A new inference method for detecting an ongoing selective sweep

Naoko T. Fujito¹, Yoko Satta^{1*}, Toshiyuki Hayakawa^{2,3} and Naoyuki Takahata¹

¹*School of Advanced Sciences, SOKENDAI (The Graduate University for Advanced Studies), Shonan Village, Hayama, Kanagawa 240-0193, Japan*

²*Graduate School of Systems Life Sciences and* ³*Faculty of Arts and Science, Kyushu University, 744 Motooka, Nishi-ku, Fukuoka 819-0395, Japan*

(Received 1 March 2018, accepted 11 June 2018; J-STAGE Advance published date: 30 September 2018)

A simple method was developed to detect signatures of ongoing selective sweeps in single nucleotide polymorphism (SNP) data. Based largely on the traditional site frequency spectrum (SFS), the method additionally incorporates linkage disequilibrium (LD) between pairs of SNP sites and uniquely represents both SFS and LD information as hierarchical “barcodes.” This barcode representation allows the identification of a hitchhiking genomic region surrounding a putative target site of positive selection, or a core site. Sweep signals at linked neutral sites are then measured by the proportion (F_c) of derived alleles within the hitchhiking region that are linked in the derived allele group defined at the core site. In measuring F_c or intra-allelic variability in an informative way, certain conditions for derived allele frequencies are required, as illustrated with the human *ST8SIA2* locus. Coalescent simulators with and without positive selection are used to assess the false-positive and false-negative rates of the F_c statistic. To demonstrate its power, the method was further applied to the *LCT*, *OCA2*, *EDAR*, *SLC24A5* and *ASPM* loci, which are known to have undergone positive selection in human populations. Overall, the method is powerful and can be used to identify core sites responsible for ongoing selective sweeps.

Key words: hitchhiking, human evolution, linkage disequilibrium, population genomics, site frequency spectrum

INTRODUCTION

Population genomics has identified a large number of genetic variants that are associated with either global or local adaptations of species, in particular of anatomically modern humans (e.g., Li et al., 2014; Schrider and Kern, 2017). As reviewed in Lachance and Tishkoff (2013), Vitti et al. (2013) and Fan et al. (2016), there are several approaches for detecting positive Darwinian selection from DNA polymorphisms. One approach is based on the allele frequency distribution or the site frequency spectrum (SFS) and uses either summary statistics (e.g., Watterson, 1975; Tajima, 1989; Fu and Li, 1993; Fay and Wu, 2000; Zeng et al., 2006; Field et al., 2016) or the entire allele frequency distribution (e.g., Sawyer and Hartl, 1992; Bustamante et al., 2001; Nielsen et al., 2005; Gutenkunst et al., 2009; Pavlidis et al., 2013). However, when positive selection is ongoing, the allele frequency at

a selected site may not be sufficiently high, which naturally results in reduced statistical power of the SFS-based methods. It is also known that the inference of positive selection made by these SFS-based methods tends to be muddled by the confounding effect of demography if they attempt to detect deviations from the standard neutral model of constant population size. In addition, these methods often assume no recombination within a locus but free recombination between loci or their loose linkage, and are extremely sensitive to these assumptions on recombination as linkage inflates the variance of allele frequencies (Bustamante et al., 2001).

Li (2011) proposed an alternative method that uses the maximum frequency of derived alleles in a sample as an indicator of unbalanced coalescent trees. This method relies on the fact that the probability of unbalanced basal branches is extremely low under neutrality and is independent of temporal changes in population size (see also Yang et al., 2018 for an extension to a varying size population, and Ferretti et al., 2017, who provided a systematic analysis of the impact of the waiting times and the branching order of coalescent events on the SFS). One

Edited by Ryosuke Kimura

* Corresponding author. E-mail: satta@soken.ac.jp

DOI: <http://doi.org/10.1266/ggs.18-00008>

challenge, however, arises from the fact that the required imbalance must be strong in a large sample. In addition, the power of the method is naturally weakened when a selective sweep is partial and incomplete.

The long-range haplotype (LRH) test is also commonly used for detecting sweep signals (Sabeti et al., 2002; Vitti et al., 2013). This approach relies on undisrupted long-range associations with neighboring polymorphisms, or the relationship between an allele's frequency at a focal site and the extent of linkage disequilibrium (LD) surrounding it (Sabeti et al., 2002), and may be classified into "haplotype" tests in contrast to SFS tests (Zeng et al., 2006). However, it appears that LRH-based methods tend to miss such signals that are localized in small genomic regions due to nearby recombination hotspots.

In the present paper, we develop a new method that permits the detection of ongoing selective sweeps. Although a sweep can be either "hard" (Maynard Smith and Haigh, 1974; Kaplan et al., 1989) or "soft" (Innan and Kim, 2004; Hermisson and Pennings, 2005; Przeworski et al., 2005), we here presume a soft sweep for generality as it can accommodate the theoretical framework to the case of a single variant on which positive selection starts to act. In particular, we are concerned with the situation under which the size of two basal branches in genealogy may not be particularly distorted, but wherein one descendant allelic lineage has expanded rapidly. We demonstrate by simulation that such lineage-specific expansion is seldom observed under the standard neutral model (a low false-positive rate) as well as even under certain demographic models of neutral variants subjected to population bottleneck and expansion. Although we do not address the reverse problem that positive selection may severely bias estimates of demographic parameters (for which readers may refer to Schrider et al., 2016), we examine whether our method can equally exhibit a low false-negative rate when positive selection in fact operates. For this purpose, we conduct coalescent simulation with selection and further apply our method to well-known cases of selective sweeps.

AN INFERENCE METHOD

We consider S bi-allelic single nucleotide polymorphism (SNP) sites in a sample of n homologous chromosomes randomly taken from a population. For simplicity, these SNP sites are numbered from the left end of a genomic region under study. To summarize such data, it is convenient to define an indicator matrix $\{\chi_{kl}\}$, the (k, l) element of which takes the value of 1 if the k -th SNP site of the l -th chromosome ($1 \leq k \leq S$ and $1 \leq l \leq n$) possesses a derived allele, otherwise being assigned as 0 (Table 1). The number of derived alleles at the k -th SNP site in a sample is then counted simply as

$$n_k = \sum_{l=1}^n \chi_{kl}. \quad (1)$$

If $n_k = 1$, the site k is a singleton (or the site has only one derived allele segregating in a sample) and the total number of such sites across the sampled genomic region is denoted by ξ_1 . Likewise, if $n_k = 2$, the site k is a doubleton (or the site has two segregating derived alleles) and the total number of such sites is denoted by ξ_2 . Here we follow Fu (1995), who used the symbol ξ_i ($1 \leq i \leq n-1$) to express the random number of SNP sites with each exhibiting i derived alleles segregating in a sample: ξ_i = the number of SNP sites at each of which $n_k = i$ for $1 \leq k \leq S$. By definition, the total number of segregating sites is equal to the assumed number of SNP sites and expressed as $S = \sum_{i=1}^{n-1} \xi_i$ and the total number of derived alleles is expressed as $L = \sum_{i=1}^{n-1} i \xi_i$ where $\xi'_i = i \xi_i$ is the scaled SFS (Fu, 1995; Ronen et al., 2013). Under the standard neutral model of constant effective population size N_e , the mean of ξ_i is given by $E\{\xi_i\} = \theta/i$ for all possible values of i (Sawyer and Hartl, 1992; Charlesworth and Charlesworth, 2012) and the mean of ξ'_i by $E\{\xi'_i\} = \theta$ where $\theta = 4N_e u$ and u is the mutation rate per region per generation. The means of S and L therefore reduce to $E\{S\} = \theta \sum_{i=1}^{n-1} 1/i = \theta a_n$ and $E\{L\} = \theta(n-1)$ (Watterson, 1975; Zeng et al., 2006). Moreover, we have $E\{\Pi\} = \theta$ for the pairwise mean of nucleotide differences, defined as $\Pi = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} i(n-i) \xi_i$ (Nei and Li, 1979; Tajima, 1989) and $E\{H\} = \theta$ for $H = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} i^2 \xi_i$ (Fay and Wu,

