer-Admetlla et al., 2014; Ronen et al., 2015; Fan et al., 2016; Schrider and Kern, 2016; Akbari et al., 2018).

Incomplete selective sweeps may provide an unprecedented opportunity to distinguish effects of positive selection from those of demography. Among many demographic causes, allelic “surfing” (Currat et al., 2006; Klopfeinstein et al., 2006) argues against positive selection invoked for the gene regions of ASPM and MCPH1 that are associated with human brain development (Kouprina et al., 2004; Evans et al., 2005; Mekel-Bobrov et al., 2005). Surfing is a phenomenon caused by a few genetically homogeneous individuals on a wave front in a geographically expanding population, who can produce disproportionate numbers of descendants by chance. Nielsen et al. (2007) doubted if it would be possible to develop methods that can distinguish between allelic surfing and positive selection with high accuracy. However, unlike positive selection, surfing should not distinguish between derived and ancestral alleles at a core site. Moreover, in the case of modern humans, surfing does not appear to be an ongoing phenomenon; rather, it is very likely associated with the past range expansions that might occur after the out-of-Africa dispersal, as presumed in Currat et al. (2006). In this case, it is interesting to determine if any incomplete selective sweeps are really ongoing. These two aspects of the target allele and timing of positive selection, at least in principle, allow the possibility of validating or invalidating the hypothesis of allelic surfing.

Apart from the above-mentioned question, a time period during which positive selection exerted or terminated its action has gained considerable attention in particular relation to human prehistory. For instance, using a rejection sampling method, Przeworski (2003) estimated the time since the fixation of a beneficial allele at a locus such as FOXP2 (Enard et al., 2002; Coop et al., 2008; Ayub et al., 2013). Based on the draft sequence of the Neanderthal genome, Green et al. (2010) identified a number of nonsynonymous substitutions that have been fixed in ancestral modern humans. Recently, using a cross-population composite likelihood ratio, Racimo (2016) inferred complete selective sweeps that occurred after the split of modern humans from Neanderthals, but before the split of Africans and Eurasians. For incomplete selective sweeps, Itan et al. (2009) and Peter et al. (2012) provided age estimates of selected alleles in an approximate Bayesian computation (ABC) framework (Chen et al., 2015, from haplotype lengths by a hidden Markov model). Furthermore, based on the comparison of SFS in multiple population genomes, Tournebize et al. (2018) recently developed a method to detect and date past and recent hard sweeps. However, to the best of our knowledge, there are few studies about the cessation of its action during the intermediate process of a selective sweep and the time elapsed since then (Garud et al., 2015; Hunter-Zinck and Clark, 2015; Pybus et al., 2015; Ferretti et al., 2017).

In this paper, we investigate a two-dimensional SFS (2D SFS) to describe IAV under an incomplete selective sweep. By coalescent simulations (Hudson, 2002; Kern and Schrider, 2016), we show how mutation, recombination, genetic drift and two modes of selective sweeps form the pattern and level of a 2D SFS. We focus on the family size and the time to the most recent common ancestor (TMRCA) of the derived alleles within a D group, as the genealogy is clearly not a star-shaped one, even under a strong selective sweep (Ferretti et al., 2017; Yang et al., 2018, and references therein). Similarly, we examine several classical and newly defined summary statistics that may facilitate the identification of the genealogical structure of IAV, evaluate their statistical performance, and apply some of them to a random sample of genomic
regions from the database (1000GD) of the 1000 Genomes Project Consortium (2015) as well as six genomic regions that are claimed to have undergone incomplete selective sweeps. Furthermore, we study the persistence time of the detectable signatures of an incomplete selective sweep that ceased in the past.

THEORIES AND METHODS

2D SFS  The 2D SFS is formulated to extract information about an incomplete selective sweep that is inscribed as biallelic SNPs in a genomic region sampled from a single panmictic diploid population. Two allele groups, the D group of size \( m \) \((1 \leq m < n)\) and the ancestral allele (A) group of size \( n-m \) where \( n \) is the entire sample size, are defined by the derived/ancestral state of a predetermined focal site or a core site, with the derived allele frequency being \( f_i = m/n \). These groups are extended to an LD region or a core region surrounding it (Fujito et al., 2018a). The 2D SFS for linked neutral polymorphism in the core region is denoted by the matrix \( \phi_{i,j} \), where the \((i,j)\) element represents the number of segregating sites at which \( i \) \((0 \leq i \leq m)\) derived alleles are found in the D group and \( j \) \((0 \leq j \leq n-m)\) derived alleles are found in the A group. The total number of derived alleles is \( 1 \leq k = i+j < n \). We are concerned only with segregating sites in a sample and ignore \( \phi_{0,0} \) and \( \phi_{m,n-m} \) or set them to 0. \( \phi_{0,0} \) denotes the number of segregating sites at which \( i \) derived alleles can be found only in the D group. For \( 1 \leq i < m \) and \( 1 \leq j \leq n-m \), the presence of \( \phi_{i,j} > 0 \) implies that, under the assumption of the infinite-sites model (Kimura, 1969), these segregating sites have experienced recombination. Note that when \( \phi_{n-m,m} > 0 \), there are only three different types of gametes specified by the core site and a linked site, despite the fact that recombination between the sites is required to produce them. In other words, when we distinguish between the derived and ancestral alleles, the presence of four different types of gametes is not a necessary condition for the involvement of recombination.

The SFS can capture rich information about the SNP sites (Sawyer and Hartl, 1992; Bustamante et al., 2001). The entire distribution was used in the composite likelihood methods (Kim and Stephan, 2002; Nielsen et al., 2005; Williamson et al., 2007) as well as in inference methods about population demographic histories (Liu and Fu, 2015; Lapierre et al., 2017; Terhorst et al., 2017). However, the SFS is usually very sparse. As a 2D SFS is even more sparse, it is sensible and convenient to consider compressed versions simultaneously:

\[
\xi_k = \sum_{i=0}^{k} \phi_{i,k-i}, \quad \text{for} \quad 1 \leq k < n, \quad (1a)
\]

\[
\zeta_i = \sum_{j=0}^{m} \phi_{i,j} = \phi_{0,0} + \xi_i, \quad \text{for} \quad 1 \leq i < m, \quad (1b)
\]

where \( \xi_k \) and \( \zeta_i \), with \( \phi_{0,0} \) and \( \zeta_i = \sum_{j=1}^{m} \phi_{i,j} \), denote the one-dimensional (1D) SFS for the entire sample and the D group subsample, respectively.

In addition to the entire SFS, we consider various summary statistics, which include \( D \) by Tajima (1989), \( H \) by Fay and Wu (2000), \( E \) by Zeng et al. (2006), \( K \) instead of the original symbol \( L \) by Ferretti et al. (2017), and \( F_c \) by Fujito et al. (2018a, 2018b). However, if \( D, H, E \) and \( K \) are applied to the entire sample, the statistical power may be weakened in the present case of incomplete or partial selective sweeps. We therefore apply them to the D group subsample and re-evaluate their statistical behavior in this circumstance. The pattern and extent of polymorphism within the D group are the key features to be estimated for detecting incomplete selective sweeps. One such feature is captured by the \( F_c \) statistic, which can be rewritten in terms of \( \phi_{i,j} \) as

\[
F_c = \sum_{i,j} (i+j) \phi_{i,j} / \sum_{i,j} (i+j) \phi_{i,j} \cdot \frac{1}{2} = \sum_{i=0}^{m} (i+1) \phi_{i,0} / \sum_{j=0}^{m} (i+1) \phi_{i,0}.
\]

Each summation is taken over all values of \( i \) and \( j \) in \( 0 \leq i \leq k_m, 0 \leq j \leq \min(n-m, k_m) \) and \( 1 \leq i + j \leq k_m \), where \( k_m \) is the largest number of \( k = i + j \) derived alleles in the frequency class \( c_i = 1 \). Here, \( c_i \) at a core site with \( m \) derived alleles \((k_m < m)\) is determined under the standard neutral model of constant size, hereafter referred to as the standard model (Fujito et al., 2018a). In a sample of size \( 503 < n < 1,367 \), \( k_m \) takes one of 1, 3, 9, 25, 68, 185 and 503 in the seven possible frequency classes, with each having nearly equal numbers of segregating sites, and we set \( c_i = 1 \) if \( m = 1 \), \( c_i = 2 \) if \( 1 < m \leq 3 \), etc. For the promoter region of STS81A2 with \( n = 1,008 \) and \( m = 349 \) in the East Asian (EAS) population (1000GD), we set \( c_i = 7 \) from 185 \( < m \leq 503 \); thus, \( k_m = 185 \) in class 6. Using \( k_m \), we can rewrite the denominator in Formula (2) as \( \sum_{j=0}^{m} k_j \).

The two vector components \( \phi_{0,0} \) and \( \zeta_i \) in Formula (1b) are most informative and convenient to describe IAV. We denote the maximum value of \( i \) by \( i_{\text{max}} \), which satisfies \( \phi_{0,0} > 0 \). \( \phi_{0,0} > 0 \) indicates two possibilities. One for a wide range of \( i \) with \( \phi_{0,0} > 0 \) (or a large \( i_{\text{max}} \)) results from a soft sweep, wherein more than one distinct derived haplotype with respect to a core site has been subjected to positive selection since it began to act. As recombination is unlikely to generate \( \phi_{0,0} > 0 \) under the circumstances discussed in this study (see below), this would provide strong evidence for soft sweeps, regardless of the \( \zeta_i \) values. However, \( \phi_{0,0} > 0 \) for some \( i \) values may occasionally be created by the cumulative effects of recombination, genetic drift and insufficient sampling processes (Fig. 1). It is then necessary to examine if \( \zeta_i > 0 \) holds, because it is an unambiguous sign of recombination with respect to the core site, as discussed above and in the four-gamete test (Hudson and Kaplan, 1985), and cannot be produced directly by any form of selective sweep. The
positive only for $\phi$-based estimate of the maximum family size with small values of $i$, as the can occur for mutations that originate within the D group and is expected to be small, we consider the larger one, denoted by $\phi_{imax}$ or out-
case for $m/2$. The former recombination are much smaller than generation, respectively. It is possible for $\phi$-rationation size and the recombination rate per core region per generation, respectively. To have such inter-
components, we divide the total number of segregating alleles in the entire sample ($\phi$-max) into two components, $S_0$ and $S_1$, as follows:

$$S_i = \sum_{i=0}^{m-1} \phi_i = \sum_{i=1}^{m-1} \phi_{0i} + \sum_{i=1}^{m-1} \phi_i, S_{01} = S_0 + S_1.$$  \hspace{1cm} (3a)

Akey et al. (2004) used $S_0$ to estimate the time elapsed since a selective sweep assuming a star-shaped genealogy, and Satta et al. (2018) studied the nonequilibrium distribution of $\phi$-normalized in individual frequency classes $c_i$ (1 ≤ $c_i$ ≤ 8, depending on the derived core allele frequency in the 1000GD). However, no-one has studied $\phi_0$ and $\phi_i$ or $S_0$ and $S_1$, separately. We also divide the total number of derived alleles ($L_0$) within the D group into two components, $L_{01}$ and $L_{11}$, as follows:

$$L_i = \sum_{i=0}^{m-1} \phi_{imax} = \sum_{i=1}^{m-1} \phi_{0i} + \sum_{i=1}^{m-1} \phi_i, L_{01} = L_0 + L_1.$$ \hspace{1cm} (3b)

where $L_i$ is the sum of the scaled SFS (Fu, 1995) of $\phi_i$, $S_i$ and $L_i$ within the D group and is expected to be small under strong selective sweeps. However, it is desirable to exclude the possibility of either background selection (Charlesworth and Charlesworth, 2012 and references therein) or an intrinsically low population mutation rate in a core region ($\rho = 4N_0r$, where $u$ is the neutral mutation rate per region per generation). To have such internal control, we may also compute the number of derived alleles in the entire sample ($L_i = \sum_{i=1}^{m-1} k\phi_i$) and use it to

rationale behind this is as follows. Suppose that a strong selective sweep is ongoing. Under this ideal situation, the genealogy among the derived alleles in the D group would be less structured with shortened basal branches, and their family size manifested by mutations would remain small in most cases. Recombination between the D and A groups can therefore make $\phi_i$, positive only for restricted $i$ values in $1 \leq i \leq i_{max}$ and $m - i_{max} \leq i < m$, provided that these maximum family sizes ($i_{max}$) revealed by recombination are much smaller than $m/2$. The former case for $1 \leq i \leq i_{max}$ can occur for mutations that originate from the A group, but are subsequently recombined with the derived core alleles. On the other hand, the latter case for $m - i_{max} \leq i < m$ can occur for mutations that are shared by all the lineages in the D group (Fig. 1). In either case, the recombining unit size must be identical to the number of descendant lineages of that recombining branch in the D group. Conversely, if a reciprocal relationship, such as $\phi_i = \phi_{m-i}$, is observed, it can be considered as a manifestation of the family size within the D group. Unfortunately, this exact relationship can be observed only when the population recombination rate of $\rho = 4N_0r$ is limited ($N_i$ and $r$ refer to the effective population size and the recombination rate per core region per generation, respectively). It is possible for $\phi_{m-i} > 0$ to be created even if $i$ individual recombination events involve only singleton branches. As this is very unlikely for $\phi_i$ with small values of $i$, $\phi_{m-i}$ declines more slowly than $\phi_i$ as $i$ increases. Therefore, $\phi_i$ with small values of $i$ reflects the genealogical structure of IAV more accurately than the reciprocal $\phi_{m-i}$. In addition, the absolute value of $\phi_i$ may differ greatly from that of $\phi_{m-i}$ because the relevant mutations involved are asymmetric in number, even within the same A group. In fact, $\phi_{m-i} > 0$ is produced only by mutations in the stem lineage of the D group, whereas $\phi_i > 0$ can be produced by a much wider range of mutations that have accumulated in the A group.

