Detecting Recent Positive Selection with a Single Locus Test Bipartitioning the Coalescent Tree

Zongfeng Yang,^{*,†} Junrui Li,^{*,‡} Thomas Wiehe,[§] and Haipeng Li^{*,1}

^{*} Key Laboratory of Computational Biology, Chinese Academy of Sciences-Max Planck Gesellschaft (CAS-MPG) Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China,

[†]University of Chinese Academy of Sciences, Beijing 100049, China, [‡]Howard Hughes Medical Institute, University of California, San Francisco, California 94143, and [§]Institut für Genetik, Universität zu Köln, 50674 Germany

ORCID IDs: 0000-0001-9477-6843 (Z.Y.); 0000-0001-7856-4488 (H.L.)

ABSTRACT Many population genomic studies have been conducted in the past to search for traces of recent events of positive selection. These traces, however, can be obscured by temporal variation of population size or other demographic factors. To reduce the confounding impact of demography, the coalescent tree topology has been used as an additional source of information for detecting recent positive selection in a population or a species. Based on the branching pattern at the root, we partition the hypothetical coalescent tree, inferred from a sequence sample, into two subtrees. The reasoning is that positive selection could impose a strong impact on branch length in one of the two subtrees while demography has the same effect on average on both subtrees. Thus, positive selection should be detectable by comparing statistics calculated for the two subtrees. Simulations demonstrate that the proposed test based on these principles has high power to detect recent positive selection even when DNA polymorphism data from only one locus is available, and that it is robust to the confounding effect of demography. One feature is that all components in the summary statistics (D_u) can be computed analytically. Moreover, misinference of derived and ancestral alleles is seen to have only a limited effect on the test, and it therefore avoids a notorious problem when searching for traces of recent positive selection.

KEYWORDS positive selection; adaptation; demography

DETECTING recent events of positive selection and identifying beneficial alleles is of continued high interest in evolutionary biology. During the last two decades, many methods for this purpose have been proposed. Recent positive selection can alter the polymorphism pattern of a neutral locus which is partially or completely linked to a selected locus (*i.e.*, the locus carrying the beneficial allele). The altered site frequency spectrum can be detected if it differs significantly from the expectation under the standard neutral model (Tajima 1989b; Fu and Li 1993; Fay and Wu 2000; Zeng *et al.* 2006; Achaz 2009; Ferretti *et al.* 2010). Recent positive selection may also reduce genetic diversity around the selected locus due to genetic hitchhiking (Maynard

Smith and Haigh 1974; Kim and Stephan 2002; Li and Stephan 2005; Nielsen *et al.* 2005), and result in an elongation of haplotype blocks (Sabeti *et al.* 2002). Several excellent reviews (Sabeti *et al.* 2006; Jensen *et al.* 2007; Fu and Akey 2013) provide more insights into the details of selection detection algorithms. Recently, also machine learning techniques have been proposed (Pavlidis *et al.* 2010; Lin *et al.* 2011; Ronen *et al.* 2013; Pybus *et al.* 2015; Schrider and Kern 2016). However, it still remains a challenging task to distinguish positive selection from demographic events since the amount and pattern of DNA polymorphism in samples from a population is typically affected not only by recent positive selection but also by demography (Nei *et al.* 1975; Watterson 1986; Tajima 1989a; Jensen *et al.* 2005; Teshima *et al.* 2006). The confounding effect of demography cannot be neglected when scanning genomes for candidate loci of recent positive selection (MacCallum and Hill 2006; Akey 2009; Hermisson 2009; Scheinfeldt and Tishkoff 2013; Vitti *et al.* 2013; Adrion *et al.* 2015).

Copyright © 2018 by the Genetics Society of America

doi: <https://doi.org/10.1534/genetics.117.300401>

Manuscript received October 14, 2017; accepted for publication December 1, 2017; published Early Online December 7, 2017.

Supplemental material is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.117.300401/-/DC1.

¹Corresponding author: Laboratory of Evolutionary Genomics, CAS-MPG Partner Institute for Computational Biology, Yue Yang Rd. 320, Shanghai 200031, China. E-mail: lihaipeng@picb.ac.cn

Recent positive selection may produce a coalescent tree which is different from that under neutral evolution (Kaplan *et al.* 1989; Barton 1998; Fay and Wu 2000). Let us consider a locus partially linked to the selected locus, which means that a few recombination events happened during the selected phase. While selection operates, a lineage could escape from the selective sweep by recombination (Kaplan *et al.* 1989; Fay and Wu 2000; Kim and Nielsen 2004). In the tree topology this is reflected by one particularly long branch emerging from the root of the tree (Figure 1). Such tree topology is often called *unbalanced* (Sibert *et al.* 2002; Blum *et al.* 2006b; Li 2011; Li and Wiehe 2013). The unbalancedness of trees has been well studied (Kirkpatrick and Slatkin 1993; Blum *et al.* 2006a), and can also be measured as the distribution of the average length of the base branches that emanate from the root (Uyenoyama 1997).