Table 1. Example of the site frequency spectrum $\{\xi_i\} = (\xi_1 = 2, \xi_2 = 3, \xi_3 = 2, \xi_4 = 2, \xi_5 = 0)$ with the total number of segregating sites $S = 9$ and the sample size $n = 6$.

	SNP sites								
	core								
	1	2	3	4	5	6	7	8	9
chr 1	1	0	1	1	0	1	1	1	1
chr 2	0	0	1	0	0	1	1	1	1
chr 3	0	0	0	0	0	1	1	1	0
chr 4	0	1	0	1	1	0	0	1	1
chr 5	0	0	0	0	1	0	0	0	1
chr 6	0	0	0	0	0	0	0	0	0
n_k	1	1	2	2	2	3	3	4	4
n_{rk}	1	0	2	1	0	3	3	3	2
g_k	3	4	3	4	5	3	3	4	5

The core SNP site (r) is arbitrarily set at $r = 6$ with $n_r = 3$. The values of n_k and n_{rk} for $1 \leq k \leq 9$ determine the barcodes at these nine SNP sites, retaining their spatial positions. The conditions for recombinants are $n_{rk} > 0$, $n_k > n_{rk}$, $n_r > n_{rk}$ and $n > g_k = n_k + n_r - n_{rk}$ as in Eq. (3a) in the text.

2000; Ferretti et al., 2017). Notably, these mean values are all independent of recombination rates, although the variances are not.

In a large sample, it is likely that $\xi_i = 0$ for many large i values. For this reason, we bin individual SNP sites into the following eight classes, although the number of such frequency classes depends on sample size n . Because of its preponderance, the first class (c_1) consists solely of singletons and is expected to have the mean of $E\{\xi_1\} = \theta$ (Fu and Li, 1993). The second class (c_2) consists of doubletons and tripletons with $E\{\xi_2 + \xi_3\} = 5\theta/6$. Likewise, the remaining classes are c_3 with $E\{\sum_{i=4}^9 \xi_i\} \approx \theta$, c_4 with $E\{\sum_{i=10}^{25} \xi_i\} \approx \theta$, c_5 with $E\{\sum_{i=26}^{68} \xi_i\} \approx \theta$, c_6 with $E\{\sum_{i=69}^{185} \xi_i\} \approx \theta$, c_7 with $E\{\sum_{i=186}^{503} \xi_i\} \approx \theta$, and c_8 with $E\{\sum_{i=504}^{1007} \xi_i\} = 0.69\theta$ if $n = 1008$ as for the East Asia meta-population in the 1000 Genomes Project database (1000 Genomes Project Consortium, 2015). However, if $n \geq 1369$ and ≥ 3720 , we additionally need classes up to c_9 and c_{10} , respectively, whereas if $n = 68$, we need only the first five classes. In any case, we sort each individual SNP site k or SFS into these classes with nearly equal weight. This strategy has been previously utilized, for example, in estimating the distribution of fitness effects for amino acid-altering mutations (Eyre-Walker et al., 2006). In our method, however, this hierarchical binning of SFS plays an essential role in extracting information on “intra-allelic variability” (Slatkin and Rannala, 1997). In this context, it is instructive to recall that the higher the allele frequency, the older the mean age of an allele (Kimura and Ohta, 1973; Slatkin and Rannala, 2000).

In human populations, the expected rough uniformity of SFS over the eight classes is often severely violated. The SFS generally reveals an excess of rare as well as intermediate or high allele frequencies, thereby supporting demographic models of recent population growth and past bottlenecks, in particular in non-African populations (e.g., Schaffner et al., 2005; Gutenkunst et al., 2009; Liu and Fu, 2015; Terhorst et al., 2017). Alternatively, positive selection may leave similar signatures at linked neutral sites (e.g., Kaplan et al., 1989; Braverman et al., 1995; Hermisson and Pennings, 2005; Evans et al., 2006; Ronen et al., 2013; Schrider and Kern, 2017). In the following analysis, we inquire whether and how we can isolate the effect of positive selection from that of demographic causes.

We pay special attention to one particular SNP site, henceforth termed a core site or the r -th SNP site, which might be a target of positive selection. This core site is assumed to be given or of interest, and does not require any *a priori* knowledge about its evolutionary aspects. Our aim is to examine whether there exists a genomic region hitchhiking and in linkage disequilibrium (LD) with derived alleles at a core site. As a rule, the

tighter the linkage, the stronger the hitchhiking effect; also, the broader the genomic region, the more recent the positive selection. There are $n_r = \sum_{l=1}^n \chi_{rl}$ derived alleles at a core site. Using these n_r derived alleles and $n - n_r = \sum_{l=1}^n (1 - \chi_{rl})$ ancestral alleles at a core site, we divide our sample of n homologous chromosomes into two mutually exclusive groups: derived and ancestral allele groups. The derived allele frequency at the core site is equal to $f_r = n_r/n$ and the frequency class (c_r) to which the core site belongs is also specified by n_r . For example, if the site is located at $r = 6$ as in Table 1, we have $n_6 = 3$, which belongs to c_2 , and divide the sample into the derived allele group of (1,2,3) chromosomes and the ancestral allele group of (4,5,6) chromosomes. This definition of allele groups and the following procedure can be extended, without any change, to the case where more than one SNP site determines two core haplotypes. Figure 1 shows one such example in which two core haplotypes, CGC vs. non-CGC, are defined at three SNP sites (rs3759916, rs3759915 and rs3759914) that are located in the promoter region of the human sialyltransferase (*ST8SIA2*) locus (Fujito et al., 2018). This gene is schizophrenia-associated and expressed preferentially in the brain, with the level being largely determined by the promoter SNPs. It is suggested that the expression level is a genetic determinant of schizophrenia risk and a non-risk SNP type (CGC-type) has significantly reduced promoter activity. In Asia and Europe, the major promoter type is not CGC but TCT.