Overall, in the absence of recombination events, we have $\phi_i = 0$ exactly for $1 \leq i < m$, irrespective of the mode of selective sweep. Under a hard sweep, $\phi_0 > 0$ is concentrated in a narrow range of $1 \leq i \leq i_{max}$ whereas under a soft sweep, $\phi_0 > 0$ spreads over a wide range of $i$; however, there is a gray zone in between. In the presence of recombination, $\phi_i > 0$ under a hard sweep is generally concentrated in $1 \leq i \leq i_{max}$ and $m - i_{max} \leq i < m$ because of the small family size within the D group. When $i_{max}$ differs from $i_{max}$, we consider the larger one, denoted by $i_{max}$, as the $\phi_i$-based estimate of the maximum family size within IAV. The size of $i_{max}$ and $i_{max}$ under a soft sweep may be much larger than those under a hard sweep, and recombination augments this trend.

Based on the different implications of $\phi_0$ and $\phi_i$, outlined above, we divide the total number of segregating sites ($S_i$) within the D group into two components, $S_0$ and $S_1$, as follows:

$$S_i = \sum_{i=1}^{m-1} \phi_i = \sum_{i=1}^{m-1} \phi_{0i} + \sum_{i=1}^{m-1} \phi_i, S_{01} = S_0 + S_1.$$  \hspace{1cm} (3a)

Akey et al. (2004) used $S_0$ to estimate the time elapsed since a selective sweep assuming a star-shaped genealogy, and Satta et al. (2018) studied the nonequilibrium distribution of $\phi_i$ normalized in individual frequency classes $c_i$ (1 ≤ $c_i$ ≤ 8, depending on the derived core allele frequency in the 1000GD). However, no-one has studied $\phi_0$ and $\phi_i$ or $S_0$ and $S_1$, separately. We also divide the total number of derived alleles ($L_0$) within the D group into two components, $L_{01}$ and $L_{11}$, as follows:

$$L_i = \sum_{i=1}^{m-1} \phi_{imax} = \sum_{i=1}^{m-1} \phi_{0i} + \sum_{i=1}^{m-1} \phi_i, L_{01} = L_0 + L_1.$$ \hspace{1cm} (3b)

where $L_i$ is the sum of the scaled SFS (Fu, 1995) of $\phi_i$, $S_i$ and $L_i$ within the D group and is expected to be small under strong selective sweeps. However, it is desirable to exclude the possibility of either background selection (Charlesworth and Charlesworth, 2012 and references therein) or an intrinsically low population mutation rate in a core region ($\rho = 4N_0r$, where $u$ is the neutral mutation rate per region per generation). To have such internal control, we may also compute the number of derived alleles in the entire sample ($L_i = \sum_{i=1}^{m-1} k\phi_i$) and use it to

Fig. 1. Ancestral recombination graph and 2D SFS $\phi_{i,j}$ for $m = 8$ and $n = 14$. Edges and haplotype bars are in green for the D group and reddish-brown for the A group. Mutations are centered in a narrow range of $1 \leq i \leq i_{max}$ whereas under a soft sweep, $\phi_0 > 0$ spreads over a wide range of $i$; however, there is a gray zone in between. In the presence of recombination, $\phi_i > 0$ under a hard sweep is generally concentrated in $1 \leq i \leq i_{max}$ and $m - i_{max} \leq i < m$ because of the small family size within the D group. When $i_{max}$ differs from $i_{max}$, we consider the larger one, denoted by $i_{max}$, as the $\phi_i$-based estimate of the maximum family size within IAV. The size of $i_{max}$ and $i_{max}$ under a soft sweep may be much larger than those under a hard sweep, and recombination augments this trend.

Based on the different implications of $\phi_0$ and $\phi_i$, outlined above, we divide the total number of segregating sites ($S_i$) within the D group into two components, $S_0$ and $S_1$, as follows:

$$S_i = \sum_{i=1}^{m-1} \phi_i = \sum_{i=1}^{m-1} \phi_{0i} + \sum_{i=1}^{m-1} \phi_i, S_{01} = S_0 + S_1.$$  \hspace{1cm} (3a)

Akey et al. (2004) used $S_0$ to estimate the time elapsed since a selective sweep assuming a star-shaped genealogy, and Satta et al. (2018) studied the nonequilibrium distribution of $\phi_i$ normalized in individual frequency classes $c_i$ (1 ≤ $c_i$ ≤ 8, depending on the derived core allele frequency in the 1000GD). However, no-one has studied $\phi_0$ and $\phi_i$ or $S_0$ and $S_1$, separately. We also divide the total number of derived alleles ($L_i$) within the D group into two components, $L_{01}$ and $L_{11}$, as follows:

$$L_i = \sum_{i=1}^{m-1} \phi_{imax} = \sum_{i=1}^{m-1} \phi_{0i} + \sum_{i=1}^{m-1} \phi_i, L_{01} = L_0 + L_1.$$ \hspace{1cm} (3b)
standardize the IAV components in Formula (3b):

\[ L_{\phi} = L_{c_{0}}/L_{c}, \quad L_{c_{1}} = L_{c_{1}}/L_{c}, \quad \text{and} \quad L_{c} = L_{c}/L_{c_{1}} = L_{0} + L_{c_{1}}. \]

(4)

We emphasize that the unstandardized \( L_{0} \) is also important. Under the assumption of a molecular clock, Thomson et al. (2000) and Tang et al. (2002) used \( L_{0}/n \) as an estimator of the scaled TMRC (\( ut_{0} \), where \( T_{0} = t_{0}/4N_{e} \)) of a random sample of size \( n \). We note that \( L_{0}/n \) is the average number of mutations per lineage that have accumulated during \( t_{0} \) generations. Hudson (2007) provided the variance formula, \( V_{ut_{0}} \), of \( ut_{0} \). We can analogously estimate the scaled TMRC (\( ut_{p} \), where \( T_{p} = t_{p}/4N_{e} \)) of the D group from the average number of mutations per lineage (C) that accumulated during \( t_{p} \) generations. However, it is essential to exclude recombinant SNP sites so that we define the estimator as follows:

\[ ut_{p} = C = \frac{L_{\phi}}{m}. \]

(5a)

Furthermore, following Hudson (2007), we define its variance as

\[ V_{ut_{p}} = \frac{L_{\phi}}{m} - \frac{m - 1}{m} \frac{\theta_{\phi}}{2}, \]

(5b)

where \( \theta_{\phi} = \frac{1}{m} \sum_{i=1}^{m-1} i(m-i) \phi_{i,0} \) is the pairwise mean number of the mutational differences at the non-recombinant SNP sites within the D group. If we write \( B = L_{0}/n - C, \) \( B \) becomes equal to the number of mutations that have accumulated in the stem lineage of the D group. Thus, an alternative estimate of \( B \) is obtained by directly counting the number in the 2D SFS:

\[ B = \sum_{i=m-a_{0}}^{m-\gamma_{0}} \sum_{j=m-\gamma_{0}}^{m-\gamma_{0}} \phi_{i,j} = \sum_{i=m-\gamma_{0}}^{m-\gamma_{0}} \sum_{j=m-\gamma_{0}}^{m-\gamma_{0}} \xi_{i,j}. \]

(6)

We may use either \( B + C \) or \( L_{0}/n \) as an estimate of \( ut_{0} \). Under the standard model, we expect \( E[L_{c}] = \theta(n - 1) \) (Watterson, 1975; Fu, 1995), where \( E[\cdot] \) is an expectation operator, and \( E[B + C] = \theta(n - 1)/n \). In Formule (5a and b), we also note that \( u \) differs from region to region because \( u = \mu l \), where \( \mu \) is the mutation rate per site and \( l \) is the length of a (core) region. The \( \rho \) insensitivity of \( ut_{p} \) and \( V_{ut_{p}} \) is crucial when applied to autosomal regions. This property will be examined later by coalescent simulations.

Next, we are concerned with the average number of derived alleles per segregating site in three categories:

\[ G_{0} = L_{c_{0}}/S_{c_{0}}, \quad G_{c_{1}} = L_{c_{1}}/S_{c_{1}}, \quad \text{and} \quad G_{c} = L_{c}/S_{c}. \]

(7)

Provided that there is at least one segregating site in each category, these \( G \) statistics are indices for the multiplicity of derived alleles per segregating site within the D group. In particular, \( G_{0} \) can be referred to, by definition, as the average family size in the D group. All three categories should be small under a hard sweep if recombination is rare. Conversely, relatively large values of \( G_{0} \) and \( G_{c} \) signify a soft sweep and recombination, respectively. The expected value of \( G_{c} = L_{c}/S_{c} \) in the entire sample can be approximated by \( E[G_{c}] = E[ut_{c}]/E[S_{c}] \) for a large number of segregating sites, where \( S_{c} = \sum_{i=1}^{n-1} \xi_{i}. \) Using the standard model and by defining \( a_{n} = \sum_{k=1}^{n-1} 1/k \), we have \( E[G_{c}] = (n - 1)/a_{n} \), which is greater than 100 for \( n = 1,000, \) or each segregating site in this large sample is represented by more than 100 derived alleles on average.

Thus, the \( G \) statistics are rated “small” if they are much smaller than 100 in the 1000GD.

In addition to \( G_{0} \), we consider the following index of the IAV multiplicity:

\[ \gamma(i_{m}) = \frac{\sum_{i=1}^{m-1} \phi_{i,0}}{\sum_{i=1}^{m-1} \phi_{i,0}} \]

(8)

Formula (8) is based on the interpretation that the ratio of \( \xi_{i}/\sum_{i=1}^{m-1} \xi_{i} \) represents the probability of observing \( i \) derived alleles at a segregating site in a sample of size \( n \) (Kim and Stephan, 2002; Charlesworth and Charlesworth, 2012). \( G_{0} \) is more sensitive to \( \phi_{i,0} \) with large values of \( i \) than \( \gamma(i_{m}) \), although \( G_{0} \) and \( \gamma(i_{m}) \) are closely related. For a hard sweep with \( f_{c} = 0.6 \) or \( m = 600 \) in a sample of size \( n = 1,000, \) \( \gamma(i_{m}) = 0.1 \) will result in \( i_{m} = 6 \), irrespective of \( \rho \), and \( G_{0} \) may become about 6–7. However, both \( G_{0} \) and \( \gamma(i_{m}) \) depend strongly on the number of singletons (\( \phi_{i,0} \)). If there are too many singletons, these values are kept small, which is undesirable for distinguishing between hard and soft sweeps. Therefore, we also examine and may interchangeably use an alternative \( \gamma'(i_{m}) \) that excludes \( \phi_{i,0} \) from the definition given in Formula (8). Under the standard model, we have \( E[\gamma'(i_{m})] = (a_{n} - a_{m})/a_{n} \) and \( E[\gamma'(i_{m})] = (a_{n} - a_{m})/(a_{n} - 1) \), which become 0.59 and 0.69, respectively, if \( i_{m} = 10 \) and \( m = 600 \). Similarly, we examine \( G_{c}^{*} \) that excludes singletons while calculating \( G_{0} \). As a soft sweep is defined by the presence of more than one distinct lineage (or haplotype) in the D group, we may impose a criterion in terms of relatively large values of \( \gamma'(10) \) and \( G_{c}^{*} \).

Many inference procedures for selective sweeps are based on differences among the unbiased estimators of \( \theta \) (Tajima, 1989; Fay and Wu, 2000; Zeng et al., 2006, 2017; Ferretti et al., 2017). Under the standard model, we can obtain an estimate of \( \theta \) as \( \theta_{\phi} = \xi_{0} \) from the observed number of singletons (Fu and Li, 1993; Fu, 1995) or one of the following formulæ:

\[ \theta_{\phi} = \frac{1}{G_{0}} \sum_{i=k=1}^{m-1} \xi_{i}. \]

(9a)
Among six possible pairwise combinations of the four estimators in Formulae (9), four difference statistics have been proposed:

\[ D_i = \frac{(\theta_i - \theta_n)}{N_o(S_i)} \]  
\[ H_i = \frac{(\theta_i - \theta_n)}{N_o(S_i)} \]  
\[ E_i = \frac{(\theta_i - \theta_n)}{N_o(S_i)} \]  
\[ K_i = \frac{(\theta_i - \theta_n)}{N_o(S_i)} \]

\( N_o(S_i) \) is the standardizing coefficient, which can be expressed as follows:

\[ N_o(S_i) = \sqrt{\frac{n}{\lambda_o S_i} + \kappa_i S_i (S_i - 1)} \]  

with \( \lambda_o \) and \( \kappa_i \) specific to the \( \Omega \) statistic (Table 4 in Ferretti et al., 2017, but with the numerator \( n - 2 \) in \( \kappa_o \) being read as \( n + 2 \), as in Tajima, 1989). To quantify IAV based on these statistics, we simply replace \( \xi_k \) by \( \zeta_i \) and \( n \) by \( m \) in Formulae (9) and (10). We denote the latter as \( D_i, H_i, E_i \), and \( K_i \). Under a hard sweep and in the absence of recombination, it is expected that \( \theta_e > \theta_o > \theta_n \), and therefore, \( D_i < 0, E_i < 0, H_i > 0, K_i > 0 \). However, the mode of selective sweeps and the presence of recombination can greatly alter these expectations.