Let us consider the tree of a sample collected from a random mating population. It is known that the probability for a tree to be unbalanced is independent of changes in population size (Tajima 1983; Hudson 1991). However, the probability of an unbalanced tree increases substantially when the neutral locus is partially linked to a positively selected site (Kaplan *et al.* 1989; Fay and Wu 2000). In this case, the majority of lineages fail to escape the selective sweep, so they will coalesce as a star-like subtree, and the internal branches are very short (Galtier *et al.* 2000) (Figure 1C). We found that the ratio between the lengths of two subtrees under the selective sweep model departs from the expectation under the standard neutral model (Figure 1). Therefore, hitchhiking also biases the average number of mutations in the two subtrees.

Recently, several groups proposed to detect selection using the information of coalescent tree, including topology of the tree (Li 2011; Disanto *et al.* 2013; Li and Wiehe 2013; Ferretti *et al.* 2017), and the distribution of coalescent times (Hunter-Zinck and Clark 2015; Ronen *et al.* 2015). Following those studies, we proposed a new test statistic, D_u , to detect recent positive selection in a varying size population. The tree can be partitioned into two parts, denoted as major and minor branches (Figure 1). Based on the infinite-sites model and the number of mutations occurred on the major and minor branches, we can calculate $\hat{\theta}_{\text{major}}$ and $\hat{\theta}_{\text{minor}}$. Under the standard neutral model, we expect the balanced tree, and $E(\hat{\theta}_{\text{major}}) = E(\hat{\theta}_{\text{minor}}) = E(\hat{\theta}_w)$, where θ_w is Watterson's θ (Watterson 1975), and estimated as the number of segregating sites divided by the tree length. However, under the selective sweep model, we expect an unbalanced tree, and $E(\hat{\theta}_{\text{major}}) < E(\hat{\theta}_{\text{minor}})$, since positive selection has a stronger impact on the major branches than on the minor branches. Thus, the unbalanced tree and a normalized difference between $\hat{\theta}_{\text{major}}$ and $\hat{\theta}_{\text{minor}}$ (denoted as D_u) is proposed as a statistic to test for recent positive selection.

To measure the robustness of this statistic with respect to different confounding factors, we examined its false positive rate. Here, the false positive rate is defined as the probability to reject the standard neutral model (the null hypothesis) when

actually no positive selection happened (Przeworski 2002; Jensen *et al.* 2005; Li 2011). A high false positive rate indicates that confounding factors should be taken into consideration when searching for loci under positive selection. We also compared our new method with other single locus tests for selection [see reviews by Sabeti *et al.* (2006), Fu and Akey (2013)]. Many of the available methods consider DNA polymorphism data from multiple loci, to first obtain an empirical background distribution, and then to perform genome scans for selective sweeps (Akey *et al.* 2002; Kim and Stephan 2002; Li and Stephan 2005, 2006; Nielsen *et al.* 2005; Sabeti *et al.* 2007). Generally, those methods require much more information than a single locus. Therefore, to conduct a fair comparison, we consider here only single locus based methods.

Methods

Statistical test

Unbalanced bifurcating trees have been introduced before (Purvis and Agapow 2002; Holman 2005; Li 2011; Li and Wiehe 2013). Here, we make a modification so that the unbalancedness can be defined for small samples. Let ψ_1 and ψ_2 be the number of left- and right-descendants of the two branches that originate from the root node of the tree, respectively, where $\psi_1 + \psi_2 = n$, and n is the number of sampled chromosomes. Furthermore, let $\Gamma = \max\{\psi_1, \psi_2\}$, such that Γ may take integer values from $\lceil n/2 \rceil$ to $(n-1)$, where $\lceil z \rceil$ means the smallest integer $\geq z$. Under neutrality, it is known that, when n is odd $P(\Gamma = x) = \frac{2}{n-1}$ for $1 \leq x \leq \frac{n-1}{2}$, and, when n is even,

$$P(\Gamma = x) = \begin{cases} \frac{2}{n-1}, & 1 \leq x < \frac{n}{2} \\ \frac{1}{n-1}, & x = \frac{n}{2} \end{cases} \quad (\text{Tajima 1983}).$$

To bound the number of false positives produced by random topologies, we set $\gamma_n = \min\{n-1, \lceil 0.95n + 0.05 \rceil\}$. We call a tree unbalanced if $n \geq 20$ and 10% probability that the tree is more unbalanced (*i.e.*, $\Gamma \geq \gamma_n$) under the standard Wright-Fisher neutral model. For smaller sample size, one can only take $\Gamma = n-1$, violating the 10% probability criterion.