Next, we compute the number (n_{rk}) of derived alleles at the k -th SNP site that simultaneously occur with derived alleles at the core site r . This number becomes non-zero only when both $\chi_{kl} = 1$ and $\chi_{rl} = 1$ hold true for some chromosome l . The n_{rk} is thus counted as

$$n_{rk} = \sum_{l=1}^n \chi_{rl} \chi_{kl} \quad (2)$$

and provides a measure of pairwise LD between the core site and any other neighboring SNP site. The n_k in Eq. (1) ranges from 1 to $n-1$ and the n_{rk} in Eq. (2) ranges from 0 to n_k or n_r , whichever is smaller. In Fig. 1, these numbers are represented as “barcodes” that depict SNP information by two-tone colored bars.

The barcode representation of n_k and n_{rk} differs from the traditional SFS in that it preserves spatial information about SNP sites and is stratified in eight layers that are on average in proportion to their ages. Figure 1 depicts derived alleles in c_6 and c_7 of *ST8SIA2*. In a central region in c_7 , 12 equally tall red bars for n_{rk} derived alleles are linked in the derived allele group; these sites may also contain a small number of derived alleles that are linked in the ancestral allele group (gray bars for $n_k - n_{rk}$ derived alleles). These barcodes indicate that the central region is in tight linkage with respect to the core site and the 12 SNP sites show 12 mutations that have accu-

Fig. 1. Barcodes for the numbers of derived alleles (C6 A and C7 A) and r^2 values for the extent of linkage disequilibrium (C6 B and C7 B) in classes c_6 and c_7 at the *ST8SIA2* locus. Magenta bars represent the number (n_{rk}) of derived alleles at the k -th SNP site that are linked in the core CGC haplotype, whereas gray bars represent the number ($n_k - n_{rk}$) of derived alleles that are linked in the core non-CGC haplotypes, as described in the text. The height of each bar n_k in c_7 ranges from 186 to 503 and that of each magenta bar ranges from 0 to 349 according to the number of CGC chromosomes in East Asians. The abscissa spans approximately 54 kb and the location of the core CGC sites is indicated by a single asterisk at the bottom (Fujito et al., 2018).

mutated in the common ancestral lineage of n_{rk} chromosomes. Later, it is more rigorously determined that this central LD region is approximately 16 kb long and ranges from 24 kb to 40 kb distant from the left end. When the central region is extended to left and right regions with each being approximately 19 kb long, it immediately becomes apparent that derived alleles in these extended regions tend to be linked with not only derived but also ancestral alleles at the core site. Under the infinite-sites model with recombination (Kimura, 1969), there are only two possibilities that can result in such allele configurations at two sites. One possibility is that derived alleles at the k -th site are of ancient origin and can therefore appear together with the derived alleles at the core site (e.g., $k = 8$ in Table 1). In this case, we have $n_{rk} = n_r < n_k$. The other is that derived alleles at the k -th site experienced recombination with respect to the core site. In this case, we have $n_{rk} < n_r$ and n_k . More thoroughly, the conditions for the latter case of recombination are that all of $\sum_{l=1}^n \chi_{rl} \chi_{kl}$, $\sum_{l=1}^n (1 - \chi_{rl}) \chi_{kl}$, $\sum_{l=1}^n \chi_{rl} (1 - \chi_{kl})$ and $\sum_{l=1}^n (1 - \chi_{rl}) (1 - \chi_{kl})$ are strictly positive as in the four-gamete test (Hudson and Kaplan, 1985):

$$n_{rk} > 0, n_k > n_{rk}, n_r > n_{rk} \text{ and } n > g_k = n_k + n_r - n_{rk}. \quad (3a)$$

In the example given in Table 1, only the two sites $k = 4$ and 9 satisfy these conditions with respect to the core site $r = 6$. In the barcode representation, the first two inequalities in Eq. (3a) are indicated by a two-tone colored bar at the k -th site, whereas the third is indicated by the difference between two red bars at the core site and the k -th site, and the last is not depicted explicitly but corresponds to the case where $n - n_r$ ancestral alleles at the core site are greater in number than derived alleles in a gray bar at the k -th site. This allele association between the core site r and the k -th site is commonly measured by the following LD formulae:

$$D = f_{rk} - f_r f_k \text{ or } r^2 = \frac{D^2}{f_r f_k (1 - f_r) (1 - f_k)} \quad (3b)$$

where $f_{rk} = n_{rk}/n$, $f_k = n_k/n$ and $f_r = n_r/n$ as before. The above “ r ” in r^2 (Hill and Robertson, 1968) should not be confused with the same symbol r used for the position of a core site or for the recombination rate given in the Discussion section. The $|D|$ or r^2 in Eq. (3b) is useful to grasp an outline of a core region: it constitutes a high LD region that includes a core site and exhibits a low level of derived allele polymorphism. Figure 1 also

plots pairwise r^2 values in c_6 and c_7 , revealing that the central region exhibits strong LD with the core site and is bounded by abruptly declined LD regions on both sides. Of importance is the examination of LD in individual classes within each of which SNPs are age-related and tend to disclose the linkage relationships. To further delimit the boundaries of the core region, we carried out exhaustive window analyses of F_c and determined the region that showed the minimum value.

In Fig. 2, we draw the entire barcode representation of the same genomic region as in Fig. 1 with the CGC haplotype and the alternative TCT haplotype as a respective core, and the regions on both sides are used for comparison with the core region and the size is taken as being almost the same as the core region. Two features emerge. Signals for tight linkage in the core region are characteristic to the CGC haplotype group, whereas several SNP sites in the TCT haplotype group fall under suspicion of recombination. It is therefore consistent

with the view that the extent of polymorphism within the TCT haplotype group is relatively high and the age is relatively old. Conversely, the pattern and extent of derived allele polymorphism in the CGC haplotype group are qualitatively and quantitatively different from those in the TCT haplotype group, suggesting rapid lineage-specific expansion of the former. This difference is captured by contrasting values of F_c , which is defined below.

Provided that we have identified a core region, we are now in a position to quantify such lineage-specific expansion by an appropriate measure. We define F_c as the ratio of the number of derived alleles in the derived allele group to that in the entire sample, where:

$$F_c = \frac{\sum_{k \in c} n_{rk}}{\sum_{k \in c} n_k} \quad (4)$$

for n_{rk} and n_k at all the SNP sites in certain frequency classes (c). The numerator and denominator are calcu-

Fig. 2. Entire barcode representation of the SNPs in a 54-kb region at the *ST8SIA2* locus. The three core SNPs are located in the right middle (an asterisk in each class) and define the CGC and TCT haplotype groups with frequency $349/1008 = 35\%$ and $515/1008 = 51\%$, respectively, in East Asians (the 1000 Genomes Project database). Note that according to the copy number of derived alleles at a site, the vertical scale differs greatly among different classes: c_1 (top) to c_8 (bottom).

lated in the same genomic region so that one can act as an internal control of the other. However, it is strongly recommended to exclude not only the own class (c_r) to which the core site belongs, but also the higher-than-own classes altogether, because these classes contain overwhelming numbers of derived alleles that accumulated in the common stem lineage of a derived allele (haplotype) group and overshadow useful information regarding the internal structure of descendant lineages. For this reason, we may rewrite Eq. (4) as $F_{<c_r}$ to express explicitly that it is computed in all the classes lower than the core allele frequency class. This F_c or $F_{<c_r}$ quantifies intra-allelic variability in the sense of Slatkin and Rannala (1997, 2000) without invoking any prior knowledge regarding intra-allelic genealogy. It is shown by simulation that the F_c statistic in Eq. (4) satisfactorily reflects genealogical structure within a derived allele group. It is important to realize that $F_c \approx f_r$ if no linkage relationship is expected between the core site and all other sites under consideration.