**The case of ST8SIA2** Before going further, let us exemplify these statistics with actual data: the promoter region of the EAS ST8SIA2 locus (Fujito et al., 2018a, 2018b). The core site consists of three promoter SNPs and divides the whole sample into CGC and non-CGC haplotype groups. The core region is ca. 15 kb long (chr 15, 92924515–92939701 in GRCh37), which is 1–3 kb shorter than that in the previous study (Fujito et al., 2018a) because of the exclusion of apparent recombinant sites located near the 5' and 3' ends. For the EAS sample in the 1000GD, we obtained the following values: \( S_i = 30, \theta_e = 0.013, \theta_o = 0.09, \theta_n = 0.002, L_{43} = 0.988 \), \( G_e = 47.0 (G_o = 1.52, G_n = 153) \), \( \kappa_0 = 4.67, \kappa_3 = 0.39 \), \( \kappa_4 = 4.05 \), \( \theta_e = 7.71, D_i = -2.47, H_i = -4.52, E_i = -0.30 \) and \( K_i = -0.93 \). We also obtained a small maximum family size of \( i_{\text{max}} = 4 \) from \( \phi_{i,0} = 1 \) and \( \phi_{i,4} = 0 \) or \( \gamma(i_m) = 0 \) for \( i_m \geq 5 \), demonstrating the reduced IAV in this genomic region. Consistent with these findings, the number of segregating sites \( (S_i) \) within the CGC haplotype group is much smaller than \( S_i = 132 \) in the entire sample. All the observed values of \( F_c, L_{0i}, G_o \), and \( D_i \) are small at highly significant levels of \( P < 0.001 \) in the simulated standard model. We observed that the small \( F_i \) value results in part from the exclusion of the derived alleles that belong to \( c = 7 \), or more precisely in the frequency range of \( m - i_{\text{max}} \leq i < m \) in Formula (2). If they had been included, \( F_i \) would have been as large as \( L_c \). Therefore, the exclusion procedure in \( F_i \) is legitimate because these derived alleles originated in the AMP group and were deceptively incorporated into the IAV by recombination. The \( G_i \) and \( L_{1i} = 40L_{0i} \) values show significant effects of recombination on the IAV. The negative \( D_i \) and \( H_i \) values are potentially useful to quantify an incomplete selective sweep, even though \( D_i \) and \( H_i \) for the entire sample lack their statistical powers (Table S5 in Fujito et al., 2018a). \( D_i < 0 \) reflects an excess of rare alleles and a lack of intermediate frequency alleles. \( H_i < 0 \) reflects a lack of intermediate frequency alleles and an excess of high frequency alleles. In the present case, an excess of high frequency alleles results from recombination, unlike their gradual increase toward fixation by positive selection. This is reflected in \( \zeta_i > 0 \) for \( m - i_{\text{max}} \leq i < m \) with \( i_{\text{max}} = 3 \). Finally, the selective sweep for the CGC haplotype group is so strong and recent that we may estimate the TMRCAs \( (t_D = C/\mu) \) of the CGC haplotype group as \( 12.2 \pm 3.9 \) thousand years (ky) from \( C = 32/349 = 0.092 \), the mutation rate per site per year of \( \mu = 0.5 \times 10^{-9} \) (Scally and Durbin, 2012), and \( \theta_e = 0.183 (B = 18 \) and \( L_{1i}/n = 15.6, \) a slightly smaller value than \( B + C \). Although the CGC haplotype originated in Africa long ago, positive selection began to exert pressure on a single ancestral haplotype lineage in East Asia around 12 ky ago (Fujito et al., 2018b). Therefore, the core region of ST8SIA2 provides a solid example of a hard sweep using standing variations (Przeworski et al., 2005; Schrider and Kern, 2016), namely hardening soft sweeps (Wilson et al., 2014).

**SIMULATED 2D SFS**

**Modes of selective sweeps and effects of recombination** We present a simulated 2D SFS in Supplementary Fig. S1 for a small sample of \( n = 68 \), under the conditions of the presence or absence of recombination and hard or soft sweeps for the D group (\( ms \) by Hudson, 2002; *disccoal* by Kern and Schrider, 2016). The methods for testing neutrality and examining statistical power are the same as described in Fujito et al. (2018a) and are not repeated here. An alternative representation of \( (\phi_i) \), as presented in Table 1, demonstrates that in the absence...
of recombination, only $\phi_{0,0}$, $\phi_{0,j}$ and $\phi_{m,j}$ elements can have non-zero values, or polymorphism is represented only by the three edges of the matrix $\{\phi_{i,j}\}$. This is because the D group is always monophyletic with respect to the A group and $\phi_{i,n-m} = 0$ holds true for all possible values of $i$. Conversely, the presence of only three different types of gametes along with $\phi_{n,m} > 0$ is a definite indication of recombination. If the D and A groups are reciprocally monophyletic, we also expect $\phi_{i,j} = 0$ for $1 \leq j < n - m$. In any case, the D and A groups evolve “independently” in the absence of recombination. It is recombination that generates the SNP sites with $\phi_{i,j} > 0$ for some values of $i$ in $1 \leq i < m$ and requires a 2D SFS to completely describe the pattern and level of polymorphism for incomplete selective sweeps.

Figure 2 shows that for hard sweeps, the expected values of $\phi_{0,0}$ and $\zeta_i$ divided by the number of their respective singletons decrease more rapidly than those of $1/h$ expected in the standard model (Fu, 1995) as $i$ increases. $\phi_{0,0}$ and $\zeta_i$ are subcritical, and they indicate that alleles in the D group do not manifest large family sizes of lineages that share particular mutations. The declining pattern of $\zeta_i$ depends on the recombination rate $\rho$. Contrastingly, $\phi_{0,0}$ is almost unchangeable by recombination. This remarkable robustness results from the fact that although recombination may occasionally contribute to $\phi_{0,0} > 0$, its effect is restricted to $i$ in $1 \leq i \leq i_{\text{max}}$ for some small number of $i_{\text{max}}$. In fact, $\phi_{0,0} > 0$ for large values of $i$ can be realized only by particular mutations in an ancestral recombination graph (Griffiths, 1991; Griffiths and Marjoram, 1997) together with fortuitous loss or no sampling of recombinant haplotypes in the A group. Under ordinary circumstances, $\phi_{0,0}$ therefore remains unaffected even by frequent recombination. Thus, $\theta$ and the variance in Formula (5) are also insensitive to recombination under selective sweeps (Supplementary Fig. S2). However, recombination dramatically increases $\zeta_i$, for $1 \leq i \leq i_{\text{max}}$ and $m - i_{\text{max}} \leq i < m$ (Fig. 2). Although $\zeta_i > 0$ in the two opposite ranges of $i$ is somehow related to the genealogical structure of IAV, the outcome may suffer from repeated occurrences of recombination, thereby no
longer reflecting the structure accurately. In fact, $\tilde{\zeta}_i > 0$ for $m - \hat{i}_{\text{max}} \leq i < m$ can occur much more frequently than $1/\alpha$, irrespective of $f_i$, indicating that it is related to the recombination processes between the D and A groups and not the coalescent processes with mutations within the D group.

**Statistical power** To further quantify $\varphi_{0,0}$, $\tilde{\zeta}_i$, and their related summary statistics, we performed extensive coalescent simulations. The results of these simulations are summarized semi-quantitatively in Table 2, which gives the empirical criteria to evaluate the significance of sweep signals, classify the mode of selective sweep, and assess the effects of recombination. The family size is independent of $\rho$, as it should be, but is strongly dependent on the mode and strength of the selective sweeps. Supplementary Table S1 presents 18 summary statistics for IAV under hard sweeps when $\theta = 20$, $n = 1,000$, $a = 2N_s = 500$ (s: genic selection intensity), and $f_i = 0.3$, 0.6 and 0.9. As mentioned above, $\varphi_{0,0}$ is insensitive to $\rho$ and so are the related $S_{\varphi 0}$, $L_0$ and $G_{\varphi 0}$. In contrast, none of the four $\zeta_i$-based estimators of $\theta$ are insensitive to $\rho$. This sensitivity is also reflected in $H_c$, $E_c$ and $K$, all of which can take positive and negative values as $\rho$ changes. However, there is an exception to this large fluctuation. $D_{\zeta}$ can keep a negative value within the examined parameter range because it depends on $\theta_i$ and $\theta_h$ and not on $\theta_0$ and $\theta_0$, which are much more sensitive to $\rho$ than the former two. In other words, recombination generates superficial signals on $\theta_i$ and $\theta_h$ through $\tilde{\zeta}_i > 0$ for large values of $i$. These results suggest that we cannot use $H_c$, $E_c$ or $K$ without knowing the accurate values of $\rho$ in actual data analyses.

Figure 3 shows the receiver operating characteristic (ROC) curves comparing the sensitivity (true-positive rate or power, $1 - \beta$) and 1 − specificity (false-positive rate, $\alpha$) of $F_c$, $L_0$, $G_0$, $G_{\varphi 0}$, as well as those of $D_{\zeta}$, $H_c$, $E_c$ and $K$. The former five are tested in left-tailed tests and the latter four are tested in two-tailed tests (Fig. 2 in Ferretti et al., 2017). Each ROC curve was constructed from 1,000 simulations under strong selection ($a = 500$) and 1,000 simulations under near neutrality ($a = 1$, $f_0 = 0.5$ and $f_i = 0.6$) to facilitate random sampling of the derived alleles with a specified frequency ($f_i$). Thus, our ROC curves are conservative compared to those when pure neutrality is assumed as a null hypothesis, and a comparison between pure neutrality (by $ms$) and near neutrality (by discoal) indicates visible effects of $a = 1$ on the 2D SFS (Supplementary Fig. S3). The area under an ROC curve, often denoted as AU(RO)C, measures the discriminative ability of a statistic. For hard sweeps (Fig. 3A) and with $\rho = 0$, $F_c$, $L_0$ and $G_{\varphi 0}$ can discriminate the hypotheses well and may be categorized as excellent, as a rule of thumb. Among the four difference statistics, $D_{\zeta}$ outperforms $H_c$, $E_c$ and $K$. However, if $\rho = 50$, $L_0$ and $G_{\varphi 0}$ outperform $F_c$, while $H_c$ and $K$ outperform $E_c$ and $D_{\zeta}$. Recombination reduces the power of $F_c$, because this statistic is not completely free from the effects of recombination, and, thus, is also reflected in the relatively low power of $G_c$ and $L_c$. The powers of $H_c$, $K$ and $E_c$ become even greater when $\rho = 50$ than when $\rho = 0$. This unexpectedly enhanced power results from the increased values of $\zeta_i$ for large values of $i$ because of the high recombination rate. Overall, for hard sweeps, (i) it is possible to infer the maximum and average family sizes among D group lineages; (ii) to evaluate IAV, it is recommended to use $\varphi_{0,0}$ and closely related summary statistics, such as $G_{\varphi 0}$, $L_0$ and $F_c$; and (iii) the power of some difference statistics is extremely sensitive to recombination, and, thus, caution is required when interpreting the application results. Regarding (iii), an alternative may be to recalculate the difference statistics based on $\varphi_{0,0}$ rather than $\zeta_i$.

A soft sweep can be modeled by a high initial frequency of the D group, such as $f_0 = 0.1$ (Ferrer-Admetlla et al., 2014; Kern and Schrider, 2016), and it generates SNP sites with $\varphi_{0,0} > 0$ for multiple intermediate values of $i$. Although effects of recombination on the pattern of $\tilde{\zeta}_i$ are similar to those of hard sweeps, IAV for

| Table 2. Semi-quantitative criteria for detecting and classifying incomplete selective sweeps and evaluating the significance of recombination effects |
|---------------------------------|---------------------------------|---------------------------------|
| **Level** | **IAV Multiplicity** | **$\tilde{\zeta}_i = 0$ for all values of $i$** | **$\tilde{\zeta}_i > 0$ for some values of $i$** |
| **Low** | **Small $i_{\text{max}}$** | **Hard sweep** | **Hard sweep** |
| $F_c < 1$ | $\gamma'(10) < 0.1$ | $G_{\varphi 0} < 9$ | No recombination | Recombination |
| $L_0 < 1$ | | | | |
| **High** | **Large $i_{\text{max}}$** | **Soft sweep** | **Soft sweep** |
| $0.1 < \gamma'(10)$ | $9 \leq G_{\varphi 0}$ | No recombination | Recombination |
A

$\rho = 0, \ f_r = 0.6$

$\rho = 50, \ f_r = 0.6$

$\rho = 0, \ f_0 = 0.1, \ f_r = 0.6$

$\rho = 0, \ f_0 = 0.05, \ f_r = 0.6$

$\rho = 0, \ f_0 = 0.1, \ f_r = 0.9$

B

$\rho = 0, \ f_0 = 0.5, \ f_r = 0.6$

$\rho = 0, \ f_0 = 0.1, \ f_r = 0.9$

Fig. 3. ROC curves in ongoing hard and soft sweeps obtained for $F_c$, $G_{c0}$, $G_c$, $L_{c0}$ and $L_c$ (left panels) and for $D_\tau$, $H_\tau$, $E_\tau$ and $K_\tau$ (right panels). The ordinate is the power $(1 - \beta)$ with a type II error rate $\beta$ under positive selection with $\alpha = 500$, and the abscissa is the false-positive rate $\alpha$ under near neutrality with $\alpha = 1$. In all cases: 1,000 replications, $\theta = 20$ and $n = 1,000$. Two-tailed tests are labeled as $\alpha/2$ and one-tailed tests are labeled as $\alpha$. (A) Hard sweeps when $f_r = 0.6$ for $\rho = 0$ and 50. (B) Soft sweeps when $\rho = 0$ for $f_0 = 0.1$ and $f_r = 0.6$; $f_0 = 0.05$ and $f_r = 0.6$; $f_0 = 0.1$ and $f_r = 0.9$. 