Then let us consider a sample from a diploid population (although the test would also hold for haploid populations). Without loss of generality, we assume $\psi_1 \geq \psi_2$, and denote v as the node that is the most recent common ancestor (MRCA) of the ψ_1 lineages that are the left-descendants of the root. Thus, the whole sample is partitioned into two subsamples by the node v (Figure 1). The branches of the subsample rooted by the node v are defined as the major branches, and the rest of the tree is defined as the minor branches. Note that the minor branches include the two basal branches immediately descended from the root. Let $E(L_{\text{minor}})$ be the expected total length of the minor branches under neutrality, scaled so

Figure 1 Genealogical trees under the standard neutral model and hitchhiking model. The branches of the subtree rooted by the node v are defined as the major (thick) branches and the rest of the tree is defined as the minor (thin) branches. (A) Balanced tree under the standard neutral model. (B) Unbalanced tree under the standard neutral model. (C) Unbalanced tree under the hitchhiking model when the sample size is small. (D) Unbalanced tree under the hitchhiking model with a large sample size n . The major part consists of $(n-2)$ lineages, and the minor part 2 lineages.

that 1 unit corresponds to $2N$ generations, where N is the effective size of a diploid population. $E(L_{\text{major}})$ is the expected total length of the major branches under neutrality. $E(L_{\text{major}})$ and $E(L_{\text{minor}})$ are calculated under the standard neutral model, and $E(L_{\text{tree}}) = E(L_{\text{major}}) + E(L_{\text{minor}})$, where L_{tree} is the tree length. Moreover, it is well-known that $\theta (= 4N\mu)$ can be estimated from the number of mutations on the tree since $E(\theta_w) = 2E(K)/E(L_{\text{tree}})$ (Watterson 1975; Hudson 1991), where μ is the mutation rate per generation, K the number of segregating sites (*i.e.*, the number of mutations occurred on the tree). Similarly, θ_{major} and θ_{minor} can also be estimated from the number of mutations on the major branches (K_{major}) and minor branches (K_{minor}), respectively, and we have $K = K_{\text{major}} + K_{\text{minor}}$. To calculate θ_{major} , θ_{minor} , and D_u , details are given as follows:

Let T_k ($2 \leq k \leq n$) be the time duration required for k sequences to coalesce to $(k-1)$ sequences, *i.e.*, the so-called k -coalescent time. Following the standard coalescent process (Hudson 1991), we have

$$E(T_k) = \frac{2}{k(k-1)}, \quad (1)$$

$$\text{Var}(T_k) = \frac{4}{k^2(k-1)^2}, \quad (2)$$

$$E(T_k^2) = \text{Var}(T_k) + E^2(T_k) = \frac{8}{k^2(k-1)^2}. \quad (3)$$

Furthermore, we can derive the expectation and variance of the branch lengths of the tree L_{tree} (Hudson 1991) from Equations (1) and (2):

$$E(L_{\text{tree}}) = E\left(\sum_{i=2}^n iT_i\right) = \sum_{i=2}^n iE(T_i) = \sum_{i=2}^n \frac{2}{i-1}, \quad (4)$$

$$\begin{aligned} \text{Var}(L_{\text{tree}}) &= \text{Var}\left(\sum_{i=2}^n iT_i\right) = \sum_{i=2}^n i^2 \text{Var}(T_i) \\ &= \sum_{i=2}^n \frac{4}{(i-1)^2}. \end{aligned} \quad (5)$$

We first take $\Gamma = n-1$ as an example to derive the statistical properties of an unbalanced tree. More general results for unbalanced trees with $\Gamma \leq n-1$ are given in the Appendix. When $\Gamma = n-1$, the expectation and variance of L_{minor} and L_{major} can be obtained from Equations (1)–(3) as follows:

$$\begin{aligned} E(L_{\text{minor}}) &= E\left(T_2 + \sum_{i=2}^n T_i\right) \\ &= E(T_2) + \sum_{i=2}^n E(T_i) \\ &= 1 + \sum_{i=2}^n \frac{2}{i(i-1)} = 3 - \frac{2}{n}, \end{aligned} \quad (6)$$