Table 2 shows the estimates of F_c in each of the eight frequency classes of *ST8SIA2* among East Asians in the 1000 Genomes Project database. These $\hat{F}_c$ are obtained separately in the central core region and, for comparison, in the neighboring region of 38 kb length as well (the rightmost column in Table 2). Here and subsequently, a caret over a statistic indicates an estimate. As the copy number of the CGC haplotype is 349 in the present sample of $n = 1008$, the core haplotype belongs to c_7 so that Eq. (4) is applied to all the classes lower than c_7 . Our estimate of $\hat{F}_{<c_7} = \frac{41}{2802} \approx 0.015$ implies that

within c_1 through c_6 , the CGC haplotype group contains only 1.5% of derived alleles relative to the whole in the corresponding frequency classes. Comparison of this estimate with $\hat{F}_{<c_7} = \frac{1513}{6305} = 0.240$ in the neighboring

region indicates that although the latter is not yet close to the CGC haplotype frequency in East Asians (0.35), it is much higher than that in the core region, suggesting that both sides are apparently in loose linkage with the core haplotype. It is also noted that $\hat{F}_{<c_7} \approx 0.015$ in the CGC haplotype group is in sharp contrast to $\hat{F}_{<c_8} = 0.349$ in the TCT haplotype group (Fig. 2). Although the large $\hat{F}_{<c_8}$ is due partly to the high core frequency $\hat{f}_r = 0.51$ of the TCT haplotype, the rest is attributed to a large number of TCT-linked derived alleles in c_6 and c_7 . In contrast, the extremely low $\hat{F}_{<c_7}$ value in the CGC haplotype stems from the virtual absence of CGC-linked derived alleles that belong to c_4 to c_6 despite abundant non-CGC-linked derived alleles (Table 2). In terms of genealogy, the deficiency of $\hat{F}_{c_4}$ to $\hat{F}_{c_6}$ corresponds to the presence of a large cluster without any solid family relationship among its members after they originated from a common ancestor. We take this unusually unstructured pattern

Table 2. Proportions of derived alleles that are linked to the CGC haplotype (n_{rk}) at the human *ST8SIA2* locus. Data are taken from East Asians ($n = 1008$) in the 1000 Genomes Project database. $F_c = \sum_{k \in c} n_{rk} / \sum_{k \in c} n_k$ where c stands for specified classes of allele frequencies.

$\hat{F}_c$	Entire 54-kb region ¹	Central 16-kb region ²	Both sides ³
$\hat{F}_{c_1}$	61/180=0.34	15/51=0.29	46/129=0.36
$\hat{F}_{c_2}$	44/154=0.29	15/42=0.36	29/112=0.26
$\hat{F}_{c_3}$	46/242=0.18	8/104=0.08	38/138=0.28
$\hat{F}_{c_4}$	77/441=0.17	0/64=0.00	77/377=0.20
$\hat{F}_{c_5}$	81/355=0.23	1/102=0.01	80/253=0.32
$\hat{F}_{c_6}$	1245/7735=0.16	2/2439=0.001	1243/5296=0.24
$\hat{F}_{c_7}$	12312/29184=0.42	4198/5853=0.72	8114/23331=0.35
$\hat{F}_{c_8}$	7799/23454=0.33	2541/9091=0.28	5258/14363=0.37
Total	21665/61745=0.35	6780/17746=0.38	14885/43999=0.34

¹ The 54-kb region is composed of the central region and both side regions.

² For the definition of the core (central) region, see the text.

³ Both sides implies the left and right side from the core. The size of each side region is approximately the same as that of the core.

and lack of derived alleles as evidence for rapid expansion of the CGC haplotype lineage. Exceptions are a few SNP sites at which there are one and two CGC-linked derived alleles in total in c_5 and c_6 , respectively (Table 2). As these CGC-linked derived alleles occur together with a large number of non-CGC-linked derived alleles at the same SNP sites, recombination is the most likely cause of these minor associations. Thus, the central core region we have just identified may not be perfectly free from recombination. Such imperfection is practically inescapable in actual data analyses and thus prompts us to be conservative in testing a null hypothesis.

Equally importantly, the observed $\hat{F}_c$ values in low frequency classes are much higher than those in high fre-

quency classes (Table 2). In fact, we have $\hat{F}_{c_{1-2}} = \frac{30}{93} = 0.323$

in classes c_{1-2} in contrast to $\hat{F}_{c_{3-6}} = \frac{11}{2709} = 0.004$ in classes

c_{3-6} . This increased level of rare variants may stem from the fact that when new mutations begin to accumulate within the derived allele group, the rate of return to equilibrium is faster in lower frequency classes than in higher frequency classes (Satta et al., 2018). Alternatively, the rate of return to equilibrium may suffer from recombination and/or the phasing problem for rare variants unless long-read sequencing is applied (cf. Field et al., 2016). In either case, inclusion of such rare variants inflates the estimate of $\hat{F}_{<c_7}$ to the level of 0.015. Before going further,

we simulated a model of selective sweeps and computed $F_{c_{1-2}}$ and $F_{c_{3-6}}$ separately to evaluate the effect of new mutations. We found that the ratio of $F_{c_{1-2}}$ to $F_{c_{3-6}}$ seldom exceeds the observed level of about 79 for a wide range of recombination rates (data not shown). Henceforth, it appears that the large value of $\hat{F}_{c_{1-2}}$ relative to $\hat{F}_{c_{3-6}}$ can be attributed to inaccurate haplotype phasing. Moreover, it is worth mentioning that under the standard neutral model, the $F_{<c_7}$ statistic is fairly insensitive to rare variants. For instance, the 1% threshold value (0.0346) of $F_{c_{3-6}}$ obtained under neutrality is approximately the same as 0.0384 of $F_{<c_7}$. This reflects the fact that under such a null equilibrium model, each frequency class contributes almost equally to $F_{<c_7}$. In other words, $F_{c_{3-6}}$ and $F_{<c_7}$ exhibit similar false-positive rates in rejecting neutrality (see below), and the use of $\hat{F}_{<c_7}$ instead of $\hat{F}_{c_{3-6}}$ may make the test conservative, depending on the quality of sequence data we use.

TESTING NEUTRALITY

To evaluate the false-positive rate (the probability α of a type I error) in our method, we performed *ms* (Hudson, 2002) and/or *fastsimcoal* (Excoffier and Foll, 2011) using the standard neutral model of constant N_e together with several demographic models proposed for modern human prehistory (e.g., Schaffner et al., 2005; Gutenkunst et al., 2009; Liu and Fu, 2015; Terhorst et al., 2017). In doing so, we restricted our examination to demographic models inferred for the East Asian or European meta-population and ignored migration with other continental populations. Moreover, these models differ from each other in detail. For instance, the model of Terhorst et al. (2017) is somewhat different from that of Schaffner et al. (2005) in the strength and timing of bottlenecks even if we discretize the population size estimated in the former. Accordingly, here we use two simplified versions of Schaffner et al. (2005) and Terhorst et al. (2017) that are subsequently referred to as SFD and TFD, respectively (Supplementary Fig. S1).