Genomic signatures of incomplete selective sweeps
soft sweeps can be considerably high such that \( \Phi_i, \rho \) and \( \zeta_i \), spread broadly in the \( 1 \leq i < m \) range. For the soft sweep in Fig. 2B and in a specified range of \( 1 \leq i \leq 30 \), for example, we observed \( \gamma(10) = 0.18 \) for \( \rho = 50, 20 \) and \( 0 \). These values are significantly larger than those for the hard sweeps shown in Fig. 2A, i.e., \( \gamma(10) = 0.03 \) for \( \rho = 50, 20 \) and \( 0 \). Supplementary Table S2 presents the 18 statistics for soft sweeps. As mentioned above, \( \Phi_i, \rho \)-related statistics are remarkably robust to recombination, although this does not guarantee their ability to identify soft sweeps. In fact, the absolute values are substantially larger than those for the corresponding hard sweeps. With increased \( \Phi_i, \rho \) and \( \zeta_i \) for intermediate values of \( i \), all other statistics are significantly affected. The ROC curves for \( f_0 = 0.1, f_1 = 0.6 \) and \( \rho = 0 \) (Fig. 3B) show lower discriminative abilities of \( F_c, L_0, \) and \( G_0 \) and almost no power of \( D, H_2, E_3 \) and \( K_3 \) for low cutoff values. Detecting soft sweeps strongly depends on the typically unknown parameter \( f_0 \). If we reduce the value of \( f_0 \) from 0.1 to 0.05, the powers of some statistics are recovered. If we increase the value of \( f_0 \) from 0.6 to 0.9, the powers of some statistics are recovered as well.

The duration time \( (T) \) of the selected phase in our simulation is approximately given by \( T = \frac{1}{4N} \ln \left( \frac{1-f_0}{f_0(1-f_0)} \right) \) in units of \( 4N \) generations (Kern and Schrider, 2016); \( T = 2.60/\alpha \) for \( f_0 = 0.1 \) and \( f_1 = 0.6, 3.35/\alpha \) for \( f_0 = 0.05 \) and \( f_1 = 0.6 \), and \( 4.39/\alpha \) for \( f_0 = 0.1 \) and \( f_1 = 0.9 \). All these duration times are shorter than 0.01 if \( \alpha \geq 500 \). Considering that \( T = \frac{\gamma\tau}{\ln(1-\tau)} \) and \( T_D = \frac{\gamma\tau}{\ln(1-\tau)} \) in Formula (5a), we may use them to formally redefine the hard and soft sweeps. A selective sweep would be hard if \( T_D \leq T \) \((t_D \leq t)\), and vice versa. Similarly, a selective sweep would be soft if \( T_D > T \) \((t_D > t)\), or, equivalently, if more than one distinct ancestral lineage is assured beyond \( T \), and vice versa. The frequency change of the \( D \) group depends on the strength and mode of selection (Wilde et al., 2014), and the same is true for \( T \) and \( T_D \). \( T \) is a major factor that affects the power of a statistic wherein \( f_0 \) becomes a more crucial determinant than \( f_1 \) (for the case of \( f_1 = 1 \), see Innan and Kim, 2004; Przeworski et al., 2005). The increased extent of the power varies largely among the statistics used, i.e., \( F_c, L_0, \) and \( G_0 \) are rather robust to changes in \( f_0 \) and \( f_1 \), whereas \( D, E_3 \) are sensitive, but \( H_2 \) is insensitive. Generally, the power of detecting soft sweeps of \( f_0 = 0.1 \) by a single SFS statistic may not be sufficiently high for a given low false-positive rate \( \alpha \). Readers may refer to Nielsen et al. (2005), Zeng et al. (2006, 2007), Grossman et al. (2010), Ayub et al. (2013), Sheehan and Song (2016) and Schrider and Kern (2017) for improved methods that combine several population genetic summary statistics and related issues.

Past incomplete selective sweeps

Any method would lose the ability to reject the null model in favor of a selective sweep with the time \( \tau \) in units of \( 4N \) generations elapsed since the fixation of a beneficial allele (Simonsen et al., 1995; Kim and Stephan, 2002; Przeworski, 2003). As sweep signals dissipate rapidly due to new mutations and recombination, the detectable period \( \tau \) of a selective sweep may not greatly exceed 0.1 (Przeworski, 2003; Ferretti et al., 2017). In the case of modern humans, this may amount to 4,000 generations or \( 10^6 \) years if \( N_e = 10^3 \) and the generation time is 25 years. Similarly, we are concerned with the possibility that positive selection would cease before a selective sweep is completed, presumably because of the ever-changing physical and/or cultural environment. We want to determine the smallest value of \( \tau \) that can be statistically distinguishable from \( \tau = 0 \) (ongoing sweep) and the statistic that is most suitable for this distinction. If \( \tau > 0 \), the condition for a past hard sweep becomes \( T_D \leq T + \tau \), and that for a past soft sweep becomes \( T_D > T + \tau \). During the time period of \( \tau \) (forward in time), the \( D \) group frequency changes from \( f_1 \) to \( f_1(\tau) \) by genetic drift. Therefore, in a sample taken at \( \tau \), we observe the \( D \) group alleles with a number that may differ greatly from \( n_f = m \) and may range from 0 (loss) to \( n \) (fixation in the sample). Theoretically, a large number of coalescent events can occur even during a time period as short as \( \tau = 0.05 \). Tavaré (1984) and Takahata and Nei (1985) derived two seemingly different formulae for the probability \( g_0(2\tau) \) of the number \( k \) of distinct ancestors \( 2\tau \) units of time ago in a sample of size \( n \) wherein time \( \tau \) is measured in units of \( 2N \) generations (Chen, 2012 for the case of time-varying population sizes). Wakeley (2009) highlighted their equivalence, provided that the denominator in his formula (3.42) is read as \( k!/(i-k)!n_{i-k} \) rather than \( i!(i-k)!n_{i-k} \). We examined \( F_c, L_0, G_0 \) and \( D_1 \) to determine how their ability to detect ongoing selective sweeps declines with \( \tau \), and we measured their power at \( \tau = 0.01, 0.03 \) and 0.05 when a type I error rate \( \alpha \) was specified as 0.01. Figure shows that for hard sweeps, the \( F_c \) power fades most rapidly, followed by \( L_0 \). However, the \( G_0 \) power remains high for a relatively long time, and the \( D_1 \) power exhibits a complicated pattern. The primary change in these statistics results from the newly created \( \Phi_0 > 0 \) for the relatively large values of \( i \) or increased \( \tau_{max} \) due to coalescence in an elongated neutral phase of a post-sweep (Supplementary Fig. S5). \( S_0 \) and \( L_0 \) increase proportionally with \( \tau \), as does \( L_0 = L_0/L_0 \); however, \( G_0 = L_0/L_0 \); \( S_0 \) increases more slowly and retains its power for a relatively long time. It is rather unexpected that \( G_0 \) does not easily distinguish between ongoing and past selective sweeps. In contrast, as the power of \( F_c \) reduces to 50% or less even at \( \tau = 0.01 \), it would be most useful to evaluate...
if $\tau > 0$. The presence of recombination enhances this decline dramatically.

Overall, if we observe significantly small values of $F_c$, $L_{c0}$, $G_{c0}$ and $D_\zeta$ we identify an ongoing hard sweep. If we observe significantly small values of $i_{\text{max}}$ and $\gamma(i_{\text{min}})$, we identify a past hard sweep. The same conclusions are applied to the case of soft sweeps (Fig. 4). However, the powers of $F_c$, $L_{c0}$, $G_{c0}$ and $D_\zeta$ for soft sweeps may be low even at $\tau = 0$ and approach the specified level of $\alpha$ rapidly with $\tau$. In hard and soft sweeps, recombination plays a significant role in diminishing the powers of $F_c$ and $D_\zeta$, whereas $G_{c0}$ and $L_{c0}$ are somewhat resistant.

APPLICATIONS TO HUMAN GENOMES

A genomic landscape revealed by 100 intronic SNPs In this section, we render a genomic landscape with a 2D SFS, although it is not exhaustive. By filtering biallelic SNP sites in the 1000GD by autosomal introns and Global MAF (ca. 40%), we randomly chose >100 core sites, and examined LD and 2D SFS in the EAS population (see Supplementary Table S3 for their SNP reference
sequence (rs numbers). To facilitate the determination of a core region for each core site, we used a distance similar to the “half length” of LD (Reich et al., 2001), which is the distance from a core site at which LD drops below 1/2. However, we assumed a more tightly linked core region than expected by the half length. For this, we used $r^2$ (Hill and Robertson, 1968), took the average over SNP sites with frequencies greater than 0.1 in a 1-kb non-overlapping window, and determined the distance at which the averaged $r^2$ value drops below 3/4. Two such distances on both sides of a core site mark the boundaries of the core region. The size determined in this way varies greatly among the core regions. In many cases, the averaged $r^2$ value decreases below 3/4 even in the first 1-kb window from a core site (for instance, rs1484214), and in other cases it extends over a 150-kb region (for instance, rs7541057). In total, we excluded 45 sites with <1-kb core regions and added the same number of sites to keep the total number of 100 core sites. We confirmed that those 45 excluded SNPs do not show any signals of selective sweeps.

Assuming near neutrality of the 100 core variants, no recombination within a core region, and the nonequilibrium demography of Schaffer et al. (2005), we computed the $P$ values of $F_c$, $L_{c0}$, $G_{c0}$ and $D_\zeta$, and examined the uniformity of their distributions by the chi-squared and Kolmogorov-Smirnov (KS with $n = 100$) tests as well as other multiple testing approaches (Fig. 5). The distribution for $F_c$ is fairly uniform and does not reject any core site ($\chi^2_{df=9} = 5.0$ with $P = 0.8$ and $g_{max} = 0.089$ with $P > 0.05$ in the KS test). Similarly, we obtained $\chi^2_{df=9} = 13.95$ ($P > 0.05$) and $g_{max} = 0.144$ ($0.05 > P > 0.01$) for $L_{c0}$, $\chi^2_{df=9} = 15.0$ ($P < 0.05$) and $g_{max} = 0.106$ ($P > 0.05$) for $G_{c0}$, and $\chi^2_{df=9} = 19.1$ ($P < 0.01$) and $g_{max} = 0.151$ ($0.05 > P > 0.01$) for $D_\zeta$. Thus, there are some indications against the null hypothesis. The multiple testing of $F_c$, $L_{c0}$, $G_{c0}$ and $D_\zeta$ (Benjamini and Hochberg, 1995) consistently rejects zero, five, eight and eight core sites, respectively, with a 10% false discovery rate (FDR). Such different performances among the statistics result partly from their different sensitivities to recombination. The least sensitive are $G_{c0}$ and $L_{c0}$, the second least sensitive is $F_c$, and the

![Histogram and cumulative distributions of 100 P values of $F_c$, $L_{c0}$, $G_{c0}$ and $D_\zeta$ at biallelic intronic SNP sites randomly chosen from EAS in the 1000GD. A significant departure from uniform distribution occurs for $P < 0.1$ except for $F_c$ (see text).](image-url)
most sensitive is $D_\gamma$. If $L_{v0}$, $G_{v0}$ and $D_\gamma$ are used, at most one core site is erroneously rejected and the remaining four to seven support false null hypotheses. Thus, with a 10% FDR, $\frac{137 - 140}{145} = 95 - 97\%$ of intronic variants in human genomes are consistent with selective neutrality (Kimura, 1968). We also applied the same multiple testing to Fisher’s combined $P$ values of $F_\gamma$, $L_{v0}$ and $G_{v0}$, and obtained 10 loci for $P < 0.01$. This is consistent with the result of the independent tests, but it is overstated because of positive dependence among the three statistics. Regarding the proportion $p_0$ of the true null hypotheses among all the hypotheses (Storey and Tibshirani, 2003; Benjamini et al., 2006), we estimated $p_0$ to be 0.90 from the multiple testing and histogram (Fig. 5). The most significant deviation from neutrality is observed at an intron variant C/T (rs1587725) that is located in a serine/threonine kinase 32 (STK32B) and associated with an autosomal recessive skeletal dysplasia. The C frequency in the African population is $<10\%$, but approximately $80\%$ in the non-African populations.

It is worth noting that these conclusions are based on the null hypothesis, which assumes the absence of recombination in addition to neutrality. Recombination affects the distributions of $F_\gamma$, $L_{v0}$, $G_{v0}$ and $D_\gamma$ by decreasing $S_{v0}$ and contrastingly increasing $S_\gamma$ in IAV (Supplementary Fig. S6). Although such effects of recombination are small for the case of selective sweeps with low levels of IAV, they can be substantial for the case of neutrality with relatively high levels of IAV. In the case of neutrality, recombination increases the expected values of $F_\gamma$ and $D_\gamma$ so that their observed values become relatively small. For the same reason, recombination decreases the expected values of $L_{v0}$ and $G_{v0}$ or makes the distributions skew toward 0. Thus, in the presence of recombination, $F_\gamma$- and $D_\gamma$-based tests become too sensitive, but $L_{v0}$- and $G_{v0}$-based tests become conservative. It is difficult to choose an appropriate recombination rate when testing a null hypothesis. However, the fact that the $P$ values of the four statistics are approximately uniformly distributed under the null hypothesis of no recombination suggests that the effect of recombination is not substantial in the core region with $r^2 \geq 3/4$ despite a large number of segregating sites that indicate recombination. Simulated $r^2$ values for various recombination rates are presented in Supplementary Fig. S7.