$$\begin{aligned} \text{Var}(L_{\text{minor}}) &= \text{Var}\left(T_2 + \sum_{i=2}^n T_i\right) \\ &= 4\text{Var}(T_2) + \sum_{i=3}^n \text{Var}(T_i) \\ &= \sum_{i=2}^n \frac{8}{i^2} - \frac{n^2 - 8n + 4}{n^2}, \end{aligned} \quad (7)$$

$$\begin{aligned} E(L_{\text{major}}) &= E\left(\sum_{i=3}^n (i-1)T_i\right) \\ &= \sum_{i=3}^n (i-1)E(T_i) \\ &= \sum_{i=3}^n \frac{2}{i}, \end{aligned} \quad (8)$$

$$\begin{aligned} \text{Var}(L_{\text{major}}) &= \text{Var}\left(\sum_{i=3}^n (i-1)T_i\right) \\ &= \sum_{i=3}^n (i-1)^2 \text{Var}(T_i) = \sum_{i=3}^n \frac{4}{i^2}. \end{aligned} \quad (9)$$

Equation (6) holds because L_{minor} is equal to the sum of the tree height and the branch length between the root and the node v (Figure 1). Since

$$\begin{aligned} \text{Var}(L_{\text{tree}}) &= \text{Var}(L_{\text{major}} + L_{\text{minor}}) \\ &= \text{Var}(L_{\text{major}}) + 2\text{Cov}(L_{\text{major}}, L_{\text{minor}}) \\ &\quad + \text{Var}(L_{\text{minor}}), \end{aligned} \quad (10)$$

we can get the covariance of L_{major} and L_{minor}

$$\begin{aligned} \text{Cov}(L_{\text{major}}, L_{\text{minor}}) &= \frac{\text{Var}(L_{\text{tree}}) - \text{Var}(L_{\text{major}}) - \text{Var}(L_{\text{minor}})}{2} \\ &= \frac{1}{2} \left(\sum_{i=2}^n \frac{4}{(i-1)^2} - \sum_{i=3}^n \frac{4}{i^2} - 4 - \sum_{i=3}^n \frac{4}{i^2(i-1)^2} \right) \\ &= 2 - \frac{4}{n} - \sum_{i=3}^n \frac{4}{i^2}. \end{aligned} \quad (11)$$

For a general value of Γ , we have $\lfloor n/2 \rfloor < \Gamma \leq n-1$, where $\lfloor z \rfloor$ means the largest integer $\leq z$. Let $x = n - \Gamma$, so we have $1 \leq x < n/2$. If the sample consists of n sequences, there are

Figure 2 Illustrations of possible coalescent event vectors for a sample when $n = 5$ and $\Gamma = 3$. There are $\binom{3}{2} = 3$ possible sequences of coalescent events. The vector elements are ordered according to the order of coalescent events; “1” stands for the coalescent event on major branches and “0” for the coalescent event on minor branches. Each coalescent event on major branches corresponds to an index e_i , where i is the number of lineages in the major part right before the coalescent event, and e_i is the number of lineages in the whole tree right before this event.

$(n - 2)$ coalescent events excluding the last one (from present to past). If we denote coalescent events that happen on major branches with “1,” and those that happen on minor branches with “0,” the sequence of coalescent events, from present to past, is a binary vector with $(\Gamma - 1)$ “1”s and $(x - 1)$ “0”s. Thus, there are $C = \binom{n-2}{\Gamma-1}$ possible binary vectors B_m , $m = 1, 2, \dots, C$. As an example, Figure 2 illustrates all the possible coalescent event vectors when $n = 5$ and $\Gamma = 3$. We use $\mathbf{B}$ to denote the set of binary vectors. Each element in $\mathbf{B}$ represents a set of trees having the same sequence of coalescent events. Each set of trees occur with the same probability ($1/C$), since tree topology is independent of coalescent time, and trees with the same topology have the same probability. In other word $p(m = i) = 1/C$ for all $i \in \{1, 2, \dots, C\}$.