In simulating a core LD region of *ST8SIA2*, we assume no recombination and set an estimated mutation rate in a core region ($\hat{\theta}$) or the observed number of segregating sites ($\hat{S}$). The choice between θ and S does not matter under the standard neutral model as long as $S = \theta a_n$ approximately holds and both parameters are sufficiently large (e.g., $S \geq 100$); however, the choice may make a difference under the SFD and TFD models. For these models of changing population size, we use the observed S value to avoid unnecessary complications. First, we examined the $F_{<c_7}$ statistic under the standard neutral model and carried out 1,000 replications with a specified range of derived allele frequencies at a core site (0.35 ± 0.015). The distribution of $F_{<c_7}$ becomes unimodal with the mean and standard deviation (SD) of 0.26 ± 0.15 (Fig.

Fig. 3. Distribution of $F_{<c_7}$ ($1 \leq n_k \leq 185$) under the standard neutral model, provided that $S = 146$ and the frequency (f_r) of derived alleles at a core site is within 0.35 ± 0.015 . With over 1,000 replications, the mean and standard deviation of F_c are 0.255 ± 0.147 .

3). The threshold value of $F_{<c_7}$ for $\alpha = 0.01$ is 0.038 and the observed value of $\hat{F}_{<c_7} \approx 0.015$ is significantly small with $\alpha < 0.001$. We can thus reject the standard neutral model with a very low false-positive rate.

Under the SFD model, we obtained the distribution of $F_{<c_7}$ with the mean and SD of 0.24 ± 0.18 , finding that it becomes somewhat broad and skewed toward 0. Because of this and if the SFD model is treated as a null hypothesis, the false-positive rate increases; however, the threshold value of $F_{<c_7}$ for $\alpha = 0.01$ is as large as 0.026 and $\hat{F}_{<c_7} \approx 0.015$ is still significantly small ($\alpha < 0.002$). Likewise, under the TFD model, we obtained the distribution of $F_{<c_7}$ with the mean and SD of 0.24 ± 0.19 and found that the threshold value for $\alpha = 0.01$ is 0.017 and that $\alpha < 0.004$ for $\hat{F}_{<c_7}$. Hence, changing population size indeed increases the false-positive rate, but the effect on the F_c statistic is surprisingly small. Conversely, if the SFD and TFD models are treated as alternatives, the power of the F_c statistic defined by $1 - \beta$ (the probability β of a type II error) is too low to detect the effect in comparison with the standard neutral model. In fact, setting $\alpha = 0.05$ with the corresponding threshold value of $F_{<c_7} = 0.07$, we found that $\beta > 0.8$ under the SFD and TFD models. This high probability of a type II error or the low power reflects the broad and largely overlapping distributions of F_c between the standard neutral model of constant size and demographic models of changing population size (Fig. 4). In our subsequent analyses, we thus use the traditional SFD model as a null hypothesis rather than the standard neutral model unless otherwise specified.

Fig. 4. Distribution of F_c under the Schaffner et al. (2005) fluctuating drift (SFD) model (blue) in comparison with that under the standard neutral model (gray): $f_r = 0.35$, $S = 146$, no recombination, and $n = 1008$. The command lines in *ms* (Hudson, 2002) are: `/ms 1008 4000 -s 160` for the standard neutral model and `/ms 1008 4000 -s 160 -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` for the SFD model.

UNDER A SELECTIVE SWEEP

To examine the statistical power of F_c when a selective sweep is an alternative, we carried out coalescent simulation with selection at a single site that incorporates a selected allelic lineage into the neutral genealogy background. We used *discoal*, which generates either deterministic or stochastic trajectories for selective sweeps (Kern and Schrider, 2016); readers may also refer to *mb*s by Teshima and Innan (2009) and *MSMS* by Ewing and Hermisson (2010). For a given value of the population-scaled selection intensity $N_e s$, *discoal* efficiently simulated soft sweeps for a given initial frequency f_0 and a given current frequency f_r of a selected allele (Supplementary Fig. S2).

To imitate a putative soft sweep at *ST8SIA2* (Fujito et al., 2018), we set f_0 as 0.008 (the CGC haplotype frequency in Africans) and f_r as 0.35 (the CGC haplotype frequency in East Asians). As expected, the distribution of F_c becomes narrow, and for $N_e s = 200$ and $\theta = 20$, it concentrates in a small range of 0.00053–0.21 with the mean and SD of 0.020 ± 0.024 (Fig. 5). The threshold value for $\alpha = 0.01$ is accordingly as small as 0.001. To evaluate the power, we further examined various values of $N_e s$. When compared with the mean and SD of F_c (0.24 ± 0.18) for $N_e s = 0$, they sharply decrease to 0.077 ± 0.042 for $N_e s = 50$, 0.050 ± 0.028 for $N_e s = 100$, and 0.039 ± 0.029 for $N_e s = 150$ (Supplementary Fig. S3). Figure 6 shows the power of the F_c statistic against $N_e s$ for three type I error rates ($\alpha = 0.1, 0.05$, and 0.01 under the null hypothesis of the standard neutral model). For $\alpha = 0.01$, the F_c statistic can exhibit the power of $1 - \beta = 0.64$ if $N_e s$ is 100. As $N_e s$ increases up to 200, the power fur-

Fig. 5. Distribution of the F_c statistic under a selective sweep model (pink) in comparison with the SFD model (blue). The parameters and conditions for a soft sweep by genic selection are set as follows: $n = 1008$, 1,000 replications, region size = 16 kb, no recombination, $\theta = 20$, $N_e s = 200$, initial frequency $f_0 = 0.008$ and current frequency $f_r = 0.35$. The command line in *discoal* (Kern and Schrider, 2016) is: `/discoal 1008 1000 16000 -t 20 -ws 0 -a 400 -x 0.5 -f 0.008 -c 0.35`.

ther increases to > 0.78 by proportionally reducing the type II error rate β . Conversely, if $N_e s$ is ≤ 10 , the power remains poor and the F_c statistic likely fails to detect such weak selection signals.

We further applied our method to three human genetic loci that are known to have undergone hard selective sweeps. The best known is perhaps the locus for the adaptation of lactase (*LCT*) persistence that has occurred independently in Africans and Europeans (Tishkoff et al., 2007). Surveying a 1-Mb region surrounding rs4988235 (Supplementary Fig. S4A), we found that although the whole region under the hitchhiking effect may exceed 500 kb, a 50-kb central region is in extremely tight linkage and exhibits marked deficiency of core-linked derived alleles in the European meta-population. Because the current frequency ($\hat{f}_r$) of the core derived allele is 0.51, the core allele group belongs to c_8 and the $\hat{F}_{c_8}$ within the 50-kb core region is calculated as 0.006. To assess the statistical significance of the $\hat{F}_{c_8}$ with the $\hat{f}_r$, we simulated the SFD model for the European meta-population and a hard sweep model as an alternative hypothesis. The SFD simulation with $S = 263$ observed in the core region yielded a threshold value of F_{c_8} for $\alpha = 0.01$ at 0.032, which indicates that the observed value of $\hat{F}_{c_8} = 0.006$ is too small to be compatible with the SFD model ($\alpha < 0.001$). Conversely, the sweep model of genic selection with $\theta = S\alpha_n = 35$ and $N_e s = 200$ yielded $\beta = 0.218$ for $\alpha = 0.01$, suggesting that the hard sweep on *LCT* is much stronger than expected from $N_e s = 200$ (Fig. 7).