**Six selected regions** We also examined six genomic regions encompassing the genetic loci of lactase ($LCT$) (Enatnah et al., 2002; Tishkoff et al., 2007), ectodysplasin A receptor ($EDAR$) (Sabeti et al., 2007; Fujimoto et al., 2008; Kimura et al., 2009; Kamberov et al., 2013), abnormal spindle-like microcephaly ($ASPM$) (Evans et al., 2005), microcephalin-1 ($MCPH1$) (Mekel-Bobrov et al., 2005), neuregulin-1 ($NRG1$), and a disintegrin and metalloproteinase domain 17 ($ADAM17$) (Pickrell et al., 2009). We selected these regions as representatives to test the power of our statistics. The six regions include two well-established cases, $LCT$ (rs48988235) and $EDAR$ (rs3827760), for positive selection, and two controversial cases, $ASPM$ (rs41310927) and $MCPH1$ (rs930557), in relation to an LRH-based outlier approach (Yu et al., 2007) as well as allelic surfing (Currat et al., 2006). Further, to extend our previous study on ST8SIA2 (Fujito et al., 2018b), the regions include $NRG1$ (rs3924999) and $ADAM17$ (rs2709591), which are variants that are known to be associated with the risks of schizophrenia and various psychiatric phenotypes (Pickrell et al., 2009). The six core regions were determined in the same way as described in the 100-intron genomic landscape (Supplementary Fig. S8) and the observed values of several statistics are summarized in Table 3. First, we examined the level of the IAV through $F_\gamma$ and $L_{v0}$, the multiplicity through $\gamma(i_{max})$ and $G_{v0}^*$, and the roles of recombination through $G_{\gamma}$ and $L_{v0}$ (Table 2). The $P$ values were computed in the null hypothesis of pure neutrality in the non-equilibrium demographic model of the EAS or European (EUR) population (Schaffner et al., 2005) with the number

| Table 3 | IAV statistics for $LCT$ and $ASPM$ in EUR ($n = 1,006$) and for $EDAR$, $MCPH1$, $NRG1$ and $ADAM17$ in EAS ($n = 1,008$) in the 1000GD |
|---|---|---|---|---|---|---|---|
| Loci (sweep mode) | $LCT$ (hard) | $EDAR$ (hard) | $ASPM$ (NS) | $MCPH1$ (soft) | $ADAM17$ (soft) | $NRG1$ (NS) |
| $m$ | $l$ | 511 (100 kb) | 880 (65 kb) | 413 (165 kb) | 770 (32 kb) | 871 (26 kb) | 765 (3 kb) |
| $S_{v0}$, $S_\gamma$ | 66, 10 | 223, 67 | 279, 78 | 196, 92 | 111, 19 | 10, 3 |
| $L_{v0}$, $L_\gamma$ | 84, 2056 | 529, 3308 | 1394, 15348 | 785, 40146 | 561, 7840 | 19, 761 |
| $F_\gamma$ | 0.0054*** | 0.257* | 0.123 | 0.103*** | 0.229* | 0.070 |
| $L_{v0}$ | 0.0015*** | 0.0091*** | 0.013 | 0.013*** | 0.023*** | 0.010 |
| $i_{max}$ | 6*** | 48*** | 187 | 85*** | 213 | 7 |
| $\gamma(10)$, $G_{v0}^*$ | 0***, 2.8*** | 0.08***, 5.5*** | 0.33, 13 | 0.27, 10.2*** | 0.31, 13.5*** | 0.3 |

The $P$ values of $F_\gamma$, $L_{v0}$, $i_{max}$, $\gamma(10)$ and $G_{v0}^*$ for each gene were obtained by $ms$ (Hudson, 2002) using the observed $S_\gamma$ value and demographic model by Schaffner et al. (2005); *** $P < 0.001$, ** $0.001 \leq P < 0.01$, * $0.01 \leq P < 0.05$ (see Supplementary Table S3 for the full table). NS: no evidence for a selective sweep.
of segregating sites being fixed to an observed one. Second, we estimated the TMRCA ($t_D$) in the D group from the observed $C$ value, the mutation rate $\mu$, and the 3/4 length ($l$) core region.

For $LCT$ in the EUR population, the core region became $l = 100$ kb long (Chr 2, 136608646–136708564). If we instead use the 1/2 length, the core spans a 600-kb genomic region. When $l = 100$ kb, there are 76 segregating sites within the D group, which are 14% (< $f_s$ = 51%) of the total in the same region. The IAV level is very low and provides the strongest evidence for a selective sweep. The estimated maximum and average family sizes are so small that the majority of the $S_{\text{fr}}$ = 66 sites are singletons ($G_{\text{fr}}$ < 3), and there is no segregating site with a multiplicity as high as $10 \leq i$ or $\gamma(10) \leq \gamma(10) = 0$. Thus, the IAV in the T-13910 group—100% association with lactase persistence in the Finnish population—is unambiguously classified as an ongoing hard sweep (Table 2; Peter et al., 2012; Ronen et al., 2015). Nevertheless, there are additional $S_{\text{fr}}$ = 10 segregating sites and $L_{\text{fr}}$ = 2,056 derived alleles that are linked to the D group by recombination. This non-negligible effect of recombination is quantified by $G_{\text{fr}}$ and $L_{\text{fr}}$. Further, we obtained $t_D = 3,280 \pm 480$ years from $C = 0.16$ as well as $V_{t_D} = 5.7 \times 10^{-4}$ with $\theta_\varphi = 0.33$ ($B = 52$ and $L_{\text{fr}}/n = 55.8$, a slightly larger value than $B + C$). As mentioned earlier, this estimate of $t_D$ for a hard sweep provides a lower bound of the time period ($t$) of the selected phase. The simulation showed that $t$ can be much longer than $t_D$ (data not shown). Our estimate of $t_D$ is as short as a recent onset of positive selection in Bronze Age Europeans (Allentoft et al., 2015), shorter than 4,000 years based on ancient DNA (Mathieson et al., 2015), much shorter than 7,500 years based on the ABC approach (Itan et al., 2009), but within a wide range of allele ages roughly estimated by LRH conservation (Bersaglieri et al., 2004).

We chose the core site of EAS EDAR at the non-synonymous variant site 1540. The 3/4 length core region becomes 65 kb long, while the 1/2 length core region spans a region of > 500 kb (Chr 2, 109465127–109530021). There are 290 segregating sites within the high-frequency D group ($f_s = 0.87$), which are about 59% of the total in the same region. The IAV level is as low as $L_{\text{fr}} = 0.0091$, but $F_r = 0.257$ is higher than 0.196 in our previous estimate based on a different core region (Fujito et al., 2018a). The IAV multiplicity with $i_{\text{max}} = 48, G_{\text{fr}}* \approx 5.5, \text{ and } \gamma(10) < 0.1$ is low, but higher than that in $LCT$. Including the two sites with $i = 120$ and $j = 48$, and with $i = 384$ and $j = 17$, the family size in the D group is fairly large, and recombination apparently plays a substantial role in increasing $G_{\text{fr}}$ and $L_{\text{fr}}$. From $C = 0.60$ and $\theta_\varphi = 1.19$ ($B = 58$ and $L_{\text{fr}}/n = 58.3$, in agreement with $B + C$), we estimated $t_D = 18.5 \pm 2.5$ ky. The derived core allele may not have originated in the main ancestral populations of East Asians (Mathieson et al., 2015). Yet all these features are best classified as a hard sweep that has been operating at least over the past 18.5 ky ($t > t_D$).

The core region of ASPM in the EUR population is 165 kb long (Chr 1, 196990327–197155021). As this region is more than five times longer than that in our previous analysis (Fujito et al., 2018a), it includes a large number of segregating sites in the D group. The IAV level is inflated in terms of $L_{\text{fr}}$ and $F_r$. The IAV multiplicity is accordingly high and exhibits a fairly large family size ($i_{\text{max}} = 187, i_{\text{max}}* = 143, G_{\text{fr}}* \approx 13, \text{ and } \gamma(10) = 0.33$). Despite the relatively low frequency, $f_s = 0.41$, significant roles of recombination are apparent. These features are consistent with previous work (Yu et al., 2007) that showed that the region is not identified empirically as an outlier in the LRH test. The estimates of $C = 3.38$ and $\theta_\varphi = 5.91$ ($B = 113$ and $L_{\text{fr}}/n = 106$, a slightly smaller value than $B + C$) led to $t_D = 41 \pm 7.9$ ky.

Some genetic variants of the $MCPH1$ gene rapidly increased their frequencies in Eurasia. The 3/4 length core region in the EAS population is 32 kb long (Chr 8, 6286532–6318958) and there are 288 segregating sites (65% of the total) in the D group that show significantly low levels of $F_r$ and $L_{\text{fr}}$. However, the core region exhibits relatively large family sizes ($i_{\text{max}} = 85, i_{\text{max}}* = 76, G_{\text{fr}}* \approx 10, \text{ and } \gamma(10) = 0.27$) and abundant recombinant haplotypes. These features are classified as an ongoing soft sweep with significant effects of recombination. From $C = 1.0$ and $\theta_\varphi = 2.0$ ($B = 64$ and $L_{\text{fr}}/n = 61.7$, a slightly smaller value than $B + C$), we have $t < t_D = 62.5 \pm 13.2$ ky, indicating that the selection began to operate after the major out-of-Africa dispersal (Bae et al., 2017; Mondal et al., 2019). However, in the EUR population, the IAV pattern and level differ significantly. The core region becomes 30 kb long (Chr 8, 6284798–6309364), and it shows a high IAV level and multiplicity with large values of $F_r$ and $L_{\text{fr}}$ as well as those of $i_{\text{max}} = 802, i_{\text{max}}* = 9, G_{\text{fr}}* \approx 141, \text{ and } \gamma(10) = 0.47$. $i_{\text{max}}$ has the largest possible value when $m = 803$, and even if this $\phi_{02,0} = 4$ is excluded, there still exist four sites of $\phi_{01,0} = 3, \phi_{05,0} = 1, \phi_{33,0} = 1, \text{ and } \phi_{01,1} = 1$. The estimate of $C = 8.7$ in the D group ($\theta_\varphi = 2.35, B = 37 \text{ but } L_{\text{fr}}/n = 37.3$) shows $t_D = 580 \pm 180$ ky, which corresponds to a time much earlier than the origin of modern humans (Bae et al., 2017). There is no indication of any selective sweep for the $MCPH1$-derived allele at rs930557 in the EUR population.

ADAM17 converts NRG1 to its active form in the NRG-ERBB4 signaling pathway, and both genomic regions exhibit iHS selection signals in the EAS population (Pickrell et al., 2009). The core region of ADAM17 is 26 kb long (Chr 2, 94900015–9515668). The IAV level is low in terms of $F_r$ and $L_{\text{fr}}$, but the IAV multiplicity is high ($i_{\text{max}} = 213, i_{\text{max}}* = 3, G_{\text{fr}}* \approx 13.5, \text{ and } \gamma(10) = 0.31$). The large value of $i_{\text{max}}$ results from the single occurrence of $\varphi_{213,0} = 1$. If that is excluded, the value of $i_{\text{max}}$ drops
substantially, but is still as large as 48. These features may well be classified as an ongoing soft sweep. Using $C = 0.64$ and $\theta_0 = 1.16$ ($B = 24$ and $L/n = 24.3$, in agreement with $B + C$), we have $t_0 = 49 \pm 20 \text{ ky} (> t)$. On the other hand, the core region of NRG1 in the EAS population becomes only 3 kb long (Chr 8, 32451286–32454158), the shortest among the six genes and containing only 13 segregating sites (72% of the total) in the D group. The small core region shows a relatively small family size ($i_{\text{max}} = 7$, $j_{\text{max}} = 6$, $G^G_{\gamma \theta} = 3.3$, and $\gamma(10) = 0$), but it results in a broad distribution of any statistic and a large type I error rate. We have $17 \pm 7.4 \text{ ky from } C = 0.025$ and $\theta_0 = 0.05$ ($B = 2$ and $L/n = 1.8$, in good agreement with $B + C$). There is no statistically significant indication of a selective sweep for NRG1.

Of the 18 summary statistics mentioned in this paper, we have paid special attention to each of $F_r$, $L{\theta_0}$, and $G{\theta_0}$ to infer significant signals of positive selection, the mode of selective sweeps, and the role of recombination. In certain circumstances, however, we may wish to evaluate the overall significance level, which is expressed by combined tests rather than individual ones. Ayub et al. (2013) used the aforementioned Fisher's combined method. Alternatively, we may follow Zeng et al. (2006, 2007), who proposed a compound test for multiple statistics such as $D_2$ and $H_\delta$, or Grossman et al. (2010), who developed a composite method of multiple signals. Of these three, as Fisher's method is straightforward, we used it here again even though the $F_r$-$L{\theta_0}$-$G{\theta_0}$-based tests are positively correlated with each other and accordingly the small combined $P$ value is generally overstated. We obtained $\chi^2_{\delta/-e} = 41.4 (P = 2.4 \times 10^{-7})$ for LCT, $\chi^2_{\delta/-e} = 34.1 (P = 6.4 \times 10^{-5})$ for EDAR, $\chi^2_{\delta/-e} = 29.3 (P = 5.2 \times 10^{-5})$ for MCPHI, $\chi^2_{\delta/-e} = 25.6 (P = 2.7 \times 10^{-4})$ for ADAM17, $\chi^2_{\delta/-e} = 13.3 (P = 0.038)$ for NRG1, and $\chi^2_{\delta/-e} = 10.2 (P = 0.117)$ for ASPM. Not surprisingly, these significance levels measured by the combined $P$ values agree well with our conclusions based on individual tests (Table 3).

We compared the above results with those of nSL (Ferrer-Admetlla et al., 2014) using selscan software (version 1.2.0a) for common biallelic variants (MAF > 1%). Specifying an optional command of “--max-extend” and checking the significant decay of the EHH curves, we set the stopping condition for nSL computations as 1,500 loci. We observed that the normalized nSL is significantly large for LCT ($P < 0.0001$), EDAR ($P = 0.0003$) and ADAM17 ($P = 0.0096$); marginal for NRG1 ($P = 0.029$); and non-significant for ASPM ($P = 0.418$) and MCPHI ($P = 0.263$). These $P$ values for LCT, EDAR, ASPM, and ADAM17 agree with our results, but those for MCPHI and NRG1 somewhat differ. The difference in MCPHI probably results from different sensitivities to coupled effects of recombination and a soft sweep. As the MCPHI core region exhibits the highest proportion ($S_{\gamma}/S_\gamma$) of segregating sites that underwent recombina-

CONCLUSIONS

In this study, we have shown that the pattern and level of IAV under an incomplete selective sweep can be satisfactorily described by 2D SFS. This allows the separation of segregating sites that have experienced recombination from those that have not. The former and latter classes of segregating sites are expressed by internal and one edge elements and by two or three marginal vectors of the 2D SFS matrix, respectively. Recombination affects genealogical relationships, but nevertheless reveals the family size of a mutation in a derived allele or haplotype group. Based on these features and findings, we have developed 2D SFS-based statistics that can have high power, be robust with respect to recombination and nonequilibrium demography, and be used to distinguish not only between hard and soft sweeps, but also between ongoing and past incomplete selective sweeps. However, any sample can lack indispensable recombination signals, thereby leading to erroneous conclusions, particularly when distinguishing the effects of recombination from those of a soft selective sweep. By applying 2D SFS to 100 intronic regions randomly sampled from the EAS population, we have concluded that about 96% of random core SNPs are selectively neutral (Kimura, 1968). In addition, among the six Eurasian genetic loci with known unusual patterns and levels of DNA polymorphism, we have classified two as ongoing hard sweeps, two as ongoing soft sweeps, and two as non-significant. We have also demonstrated that 2D SFS is useful to date an incomplete selective sweep in relation to the TMRCA in a selected allele group.

We thank two anonymous reviewers for their critical reading and useful comments. This work was supported in part by the Japan Society for the Promotion of Science grant 16H04821 to Y. S. and the Scientific Research on Innovative Areas, a MEXT Grant-in-Aid Project FY2016-2020 to N. T.

REFERENCES

1000 Genomes Project Consortium (2015) A global reference for human genetic variation. Nature 526, 68–74.

Akbari, A., Vitti, J. J., Irannehr, A., Bakhtiari, M., Sabeti, P. C., Mirarab, S., and Bafna, V. (2018) Identifying the favored mutation in a positive selective sweep. Nat. Methods 15, 279–282.

Akey, J. M. (2009) Constructing genomic maps of positive selection in humans: where do we go from here? Genome Res. 19, 711–722.

Akey, J. M., Eberle, M. A., Rieder, M. J., Carlson, C. S., Shriver, M. D., Nickerson, D. A., and Kruglyak, L. (2004) Population history and natural selection shape patterns of genetic
variation in 132 genes. PLoS Biol. 2, e286.
Allentoft, M. E., Sikora, M., Sjögren, K.-G., Rasmussen, S., Rasmussen, M., Stenderup, J., Damgaard, P. B., Schroeder, H., Ahlstrøn, T., Vinner, L., et al. (2015) Population genomics of Bronze Age Europe. Nature 522, 167–172.
Ayub, Q., Yngvadottir, B., Chen, Y., Xue, Y., Hu, M., Vernes, S. C., Fisher, S. E., and Tyler-Smith, C. (2013) FOXP2 targets show evidence of positive selection in European populations. Am. J. Hum. Genet. 92, 696–706.
Bae, C. J., Douka, K., and Pettaglia, M. D. (2017) On the origin of modern humans: Asian perspectives. Science 358, eaai9067.
Benjamini, Y., and Hochberg, Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. R. Stat. Soc. Series B Stat. Methodol. 57, 289–300.
Benjamini, Y., Krieger, A. M., and Yekutieli, D. (2006) Adaptive linear step-up procedures that control the false discovery rate. Biometrika 93, 491–507.
Bersaglieri, T., Sabeti, P. C., Patterson, N., Vanderploeg, T., Schaffner, S. F., Drake, J. A., Rhodes, M., Reich, D. E., and Hirschhorn, J. N. (2004) Genetic signatures of strong recent positive selection at the lactase gene. Am. J. Hum. Genet. 74, 1111–1120.
Braverman, J. M., Hudson, R. R., Kaplan, N. L., Langley, C. H., and Stephan, W. (1995) The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics 140, 783–796.
Bustamante, C. D., Wakeley, J., Sawyer, S., and Hartl, D. L. (2001) Directional selection and the site-frequency spectrum. Genetics 159, 1779–1788.
Charlesworth, B., and Charlesworth, D. (2012) Elements of Evolutionary Genetics (2nd edition). Roberts and Company Publishers, Colorado, USA.
Chen, H. (2012) The joint allele frequency spectrum of multiple populations: a coalescent theory approach. Theor. Popul. Biol. 81, 179–195.
Chen, H., Hay, J., and Slatkin, M. (2015) A hidden Markov model for investigating recent positive selection through haplotype structure. Theor. Popul. Biol. 99, 18–30.
Coop, G., Bullaughay, K., Luca, F., and Przeworski, M. (2008) The timing of selection at the human FOXF2 gene. Mol. Biol. Evol. 25, 1257–1259.
Currat, M., Excoffier, L., Maddison, W., Otto, S. P., Ray, N., Whittlock, M. C., and Yeaman, S. (2006) Comment on “Ongoing adaptive evolution in ASPM, a brain size determinant in Homo sapiens” and “Microcephalin, a gene regulating brain size, continues to evolve adaptively in humans”. Science 313, 172.
Enard, W., Przeworski, M., Fisher, S. E., Lai, C. S., Wiebe, V., Kitano, T., Monaco, A. P., and Paiábo, S. (2002) Molecular evolution of FOXF2, a gene involved in speech and language. Nature 418, 869–872.
Enattah, N. S., Sahi, T., Saviähti, E., Terwilliger, J. D., Peltonen, L., and Järvelä, I. (2002) Identification of a variant associated with adult-type hypolactasia. Nat. Genet. 30, 233–237.
Evans, P. D., Gilbert, S. L., Mekel-Bobrov, N., Vallender, E. J., Anderson, J. R., Vaziri-Azizi, L. M., Tishkoff, S. A., Hudson, R. R., and Lahn, B. T. (2005) Microcephalin, a gene regulating brain size, continues to evolve adaptively in humans. Science 309, 1717–1720.
Fan, S., Hansen, M. E., Lo, Y., and Tishkoff, S. A. (2016) Going global by adapting local: a review of recent human adaptation. Science 354, 54–59.
Fay, J. C., and Wu, C.-I. (2000) Hitchhiking under positive Darwinian selection. Genetics 155, 1405–1413.
Ferrer-Admetlla, A., Liang, M., Korneliusen, T., and Nielsen, R. (2014) On detecting incomplete soft or hard sweeps using haplotype structure. Mol. Biol. Evol. 31, 1275–1291.
Ferretti, L., Ledda, A., Wiehe, T., Achaz, G., and Ramos-Onsins, S. E. (2017) Decomposing the site frequency spectrum: the impact of tree topology on neutrality tests. Genetics 207, 229–240.
Field, Y., Boyle, E. A., Nel, N., Gao, Z., Gauton, K. J., Golan, D., Yengo, L., Rocheleau, G., Froguel, P., McCarthy, M. I., et al. (2016) Detection of human adaptation during the past 2000 years. Science 354, 760–764.
Fu, Y.-X. (1995) Statistical properties of segregating sites. Theor. Popul. Biol. 48, 172–197.
Fu, Y.-X., and Li, W.-H. (1993) Statistical tests of neutrality of mutations. Genetics 133, 693–709.
Fujimoto, A., Kimura, R., Ohashi, J., Omi, K., Yuliwulandari, R., Batubara, L., Mustofa, M. S., Samakkam, U., Settheetham-Ishia, W., Ishida, T., et al. (2008) A scan for genetic determinants of human hair morphology: EDA1 is associated with Asian hair thickness. Hum. Mol. Genet. 17, 835–843.
Fujito, N. T., Satta, Y., Hayakawa, T., and Takahata, N. (2018a) A new inference method for ongoing selective sweep. Genes Genet. Syst. 93, 149–161.
Fujito, N. T., Satta, Y., Hane, M., Matsui, A., Yashima, K., Kitajima, K., Sato, C., Takahata, N., and Hayakawa, T. (2018b) Positive selection on schizophrenia-associated ST8SIA2 gene in post-glacial Asia. PLoS One 13, e0200278.
Garud, N. R., Messer, P. W., Buzbas, E. O., and Petrov, D. A. (2015) Recent selective sweeps in North American Drosophila melanogaster show signatures of soft sweeps. PLoS Genet. 11, e1005004.
Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H.-Y., et al. (2010) A draft sequence of the Neandertal genome. Science 328, 710–722.
Griffiths, R. C. (1991) The two-locus ancestral graph. In: Selected Proceedings of the Symposium on Applied Probability. (eds.: Basawa, I. V., and Taylor, R. L.), pp. 100–117. Institute of Mathematical Statistics, Hayward, California, USA.
Griffiths, R. C., and Marjoram, P. (1997) An ancestral recombination graph. In: Progress in Population Genetics and Human Evolution (IMA Volumes in Mathematics and Its Applications, vol. 87) (eds.: Donnelly, P., and Tavaré, S.), pp. 257–270, Springer-Verlag, New York, USA.
Grossman, S. R., Shylakhter, I., Karlsson, E. K., Byrne, E. H., Morales, S., Frieden, G., Hostetter, E., Angelino, E., Garber, M., Zuk, O., et al. (2010) A composite of multiple signals distinguishes causal variants in regions of positive selection. Science 327, 883–886.
Harris, A. M., Garud, N. R., and DeGiorgio, M. (2018) Detection and classification of hard and soft sweeps from unphased genotypes by multilocus genotype identity. Genetics 210, 1429–1452.
Herisson, J., and Pennings, P. S. (2005) Soft sweeps: Molecular population genetics of adaptation from standing genetic variation. Genetics 169, 2335–2352.
Hill, W. G., and Robertson, A. R. (1968) Linkage disequilibrium in finite populations. Theoret. Appl. Genet. 38, 226–231.
Hudson, R. R. (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18, 337–338.
Hudson, R. R. (2007) The variance of coalescent time estimates
from DNA sequences. J. Mol. Evol. 64, 702–705.
Hudson, R. R., and Kaplan, N. L. (1985) Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111, 147–164.
Hunter-Zingck, H., and Clark, A. G. (2015) Arrant time to most recent common ancestor as a signature of natural selection. Mol. Biol. Evol. 32, 2784–2797.
Innan, H., and Kim, Y. (2004) Pattern of polymorphism after strong artificial selection in a domestication event. Proc. Natl. Acad. Sci. USA 101, 10667–10672.
Itan, Y., Powell, A., Beaumont, M. A., Burger, J., and Tomas, M. G. (2009) The origins of lactase persistence in Europe. PLoS Comput. Biol. 5, e1000491.
Kamberov, Y. G., Wang, S., Tan, J., Gerbault, P., Wark, A., Tan, L., Yang, Y., Li, S., Tung, K., Chen, H., et al. (2013) Modeling recent human evolution in mice by expression of a selected EDAR variant. Cell 152, 691–702.
Kaplan, N. L., Hudson, R. D., and Langley, C. H. (1989) The “hitchhiking effect” revisited. Genetics 123, 887–899.
Kelly, J. K. (1997) A test of neutrality based on interlocus associations. Genetics 146, 1197–1206.
Kern, A. D., and Schrider, D. R. (2016) Discoal: flexible coalescent simulations with selection. Bioinformatics 32, 3839–3841.
Kim, Y., and Stephan, W. (2002) Detecting a local signature of genetic hitchhiking along a recombinating chromosome. Genetics 160, 765–777.
Kimura, M. (1968) Evolutionary rate at the molecular level. Nature 217, 624–626.
Kimura, M. (1969) The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics 61, 893–901.
Kimura, R., Yamaguchi, T., Takeda, M., Kondo, O., Toma, T., Hansel, K., Hanihara, T., Matsukusa, H., Kawamura, S., Maki, K., et al. (2009) A common variation in EDAR is a genetic determinant of shovel-shaped incisors. Am. J. Hum. Genet. 85, 528–535.
Klopfstein, S., Currat, M., and Excoffier, L. (2006) The fate of mutations surging on the wave of a range expansion. Mol. Biol. Evol. 23, 482–490.
Kouprina, N., Pavlicev, A., Mochida, G. H., Solomon, G., Gersch, W., Yoon, Y.-H., Collura, R., Ruvolo, M., Barrett, J. C., Woods, C. G., et al. (2004) Accelerated evolution of the ASPM controlling brain size begins prior to human brain expansion. PLoS Biol. 2, e126.
Lapierre, M., Lambert, A., and Achaz, G. (2017) Accuracy of demographic inferences from the site frequency spectrum: the case of the Yoruba population. Genetics 206, 439–449.
Liu, X., and Fu, Y.-X. (2015) Exploring population size changes using SNP frequency spectra. Nat. Genet. 47, 555–559.
Mathieson, I., Lazaridis, I., Rohlhan, N., Mallick, S., Patterson, S., Roodenberg, S. A., Harney, E., Stewartson, K., Fernandez, D., Novak, M., et al. (2015) Genome-wide patterns of selection in 230 ancient Eurasians. Nature 528, 499–503.
Maynard Smith, J., and Haigh, J. (1974) The hitch-hiking effect of a favourable gene. Genet. Res. 23, 23–35.
Mekel-Bobrov, N., Gilbert, S. L., Evans, P. D., Vallender, E. J., Anderson, J. R., Hudson, R. R., Tishkoff, S. A., and Lahn, B. T. (2005) Ongoing adaptive evolution of ASPM, a brain size determinant in Homo sapiens. Science 309, 1720–1722.
Mondal, M., Bertranpetit, J., and Lao, O. (2019) Approximate Bayesian computation with deep learning supports a third archaic introgression in Asia and Oceania. Nat. Commun. 10, 246.
Nei, M., and Li, W.-H. (1979) Mathematical model for studying genetic variation in terms of restriction endonuclease.
ases. Proc. Natl. Acad. Sci. USA 76, 5269–5273.
Nielsen, R., Hellmann, I., Hubisz, M., Bustamante, C., and Clark, A. G. (2007) Recent and ongoing selection in the human genome. Nat. Rev. Genet. 8, 857–868.
Nielsen, R., Williamson, S., Kim, Y., Hubisz, M. J., Clark, A. D., and Bustamante, C. (2005) Genomic scans for selective sweeps using SNP data. Genome Res. 15, 1566–1575.
Orr, H. A., and Betancourt, A. J. (2001) Haldane’s sieve and adaptation from the standing genetic variation. Genetics 157, 875–884.
Pennings, P. S., and Hermisson, J. (2006) Soft sweeps III: the signature of positive selection from recurrent mutation. PLoS Genet. 2, e186.
Peter, B. M., Huerta-Sanchez, E., and Nielsen, R. (2012) Distinguishing between selective sweeps from standing variation and from a de novo mutation. PLoS Genet. 8, e1003011.
Pickrell, J. K., Coop, G., Novembre, J., Kudaravalli, S., Li, J. Z., Absher, D., Jin, L., Drennan, C. S., and另有6篇引文。
of human genome sequence variation. Genome Res. **15**, 1576–1583.
Scheinfeldt, L. B., Biswas, S., Madey, J., Connelly, C. F., Schadt, E. E., and Akey, J. M. (2009) Population genomics analysis of ALMS1 in humans reveals a surprisingly complex evolutionary history. Mol. Biol. Evol. **26**, 1357–1367.
Schrider, D. R., and Kern, A. D. (2016) S/HIC: robust identification of soft and hard sweeps using machine learning. PLoS Genet. **12**, e1005928.
Schrider, D. R., and Kern, A. D. (2017) Soft sweeps are the dominant mode of adaptation in the human genome. Mol. Biol. Evol. **34**, 1863–1877.
Sheehan, S., and Song, Y. S. (2016) Deep learning for population genetic inference. PLoS Comput. Biol. **12**, e1004845.
Simonsen, K. L., Churchill, G. A., and Aquadro, C. F. (1995) Properties of statistical tests of neutrality for DNA polymorphism data. Genetics **141**, 413–429.
Slatkin, M., and Rannala, B. (1997) Estimating the age of alleles by use of intraallelic variability. Am. J. Hum. Genet. **60**, 447–458.
Slatkin, M., and Rannala, B. (2000) Estimating allele age. Annu. Rev. Genomics Hum. Genet. **1**, 225–249.
Stephan, W., Wiehe, T. H. E., and Lenz, M. W. (1992) The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. Theor. Popul. Biol. **41**, 237–254.
Storey, J. D., and Tibshirani, R. (2003) Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. USA **100**, 9440–9445.
Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123**, 585–595.
Takahata, N., and Nei, M. (1985) Gene genealogy and variance of interpopulational nucleotide differences. Genetics **110**, 325–344.
Tang, H., Siegmund, D. O., Shen, P., Oefner, P. J., and Feldman, M. W. (2002) Frequentist estimation of coalescence times from nucleotide sequence data using a tree-based partition. Genetics **161**, 447–459.
Tavaré, S. (1984) Line-of-descent and genealogical processes, and their applications in population genetics models. Theor. Popul. Biol. **26**, 119–164.
Terhorst, J., Kamm, J. A., and Song, Y. S. (2017) Robust and scalable inference of population history from hundreds of unphased whole-genomes. Nat. Genet. **49**, 303–309.
Teshima, K. M., Coop, G., and Przeworski, M. (2006) How reliable are empirical genomic scans for selective sweeps? Genome Res. **16**, 702–712.
Thomson, R., Pritchard, J. K., Shen, P., Oefner, P. J., and Feldman, M. W. (2000) Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. Proc. Natl. Acad. Sci. USA **97**, 7360–7365.
Tishkoff, S. A., Reed, F. A., Ranciaro, A., Voight, B. F., Babbitt, C. C., Silverman, J. S., Powell, K., Mortensen, H. M., Hirbo, J. B., Osman, M., et al. (2007) Convergent adaptation of human lactase persistence in Africa and Europe. Nat. Genet. **39**, 31–40.
Tournebize, R., Poncet, V., Jakobsson, M., Vigouroux, Y., and Manel, S. (2018) McSwa: a joint site frequency spectrum method to detect and date selective sweep across multiple population genomes. Mol. Ecol. Resour. **19**, 283–295.
Vitti, J. J., Grossman, S. R., and Sabeti, P. C. (2013) Detecting natural selection in genomic data. Annu. Rev. Genet. **47**, 97–120.
Wakeley, J. (2009) Coalescent Theory: An Introduction. Roberts and Company Publishers, Greenwood Village, Colorado, USA.
Watterson, G. A. (1975) On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7**, 256–276.
Wilde, S., Timpson, A., Kirsanow, K., Kaiser, E., Kayser, M., Unterländer, M., Höffelder, N., Potekhina, I. D., Thomas, M. G., et al. (2014) Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y. Proc. Natl. Acad. Sci. USA **111**, 4832–4837.
Williamson, S. H., Hubisz, M. J., Clark, A. G., Payseur, B. A., Bustamante, C. D., and Nielsen, R. (2007) Localizing recent adaptive evolution in the human genome. PLoS Genet. **3**, e90.
Wilson, B. A., Petrov, D. A., and Messer, P. W. (2014) Soft selective sweeps in complex demographic scenarios. Genetics **198**, 669–684.
Yang, Z., Li, J., Wiehe, T., and Li, H. (2018) Detecting recent positive selection with a single locus test bipartitioning the coalescent tree. Genetics **208**, 791–805.
Yu, F., Hill, R. S., Schaffner, S. F., Sabeti, P. C., Wang, E. T., Mignault, A. A., Ferland, R. J., Mozyk, R. K., Walsh, C. A., and Reich, D. (2007) Comment on “Ongoing adaptive evolution of ASPM, a brain size determinant in Homo sapiens.” Science **316**, 370.
Zeng, K., Su, Y.-X., Shi, S., and Wu, C.-I. (2006) Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics **174**, 1431–1439.
Zeng, K., Shi, S., and Wu, C.-I. (2007) Compound tests for the detection of hitchhiking under positive selection. Mol. Biol. Evol. **24**, 1898–1908.
Supplementary Table S1. Simulated 18 summary statistics with ongoing hard sweeps when \( f_r = 0.3, 0.6 \) and \( 0.9 \) and \( n = 1,000 \)