As the second example, we studied oculocutaneous albinism type II (*OCA2*) eye color variation (rs1800414) with $\hat{f}_r = 0.60$ in Asians (e.g., Sturm and Duffy, 2012). This derived allele is also associated with skin pigmenta-

tion and acts as a skin-lightening allele under epistatic interactions with the melanocortin-1 receptor gene. As with *LCT*, we first performed the barcode analysis for an approximately 1-Mb region surrounding rs1800414 (Supplementary Fig. S4B). The core region becomes 49 kb long and yields $\hat{F}_{<c_8} = 0.086$. We simulated the SFD model with the $\hat{f}_r$ and $S = 320$ for the East Asian meta-population as well as the same sweep model with $\theta =$

Fig. 6. Power of the F_c statistic (ordinate) against $N_e s$ (abscissa) and the three curves for three α values of 0.1 (green), 0.05 (blue) and 0.01 (red). These are evaluated by *discoal* with the same command line as in Fig. 5. Genic selection follows stochastic frequency trajectories with selection intensity s and begins at initial frequency f_0 at time t_0 and ends at current frequency f_r at time t_r . The time required to change the variant frequency from f_0 to f_r is well approximated by

$$t = t_r - t_0 \approx \frac{2}{s} \ln \frac{(1-f_0)f_r}{(1-f_r)f_0} \approx \frac{8.4}{s}.$$

43 and $N_e s = 200$. For $\alpha = 0.01$, the threshold value of $F_{<c_8}$ is 0.041, which is smaller than the observed value of $\hat{F}_{<c_8} = 0.086$ (Fig. 7). Conversely, the $\hat{F}_{<c_8}$ becomes too high to reject the SFD model at the 1% level. Thus the statistic on *OCA2* is only suggestive of positive selection.

As the third example, we examined the gene encoding ectodysplasin A receptor (*EDAR*), which is associated with Asian hair thickness (Fujimoto et al., 2008). At the core site 1540 (rs3827760) within the coding region, the Asian-specific derived allele C resulted in a non-synonymous change and increased its frequency to $\hat{f}_r = 0.85-0.90$ in CHB (Han Chinese Beijing) and JPT (Japanese Tokyo) populations, most likely after East Asians diverged from Europeans. The slow rate of LD decay measured by extended haplotype homozygosity (EHH) in Sabeti et al. (2002) and high extents of population differentiation measured by F_{ST} both strongly supported an ongoing selective sweep (Fujimoto et al., 2008). Consistent with this, we found a 73-kb core region with $\hat{F}_{<c_8} = 0.196$ (Supplementary Fig. S4C) and noted that this relatively high value of the $\hat{F}_{<c_8}$ is correlated with the high incidence of the 1540C allele in the Chinese and Japanese populations. The simulation revealed that for $\alpha = 0.01$, the threshold value of $F_{<c_8}$ is 0.233 under the SFD model with $S = 543$ and $\beta = 0.107$ under the sweep model with $\theta = 72$ and $N_e s = 200$ (Fig. 7). Thus, the $\hat{F}_{<c_8}$ is significantly small ($\alpha = 0.004$) and supports the notion that the hard sweep on *EDAR* is not much weaker than expected from $N_e s = 200$.

Fig. 7. Cumulative distribution of the F_c statistic under the neutral SFD model (blue) and the hard sweep model with $N_e s = 100, 200$ and 300 for *LCT*, *OCA2*, *EDAR* and *ASPM*. Simulation with 1,000 replications is performed for a fixed S value under the neutral SFD model, whereas it is performed for a specified θ value under the sweep model. A red triangle at the bottom of each panel shows an observed F_c value.

DISCUSSION

In our illustrative example, we examined the significance of the F_c statistic under the assumption of a fixed core allele frequency $f_r = 0.35$. However, as with EHH (Sabeti et al., 2002) and other statistics of the LRH (Vitti et al., 2013), F_c also depends noticeably on f_r . It is therefore necessary to carry out appropriate simulation studies to evaluate the statistical significance of observed F_c and/or to examine the empirical distribution for a number of its realized values in a genome (see Fujito et al., 2018 for comparison of various methods in their statistical power when applied to *ST8SIA2*). Nevertheless, to obtain a rough idea, we considered it useful to tabulate the f_r dependency of threshold values of F_c and other related statistics under the standard neutral model. In making such a table, we further assumed fairly large values of θ or S to reduce stochastic errors that might be induced by low mutation rates and, as a consequence, become more appropriate for *EDAR* with a larger core region than that of *ST8SIA2*. Table 3 shows that the threshold value of F_c is very sensitive to f_r while simultaneously illustrating the atypical characteristic of *LCT* with $\hat{f}_r = 0.51$ and $\hat{F}_{c_{c_8}} = 0.006$. In general, if $f_r \leq 0.6$ or $f_r \geq 0.7$, the mean ($\bar{F}_c$) of F_c is greater or smaller than the median, and the distribution is positively or negatively skewed accordingly.

Although the distribution of F_c is generally very broad under neutrality, it may be claimed that $\bar{F}_c < f_r$ can be expected only when recombination is rare or absent, because only particular branches in a coalescent tree can produce a cluster of chromosomes with specified frequency f_r (Fu, 1995; Li, 2011; Ferretti et al., 2017; Yang et al., 2018) and mutations in the common stem lineage are not counted in F_c . To illustrate this, let us consider a simple case of $n = 4$ (Supplementary Fig. S5). For $f_r = 3/4$, we can uniquely determine the tree topology in which only one suitable internal branch satisfies the frequency condition. Assuming the number of nucleotide changes is ideally in proportion to the average coalescent branch length, we can compute $\bar{F}_{c_{c_3}} = 0.69$ rather than $3/4$. Similarly, for $f_r = 1/2$, we obtain $\bar{F}_{c_1} = 0.38$ rather than $1/2$. Hence, $\bar{F}_c$ is slightly smaller than f_r in a coalescent tree. However, it appears that the difference becomes even larger for a large sample size n and for a high current frequency f_r (Table 3). Alternatively, if recombination occurs frequently, $\bar{F}_c$ should approach f_r because a core site and any other site tend to be in linkage equilibrium regardless of frequency classes. Simulation confirmed this assertion (Supplementary Fig. S6). Conversely, the observation of a large difference between f_r and F_c thus constitutes a first favorable signal of LD being worthy of further scrutiny.