| \( f_r = 0.3 \) | \( \rho = 0 \) | \( \rho = 20 \) | \( \rho = 50 \) |
|-----------------|----------------|----------------|----------------|
| \( F_c \)       | 0.012 ± 0.015  | 0.018 ± 0.016  | 0.031 ± 0.024  |
| \( S_\xi \)     | 148 ± 28.4     | 147 ± 17.6     | 147 ± 15.9     |
| \( S_{\xi_0} \) | 6.69 ± 2.65    | 6.63 ± 2.60    | 6.65 ± 2.70    |
| \( S_{\xi_1} \) | 0              | 18.4 ± 10.7    | 33.0 ± 10.6    |
| \( L_\xi \)     | 20415 ± 11539  | 20117 ± 5856   | 20375 ± 4680   |
| \( L_{\xi_0} \) | 38.0 ± 50.6    | 46.6 ± 77.6    | 38.5 ± 65.3    |
| \( L_{\xi_1} \) | 0              | 2388 ± 1540    | 3823 ± 1446    |
| \( G_{c_0} \)   | 5.56 ± 7.06    | 6.99 ± 12.6    | 5.55 ± 9.05    |
| \( G_{c_1} \)   | 0              | 129 ± 47.4     | 116 ± 28       |
| \( G_c \)       | 5.56 ± 7.06    | 91.1 ± 34.0    | 96.9 ± 24.5    |
| \( \theta_w \)  | 1.07 ± 0.42    | 3.99 ± 1.79    | 6.32 ± 1.75    |
| \( \theta_\pi \)| 0.20 ± 0.19    | 1.02 ± 1.26    | 2.24 ± 1.69    |
| \( \theta_L \)  | 0.13 ± 0.17    | 8.16 ± 5.17    | 12.9 ± 4.82    |
| \( \theta_H \)  | 0.05 ± 0.19    | 15.3 ± 9.86    | 23.6 ± 9.09    |
| \( D_\xi \)     | -1.58 ± 0.36   | -1.96 ± 0.61   | -1.82 ± 0.63   |
| \( H_\xi \)     | 0.28 ± 0.30    | -9.34 ± 4.20   | -10.0 ± 3.24   |
| \( E_\xi \)     | -1.32 ± 0.29   | 2.04 ± 1.55    | 2.47 ± 1.22    |
| \( K_\xi \)     | 0.90 ± 0.21    | -3.58 ± 2.00   | -4.05 ± 1.54   |
| \( f_r = 0.6 \) | \( p = 0 \) | \( p = 20 \) | \( p = 50 \) |
|---|---|---|---|
| \( F_c \) | 0.025 ± 0.029 | 0.064 ± 0.063 | 0.117 ± 0.084 |
| \( S_\xi \) | 139 ± 27.5 | 139 ± 18.1 | 140 ± 16.2 |
| \( S_{\xi 0} \) | 14.6 ± 3.80 | 14.8 ± 3.80 | 14.8 ± 4.00 |
| \( S_{\xi 1} \) | 0 | 27.3 ± 11.9 | 43.0 ± 11.6 |
| \( L_\xi \) | 20539 ± 10761 | 20593 ± 6050 | 20667 ± 5086 |
| \( L_{\xi 0} \) | 81.6 ± 86.0 | 114 ± 120 | 89.3 ± 129 |
| \( L_{\xi 1} \) | 0 | 6648 ± 3343 | 8980 ± 3086 |
| \( G_{c 0} \) | 5.59 ± 5.96 | 7.42 ± 12.2 | 6.10 ± 9.53 |
| \( G_{c 1} \) | 0 | 243 ± 69.0 | 209 ± 48.0 |
| \( G_c \) | 5.59 ± 5.96 | 154 ± 49.5 | 156 ± 38.2 |
| \( \theta_w \) | 2.10 ± 0.54 | 6.04 ± 1.82 | 8.29 ± 1.76 |
| \( \theta_\pi \) | 0.23 ± 0.17 | 1.26 ± 1.30 | 2.59 ± 1.73 |
| \( \theta_L \) | 0.14 ± 0.14 | 11.3 ± 5.60 | 15.1 ± 5.10 |
| \( \theta_H \) | 0.04 ± 0.15 | 21.3 ± 10.8 | 27.7 ± 9.70 |
| \( D_\xi \) | −2.01 ± 0.21 | −2.13 ± 0.47 | −1.91 ± 0.50 |
| \( H_\xi \) | 0.23 ± 0.17 | −9.43 ± 3.53 | −9.17 ± 2.70 |
| \( E_\xi \) | −1.70 ± 0.18 | 1.83 ± 1.32 | 1.93 ± 1.05 |
| \( K_\xi \) | 1.14 ± 0.12 | −3.48 ± 1.69 | −3.50 ± 1.33 |
\[ f_r = 0.9 \]

|       | \( \rho = 0 \)       | \( \rho = 20 \)       | \( \rho = 50 \)       |
|-------|---------------------|---------------------|---------------------|
| \( F_c \) | 0.152 ± 0.098       | 0.269 ± 0.128       | 0.405 ± 0.138       |
| \( S_\xi \) | 120 ± 29.2          | 122 ± 17.2          | 122 ± 15.0          |
| \( S_{\xi 0} \) | 28.2 ± 5.50         | 29.2 ± 5.39         | 30.2 ± 5.45         |
| \( S_{\xi 1} \) | 0                   | 27.8 ± 11.8         | 41.7 ± 10.9         |
| \( L_\xi \) | 20691 ± 12572       | 20795 ± 6335        | 20560 ± 5435        |
| \( L_{\xi 0} \) | 159 ± 127           | 363 ± 629           | 263 ± 365           |
| \( L_{\xi 1} \) | 0                   | 10149 ± 4791        | 13328 ± 4512        |
| \( G_{c0} \) | 5.63 ± 4.44         | 12.3 ± 20.4         | 8.78 ± 12.3         |
| \( G_{c1} \) | 0                   | 368 ± 100           | 319 ± 69.3          |
| \( G_c \) | 5.63 ± 4.44         | 179 ± 58.3          | 187 ± 47.0          |

|       | \( \theta_w \) | \( \theta_\pi \) | \( \theta_L \) | \( \theta_H \) | \( D_\xi \) | \( H_\xi \) | \( E_\xi \) | \( K_\xi \) |
|-------|----------------|----------------|-------------|-------------|----------|--------|-------|--------|
|       | 3.82 ± 0.75   | 0.31 ± 0.18    | 0.17 ± 0.15 | 0.04 ± 0.16 | -2.27 ± 0.13 | 0.20 ± 0.12 | -1.97 ± 0.11 | 1.31 ± 0.07 |
|       | 7.73 ± 1.80   | 1.31 ± 1.10    | 11.7 ± 5.38 | 22.1 ± 10.4 | -2.24 ± 0.32 | -7.88 ± 2.93 | 1.09 ± 1.11  | -2.63 ± 1.42 |
|       | 9.74 ± 1.64   | 2.69 ± 1.67    | 15.1 ± 5.04 | 27.5 ± 9.62 | -2.00 ± 0.42 | -7.79 ± 2.44 | 1.27 ± 0.92  | -2.74 ± 1.17 |

Command line: `./discoal2 1000 1000 16000 -t 20 -ws 0.5 -x 0.30 -a 500 -r \( \rho \).`
**Supplementary Table S2.** Simulated summary statistics with ongoing soft sweeps under the *discoal* command line for $n = 1,000$, $f_r = 0.3, 0.6$ and 0.9, $f_0 = 0.1$, $a = 500$, $\theta = 20$ and 1,000 replications

| $f_r$ = 0.3 | $\rho = 0$ | $\rho = 20$ | $\rho = 50$ |
|------------|------------|------------|------------|
| $F_c$      | 0.106 ± 0.100 | 0.148 ± 0.098 | 0.184 ± 0.089 |
| $S_\xi$    | 151 ± 28.2 | 150 ± 18.4 | 151 ± 15.6 |
| $S_{\xi_0}$ | 13.7 ± 6.02 | 13.2 ± 4.77 | 11.7 ± 3.99 |
| $S_{\xi_1}$ | 0 | 32.0 ± 12.9 | 47.1 ± 12.0 |
| $L_\xi$     | 21203 ± 11294 | 20802 ± 5720 | 20711 ± 4571 |
| $L_{\xi_0}$ | 483 ± 569 | 449 ± 459 | 244 ± 228 |
| $L_{\xi_1}$ | 0 | 3605 ± 1626 | 4685 ± 1438 |
| $G_{c0}$    | 30.6 ± 21.0 | 31.2 ± 25.1 | 20.0 ± 16.4 |
| $G_{c1}$    | 0 | 113 ± 27.9 | 99.7 ± 18.6 |
| $G_c$       | 30.6 ± 21.0 | 88.0 ± 21.1 | 83.7 ± 16.1 |

| $\theta_w$  | 2.19 ± 0.96 | 7.20 ± 2.36 | 9.36 ± 2.12 |
| $\theta_{\pi}$ | 1.72 ± 1.60 | 7.14 ± 4.61 | 10.5 ± 4.51 |
| $\theta_L$  | 1.61 ± 1.87 | 13.6 ± 5.82 | 16.5 ± 4.83 |
| $\theta_H$  | 1.50 ± 2.40 | 20.0 ± 9.16 | 22.5 ± 7.63 |
| $D_\xi$     | -0.63 ± 0.88 | -0.15 ± 1.20 | 0.29 ± 0.99 |
| $H_\xi$     | 0.35 ± 1.30 | -5.417 ± 3.36 | -4.08 ± 2.72 |
| $E_\xi$     | -0.63 ± 0.83 | 2.07 ± 1.06 | 1.91 ± 0.82 |
| $K_\xi$     | 0.48 ± 0.74 | -2.67 ± 1.37 | -2.22 ± 1.09 |
\[ f_r = 0.6 \]

| \( f_r = 0.6 \) | \( \rho = 0 \) | \( \rho = 20 \) | \( \rho = 50 \) |
|----------------|----------------|----------------|----------------|
| \( F_c \)     | 0.140 ± 0.113  | 0.327 ± 0.156  | 0.427 ± 0.121  |
| \( S_\xi \)   | 146 ± 30.1     | 144 ± 18.6     | 144 ± 15.2     |
| \( S_{\xi0} \)| 22.1 ± 6.79    | 21.5 ± 5.47    | 20.2 ± 5.05    |
| \( S_{\xi1} \)| 0              | 36.8 ± 13.3    | 51.9 ± 11.3    |
| \( L_\xi \)   | 22040 ± 11963  | 20797 ± 5929   | 20650 ± 4682   |
| \( L_{\xi0} \)| 1093 ± 1212    | 916 ± 889      | 497 ± 471      |
| \( L_{\xi1} \)| 0              | 7851 ± 3272    | 9682 ± 2825    |
| \( G_{c0} \)  | 44.0 ± 33.9    | 40.0 ± 34.1    | 23.8 ± 20.8    |
| \( G_{c1} \)  | 0              | 214 ± 49.7     | 186 ± 34.3     |
| \( G_c \)     | 44.0 ± 33.9    | 147 ± 35.9     | 140 ± 27.7     |

\[ \theta_w \]
\[ \theta_\pi \]
\[ \theta_L \]
\[ \theta_H \]
\[ D_\xi \]
\[ H_\xi \]
\[ E_\xi \]
\[ K_\xi \]

\[ \theta_w \] | 3.17 ± 0.97 | 8.37 ± 2.18 | 10.3 ± 1.83 |
| \( \theta_\pi \) | 1.91 ± 1.74 | 7.22 ± 4.59 | 10.5 ± 4.26 |
| \( \theta_L \)  | 1.82 ± 2.02 | 14.6 ± 5.80 | 17.0 ± 4.76 |
| \( \theta_H \)  | 1.73 ± 2.56 | 22.0 ± 9.68 | 23.5 ± 8.03 |
| \( D_\xi \)    | -1.07 ± 0.82 | -0.47 ± 1.12 | -0.004 ± 0.92 |
| \( H_\xi \)    | 0.22 ± 1.11 | -5.41 ± 3.21 | -3.96 ± 2.52 |
| \( E_\xi \)    | -0.96 ± 0.80 | 1.71 ± 0.99 | 1.55 ± 0.78 |
| \( K_\xi \)    | 0.66 ± 0.69 | -2.42 ± 1.30 | -1.97 ± 1.01 |
| $f_r = 0.9$ | $\rho = 0$ | $\rho = 20$ | $\rho = 50$ |
|------------|-------------|-------------|-------------|
| $F_c$      | $0.375 \pm 0.187$ | $0.633 \pm 0.164$ | $0.763 \pm 0.102$ |
| $S_\xi$    | $128 \pm 30.0$ | $129 \pm 18.0$ | $129 \pm 15.1$ |
| $S_{\xi_0}$ | $35.1 \pm 7.45$ | $36.6 \pm 7.1$ | $36.3 \pm 6.79$ |
| $S_{\xi_1}$ | $0$ | $35.7 \pm 12.7$ | $50.2 \pm 11.1$ |
| $L_\xi$    | $21612 \pm 12373$ | $21103 \pm 6138$ | $20934 \pm 5133$ |
| $L_{\xi_0}$ | $1500 \pm 1607$ | $1929 \pm 1733$ | $1213 \pm 971$ |
| $L_{\xi_1}$ | $0$ | $11507 \pm 4920$ | $14380 \pm 4268$ |
| $G_{c_0}$  | $39.6 \pm 33.8$ | $50.6 \pm 41.5$ | $32.6 \pm 24.4$ |
| $G_{c_1}$  | $0$ | $321 \pm 72.7$ | $286 \pm 52.6$ |
| $G_c$      | $39.6 \pm 33.8$ | $182 \pm 46.4$ | $179 \pm 36.0$ |

| $\theta_w$ | $4.76 \pm 1.01$ | $9.79 \pm 2.09$ | $11.7 \pm 1.86$ |
| $\theta_\pi$ | $1.83 \pm 1.56$ | $7.01 \pm 4.47$ | $10.7 \pm 4.37$ |
| $\theta_L$ | $1.67 \pm 1.79$ | $14.9 \pm 5.85$ | $17.3 \pm 4.92$ |
| $\theta_H$ | $1.51 \pm 2.33$ | $22.9 \pm 9.73$ | $24.0 \pm 8.06$ |
| $D_\xi$    | $-1.62 \pm 0.60$ | $-0.85 \pm 0.96$ | $-0.29 \pm 0.82$ |
| $H_\xi$    | $0.21 \pm 0.84$ | $-4.96 \pm 2.72$ | $-3.59 \pm 2.22$ |
| $E_\xi$    | $-1.44 \pm 0.59$ | $1.17 \pm 0.91$ | $1.13 \pm 0.72$ |
| $K_\xi$    | $0.98 \pm 0.52$ | $-1.97 \pm 1.14$ | $-1.62 \pm 0.90$ |
### Supplementary Table S3

**rs numbers of 145 SNP sites randomly chosen from the intronic regions in EAS in the 1000 GD**

| rs number   | rs number   | rs number   | rs number   |
|-------------|-------------|-------------|-------------|
| rs287103    | rs236316    | rs580057    | rs803922    |
| rs1540932   | rs238154    | rs580909    | rs852976    |
| *rs1484214  | *rs241546   | *rs599988   | *rs857845   |
| rs1947274   | rs268601    | *rs618369   | *rs877268   |
| rs2251954   | *rs301506   | *rs621196   | *rs885603   |
| rs2435343   | rs333292    | *rs653465   | *rs891698   |
| rs2868177   | rs336296    | rs658965    | rs919823    |
| rs3765069   | rs337445    | rs660214    | *rs935746   |
| rs6075382   | rs347112    | *rs683023   | *rs941503   |
| rs6464926   | *rs349536   | rs683741    | *rs942878   |
| rs8026435   | *rs350261   | rs685472    | *rs954124   |
| *rs10246477 | rs857845    | rs690176    | *rs960270   |
| rs11042714  | rs380512    | *rs691656   | *rs969876   |
| rs12310399  | rs391470    | rs710983    | *rs985758   |
| rs12753665  | rs395897    | rs717244    | *rs999431   |
| *rs12972621 | rs403732    | rs721053    | *rs100205   |
| *rs33428    | rs409636    | rs730097    | rs1001460   |
| rs33685     | rs425036    | rs738372    | rs1150048   |
| rs130145    | rs439643    | rs740067    | rs1180020   |
| rs161020    | *rs456026   | rs740306    | rs1214557   |
| rs171131    | rs483894    | rs743870    | rs1217631   |
| rs176961    | rs517412    | rs759495    | rs1227243   |
| rs217207    | rs517738    | rs772152    | *rs1230313  |
| rs223573    | rs532097    | rs778444    | rs1283686   |
| rs231602    | rs537145    | rs786921    | *rs1373806  |

There are 2,985 SNPs in introns with Global MAF ranging from 0.399 to 0.400. Shown here are 145 biallelic SNPs validated by the 1000GD. These SNPs were filtered by \( f_r = 0.25\)~0.75 in EAS. The 45 rs numbers with asterisks show a rapid decay of \( r^2 \) values, and do not show any LD signal.
of a selective sweep or hitchhiking effects: they were not included in further analysis.
Hard sweep without recombination ($\rho = 0$)

Soft sweep without recombination ($\rho = 0$)

Supplementary Fig. S1.
Simulated 2D SFS under hard and soft sweeps with and without recombination ($n = 68$).
Supplementary Fig. S1.
Supplementary Fig. S2.
Compared with the case of neutrality (red line) in a nonequilibrium population model (SFD), the C value under positive selection (blue line) is almost independent of $\rho = 4Nrl$: $\rho$ - insensitivity of $T_D = L\zeta_0/m$ under (a) a hard sweep with $a = 2N_e s = 350$ ($s$: genic selection intensity) in contrast to the sensitivity under (b) neutrality. The command lines for (a) and (b) are, respectively: (a) ./discoal2 1008 1000 100000 -t 67 -r $\rho$ -ws 0 -a 350 -x 0.5 -c 0.60 -en 0.001 0 0.077 -en 0.004745 0 0.007 -en 0.004995 0 0.077 -en 0.0084975 0 0.006 -en 0.0087475 0 0.24 -en 0.0425 0 0.125, and (b) ./ms 1008 5000 -s 89 -r $\rho$ 100000 -eN 0.001 0.077 -eN 0.004745 0.007 -eN 0.004995 0.077 -eN 0.0084975 0.006 -eN 0.0087475 0.24 -eN 0.0425 0.125
Supplementary Fig. S3.
2D SFS under pure ($a = 0$) and near ($a = 1$) neutral models simulated by $ms$ (Hudson, 2002) and $discoal$ (Kern and Schrider, 2016), respectively.
Supplementary Fig. S4.
Probability $g_{n,k}(2\tau)$ of the number ($k$) of distinct ancestors $2\tau$ units of time ago in a sample of size $n$.
Formulae (7a) – (7c) in Takahata and Nei (1985) for the probability (ordinate) of the number of distinct ancestors (abscissa) $t = 0.05$ units of time ago for a sample of size 100 (left), 200 (middle) and 600 (right).
Supplementary Fig. S5.
NJ trees for a sample that underwent neutral drift before a selective sweep was completed.
Supplementary Fig. S6.
Distributions of $F_c$, $L_{c0}$, $G_{c0}$ and $D_\zeta$ in the presence and absence of recombination.
Supplementary Fig. S7.
ρ-dependency of \(r^2\) values under (a) neutrality and (b) hard selective sweeps.
(b) Hard sweep

Supplementary Fig. S7
Supplementary Fig. S8. Average $r^2$ values in 1-kb non-overlapping windows for six core regions. An arrow indicates the core SNP in each region. Dotted lines show 5' and 3' boundaries of the core region. A square bracket marks the location of each core region defined by these boundaries.
**ASPM** (rs41310927: at 197070697)

![Graph showing SNP data for ASPM](image)

165 kb (196900000-197150000)

**MCPH1** (rs930557: at 6302183)

![Graph showing SNP data for MCPH1](image)

32 kb (6286532-6318958)

Supplementary Fig. S8
**ADAM17** (rs2709591: at 9496291)

![ADAM17 plot](image)

26 kb (9490015-9515668)

**NRG1** (rs3924999: at 32453358)

![NRG1 plot](image)

3 kb (32451286-32454158)