When a derived allele at a core site increases in frequency and is eventually fixed in a population, we can no

Table 3. f_r dependency of the threshold values of the F_c statistic for specified type I error rates α and the same dependency of the mean and standard deviation (SD). As a window for a specified value of f_r must be set, it is assumed to be $f_r \pm \sqrt{f_r(1-f_r)/n}$. Under the assumption of the standard neutral model with $n = 1008$ and $\theta = 100$ or $S = 1000$, the F_c is computed for $<c_6$ when $f_r = 0.1$, for $<c_7$ when $f_r = 0.2$ to 0.4 and for $<c_8$ when $f_r \geq 0.6$. The exception is the case of $f_r = 0.5$, for which the average of F_c s is taken over realized values for $<c_7$ and $<c_8$. For each set of parameters, the number of replications is 1000. The differences between the two tables can likely be attributed to replication errors.

$\theta = 100$									
α	f_r								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.01	0.009	0.017	0.037	0.064	0.074	0.096	0.147	0.245	0.387
0.05	0.016	0.028	0.056	0.097	0.116	0.151	0.219	0.316	0.517
0.1	0.021	0.037	0.073	0.121	0.143	0.181	0.263	0.382	0.574
0.5	0.054	0.101	0.168	0.258	0.279	0.345	0.444	0.577	0.746
mean	0.072	0.127	0.197	0.283	0.300	0.345	0.440	0.562	0.723
SD	0.062	0.096	0.120	0.145	0.136	0.122	0.132	0.130	0.111

$S = 1000$									
α	f_r								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.01	0.010	0.017	0.039	0.071	0.082	0.109	0.125	0.222	0.428
0.05	0.017	0.029	0.062	0.101	0.117	0.159	0.206	0.327	0.527
0.1	0.023	0.039	0.075	0.122	0.143	0.193	0.254	0.386	0.585
0.5	0.053	0.099	0.172	0.262	0.280	0.343	0.438	0.578	0.753
mean	0.068	0.127	0.207	0.290	0.300	0.351	0.436	0.566	0.733
SD	0.053	0.097	0.132	0.146	0.132	0.121	0.136	0.133	0.106

longer compute the F_c statistic. *SLC24A5* is one of two known loci strongly affecting light skin pigmentation in Europeans and shows a fixed derived allele in the CEU population (Utah residents with Northern and Western European ancestry; Olalde et al., 2014). Although the variant is fixed in the population, the hierarchical barcodes of an approximately 20-kb region surrounding rs1426654 still indicate marked deficiency of SNPs in c_6 and c_7 in addition to a notable recovery of polymorphism in the lower classes (data not shown). As these features are likely footprints of a complete selective sweep, it is both worthwhile and possible to quantify such characteristics by some other measures than F_c . Kimura et al. (2007) considered two measures of the LRH and developed a method for when a selected allele has reached fixation in a local population (see also Maynard Smith and Haigh, 1974; Kaplan et al., 1989; Ronen et al., 2013; Schrider and Kern, 2017). Alternatively, we may apply the F_c statistic in a polymorphic population such as South East Asians

(SAS), although this may not work for the case of local adaptation. In SAS, we actually found a 134-kb core region with $\hat{f}_r = 0.60$ and $\hat{F}_{c_6} = 0.136$. Although the probability of a type I error ($\alpha = 0.03$) may not be sufficiently small, this finding suggests that positive selection has acted on *SLC24A5* in SAS as well.

In most simulations, we have assumed no recombination within a core region, not addressing the effect of recombination on neighboring regions. Clearly, it is worth demonstrating the effect on both the barcode representation and the F_c statistic. For this purpose, we ran *discoal* with recombination under $\theta = 20$, $N_e s = 200$ and $\hat{f}_r = 0.35$. Although it is necessary to assume a rather small sample size of $n = 200$ or less for technical reasons, the c_6 barcode thus generated reveals patterns similar to the c_7 barcode in Fig. 1 (Supplementary Fig. S7). We also confirmed that the F_{c_6} becomes approximately 0.01 in the central region, but increases to 0.34 on both sides. In short, simulation with selection and recombination can reproduce the pattern and level of polymorphism that resemble those observed in some human genomic regions under selective sweeps. More intriguingly, we found that the F_c statistic is little affected even when the recombination rate (r) per region is fairly high (e.g., $\rho = 4N_e r \geq 10$). This unexpected robustness certainly depends on how to evaluate intra-allelic variability under selective sweeps. As mentioned earlier, the F_c statistic is intended to exclude SNP mutations with $n_k \geq n_{rk} = n_r$ at site k because they have accumulated in the stem lineage of a derived allele group. However, recombination, if it occurs in one or a few lineages in the group, may act so as to incorporate some of those stem-lineage mutations into intra-allelic variability and substantially enhance it: this is because although the number of remaining unrecombined lineages is smaller than n_r , they can be still large in size (large n_{rk}). Fortunately, it is likely that SNP sites thus created within the derived allele group remain in the same frequency class as the core site c_r . In this case, the F_c statistic in Eq. (4) simply ignores those n_{rk} mutations from intra-allelic variability.

Next, we briefly discuss one issue concerning the distinction between hard sweeps and soft sweeps, the latter being the more dominant mode of adaptation (Schridder and Kern, 2017). As a soft sweep may drive multiple standing variants at linked neighboring sites from the beginning and retain the polymorphism to some extent even if selection has been completed, it is in general more challenging to detect a soft sweep than a hard sweep. Peter et al. (2012) concluded that whereas *LCT* and *EDAR* are subjected to hard sweeps, *ASPM* (*abnormal spindle-like microcephaly-associated*) (Mekel-Bobrov et al., 2005) is likely affected by a soft sweep. Owing to the similarity to *ST8SIA2*, we examined a core SNP site in *ASPM* with $\hat{f}_r = 0.41$ in Europeans and found that there exists a 26-kb core region with $\hat{F}_{c_{c_7}} = 0.052$. Under the

assumption of a hard sweep with $\theta = 20$ and $N_e s = 200$, contrary to Peter et al. (2012), the threshold value of $F_{c_{c_7}}$ becomes 0.029 for $\alpha = 0.01$ and the corresponding value of β becomes 0.186 (Fig. 7). These results may be taken as evidence for positive selection on *ASPM*, though not very strong. However, unlike that for a hard sweep, the power $(1 - \beta)$ for detecting a soft sweep depends strongly on the initial frequency f_0 of a selected allele (Innan and Kim, 2004; Hermisson and Pennings, 2005; Przeworski et al., 2005). For this reason, we examined the distribution of F_c for various values of f_0 under the conditions of $\theta = 20$, $n = 1008$ and $N_e s = 200$. Although the power is still as high as 0.74 when $f_0 = 0.01$, it sharply declines as f_0 further increases: 0.26 for $f_0 = 0.05$, 0.12 for $f_0 = 0.1$, and almost 0 for $f_0 = 0.2$, which is nearly halfway to the current frequency f_r . Without knowing the initial frequency of a selected allele, as is usually the case, it seems almost impossible to make definite conclusions about the power of a method for detecting soft sweeps (but see Satta et al., 2018 for one possibility). Nevertheless, if *ASPM* showed more unambiguous non-neutrality (i.e., $\alpha < 0.01$) concerning intra-allelic variability, it could have also been suggested that the initial frequency of the selected allele was no greater than 1% when a soft sweep began to operate; such a low initial frequency appears consistent with the geographic distribution of the *ASPM* variant, which is confined to Eurasia.

As a final caveat, whereas we have shown that changing population size does not markedly affect the F_c statistic, the demographic models examined here are not exhaustive. Many other demographic models or events potentially influence the F_c statistic. These include range expansion with so-called “surfing” in a wave front that may mimic a selective sweep and be relevant to the present context (Edmonds et al., 2004; Klopstein et al., 2006). However, this and other demographic causes such as population structure and admixture are beyond the scope of the present paper.

Source code for F_c statistic The source code (python) for the F_c statistic is available freely from our web site (<https://sites.google.com/site/sattalab/research>).

We thank Dr. Andy D. Kern for his kind and prompt response to our inquiry and request to utilize *discoal*. We also thank the editor and reviewers for their constructive criticisms, and Editage (www.editage.jp) for English language editing on early versions of this manuscript. This work was supported in part by the Japan Society for Promotion of Science (JSPS) grant 16H04821 to Y. S., the JSPS grants 23570271, 25101705 and 16K07535 to T. H., 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.
- 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, A., 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.
- Edmonds, C. A., Lillie, A. S., and Cavalli-Sforza, L. L. (2004) Mutations arising in the wave front of an expanding population. *Proc. Natl. Acad. Sci. USA* **101**, 975–979.
- Evans, P. D., Mekel-Bobrov, N., Vallender, E. J., Hudson, R. R., and Lahn, B. T. (2006) Evidence that the adaptive allele of the brain size gene *microcephalin* introgressed into *Homo sapiens* from archaic *Homo* lineage. *Proc. Natl. Acad. Sci. USA* **103**, 18178–18183.
- Ewing, G., and Hermisson, J. (2010) *MSMS*: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics* **26**, 2064–2065.
- Excoffier, L., and Foll, M. (2011) fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrary complex evolutionary scenarios. *Bioinformatics* **27**, 1332–1334.
- Eyre-Walker, A., Woolfit, M., and Phelps, T. (2006) The distribution of fitness effects on new deleterious amino acid mutations in humans. *Genetics* **173**, 891–900.
- 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.
- 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., Telis, N., Gao, Z., Gaulton, 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., Samakkarn, U., Settheetham-Ishida, W., Ishida, T., et al. (2008) A scan for genetic determinants of human hair morphology: *EDAR* is associated with Asian hair thickness. *Hum. Mol. Genet.* **17**, 835–843.
- Fujito, N. T., Satta, Y., Hane, M., Matsui, A., Yashima, K., Kitajima, K., Sato, C., Takahata, N., and Hayakawa, T. (2018) Positive selection on schizophrenia-associated *ST8SIA2* gene in post-glacial Asia. *PLoS One* **13**, e0200278.
- Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., and Bustamante, C. D. (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* **5**, e1000695.
- Hermisson, 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., 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.
- 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.
- Kaplan, N. L., Hudson, R. R., and Langley, C. H. (1989) The “hitchhiking effect” revisited. *Genetics* **123**, 887–899.
- Kern, A. D., and Schrider, D. R. (2016) Discoal: flexible coalescent simulations with selection. *Bioinformatics* **32**, 3839–3841.
- Kimura, M. (1969) The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* **61**, 893–903.
- Kimura, M., and Ohta, T. (1973) The age of a neutral mutant persisting in a finite population. *Genetics* **75**, 199–212.
- Kimura, R., Fujimoto, A., Tokunaga, K., and Ohashi, J. (2007) A practical genome scan for population-specific strong selective sweeps that have reached fixation. *PLoS One* **2**, e286.
- Klopfstein, S., Currat, M., and Excoffier, L. (2006) The fate of mutations surfing on the wave of a range expansion. *Mol. Biol. Evol.* **23**, 482–490.
- Lachance, J., and Tishkoff, S. A. (2013) Population genomics of human adaptation. *Annu. Rev. Ecol. Evol. Syst.* **44**, 123–143.
- Li, H. (2011) A new test for detecting recent positive selection that is free from the confounding impacts of demography. *Mol. Biol. Evol.* **28**, 365–375.
- Li, M. J., Wang, L. Y., Xia, Z., Wong, M. P., Sham, P. C., and Wang, J. (2014) dbPSHP: a database of recent positive selection across human populations. *Nucleic Acids Res.* **42**, D910–D916.
- Liu, X., and Fu, Y. X. (2015) Exploring population size changes using SNP frequency spectra. *Nat. Genet.* **47**, 555–559.
- 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.
- Nei, M., and Li, W.-H. (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. USA* **76**, 5269–5273.
- 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.
- Olalde, I., Allentoft, M. E., Sánchez-Quinto, F., Santpere, G., Chiang, C. W. K., DeGiorgio, M., Prado-Martínez, J., Rodríguez, J. A., Rasmussen, S., Quilez, J., et al. (2014) Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European. *Nature* **507**, 225–228.
- Pavlidis, P., Živković, D., Stamatakis, A., and Alachiotis, N. (2013) SweeD: likelihood-based detection of selective sweeps in thousands of genomes. *Mol. Biol. Evol.* **30**, 2224–2234.
- Peter, B. M., Huerta-Sánchez, E., and Nielsen, R. (2012) Distinguishing between selective sweeps from standing variation and from a *de novo* mutation. *PLoS Genet.* **8**, e1003011.
- Przeworski, M., Coop, G., and Wall, J. D. (2005) The signature of positive selection on standing genetic variation. *Evolution* **59**, 2312–2323.
- Ronen, R., Udpa, N., Halperin, E., and Bafna, V. (2013) Learning natural selection from the site frequency spectrum. *Genetics* **195**, 181–193.

- Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., et al. (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832–837.
- Satta, Y., Fujito, N. T., and Takahata, N. (2018) Nonequilibrium neutral theory for hitchhikers. *Mol. Biol. Evol.* **35**, 1362–1365.
- Sawyer, S. A., and Hartl, D. L. (1992) Population genetics of polymorphism and divergence. *Genetics* **132**, 1161–1176.
- Schaffner, S. F., Foo, C., Gabriel, S., Reich, D., Daly, M. J., and Altshuler, D. (2005) Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* **15**, 1576–1583.
- 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.
- Schrider, D. R., Shanku, A. G., and Kern, A. D. (2016) Effects of linked selective sweeps on demographic inference and model selection. *Genetics* **204**, 1207–1223.
- 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.
- Sturm, R. A., and Duffy, D. L. (2012) Human pigmentation genes under environmental selection. *Genome Biol.* **13**, 248.
- Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595.
- 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., and Innan, H. (2009) mbs: modifying Hudson's ms software to generate samples of DNA sequences with a biallelic site under selection. *BMC Bioinformatics* **10**, 166.
- 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.
- Vitti J. J., Grossman, S. R., and Sabeti, P. C. (2013) Detecting natural selection in genomic data. *Annu. Rev. Genet.* **47**, 97–120.
- Watterson, G. A. (1975) On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**, 256–276.
- 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.
- Zeng, K., Fu, Y. X., Shi, S., and Wu, C. I. (2006) Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics* **174**, 1431–1439.