For convenience, we introduced a series of indices $e_2, e_3, \dots, e_{n-x}$. Each index e_i corresponds to a coalescent event on major branches. The subscript i in e_i denotes the number of lineages in major branches right before this coalescent event (from present to past), and the value e_i represents the number of lineages in the whole tree at the same time (Figure 2). With these indices, the expectation of L_{major} can be calculated following the law of total expectation (details are given in the Appendix):

$$E(L_{\text{major}}) = E(E(L_{\text{major}}|m)) = \frac{1}{C} \sum_{k=3}^n \frac{2d_k}{k(k-1)}, \quad (12)$$

where

$$d_k = \sum_{i=3}^{x+2} \sum_{j=1, k-x-1 \leq j \leq k-i+1}^{n-x-1} \binom{k-i}{j-1} \binom{n-k}{n-x-j-1} (j+1)$$

and $m = 1, 2, \dots, C$. Since $L_{\text{tree}} = L_{\text{major}} + L_{\text{minor}}$, we can easily obtain the expectation of L_{minor} from Equations (4) and (12),

$$E(L_{\text{minor}}) = E(L_{\text{tree}}) - E(L_{\text{major}}) = \sum_{i=2}^n \frac{2}{i-1} - E(L_{\text{major}}). \quad (13)$$

Similar to the derivation of $E(L_{\text{major}})$, we derived the variance of L_{major} and L_{minor} (details are presented in the Appendix),

$$\begin{aligned} \text{Var}(L_{\text{major}}) &= \frac{1}{C} \sum_{e_2=3}^{x+2} \sum_{e_3=e_2+1}^{x+3} \dots \sum_{e_{n-x}=e_{n-x-1}+1}^n \\ &\times \left(\sum_{i=e_2}^n \sum_{j=i}^n \frac{8a_i a_j}{ij(i-1)(j-1)} \right) - E^2(L_{\text{major}}), \end{aligned} \quad (14)$$

where a_i is determined by: $a_i = k$ if $e_k \leq i < e_{k+1}$ ($k \in \{2, 3, \dots, n-x\}$ and $e_{n-x+1} = n+1$).

$$\begin{aligned} \text{Var}(L_{\text{minor}}) &= \frac{1}{C} \sum_{e_2=3}^{x+2} \sum_{e_3=e_2+1}^{x+3} \dots \sum_{e_{n-x}=e_{n-x-1}+1}^n \\ &\times \left(\sum_{i=2}^n \sum_{j=i}^n \frac{8b_i b_j}{ij(i-1)(j-1)} \right) - E^2(L_{\text{minor}}), \end{aligned} \quad (15)$$

where b_i is determined by $b_i = \begin{cases} i, & 2 \leq i < e_2 \\ i-k, & e_k \leq i < e_{k+1} \end{cases}$, where

$2 \leq k \leq n-x$, $e_{n-x+1} = n+1$. The covariance of L_{major} and L_{minor} is

$$\begin{aligned} \text{Cov}(L_{\text{major}}, L_{\text{minor}}) &= \left(\text{Var}(L_{\text{tree}}) - \text{Var}(L_{\text{major}}) \right. \\ &\quad \left. - \text{Var}(L_{\text{minor}}) \right) / 2, \end{aligned} \quad (16)$$

where $\text{Var}(L_{\text{tree}})$, $\text{Var}(L_{\text{major}})$, and $\text{Var}(L_{\text{minor}})$ are given by Equations (5), (14), and (15), respectively.

Define $\theta = 4N\mu$, where μ is the mutation rate of the locus, and let K_{minor} be the number of mutations on the minor branches. Then, we have $E(K_{\text{minor}}) = E(L_{\text{minor}})\theta/2$. Similarly, we have $E(K_{\text{major}}) = E(L_{\text{major}})\theta/2$, where K_{major} is the number of mutations on the major branches. Thus, under the infinite-sites model, the population parameter θ can be estimated from K_{minor} and K_{major} separately as

$$\hat{\theta}_{\text{minor}} = 2K_{\text{minor}}/E(L_{\text{minor}}), \quad (17)$$

$$\hat{\theta}_{\text{major}} = 2K_{\text{major}}/E(L_{\text{major}}). \quad (18)$$

Similar to Tajima's D (Tajima 1989b), we denote the normalized difference between $\hat{\theta}_{\text{major}}$ and $\hat{\theta}_{\text{minor}}$ as D_u . That is,

$$D_u = \frac{\hat{\theta}_{\text{major}} - \hat{\theta}_{\text{minor}}}{\sqrt{\text{Var}(\hat{\theta}_{\text{major}} - \hat{\theta}_{\text{minor}})}}. \quad (19)$$

Then, we have $\text{Var}(\hat{\theta}_{\text{major}} - \hat{\theta}_{\text{minor}}) = \text{Var}(\hat{\theta}_{\text{major}}) - 2\text{Cov}(\hat{\theta}_{\text{major}}, \hat{\theta}_{\text{minor}}) + \text{Var}(\hat{\theta}_{\text{minor}})$, where $\text{Var}(\hat{\theta}_{\text{minor}})$, $\text{Var}(\hat{\theta}_{\text{major}})$ and $\text{Cov}(\hat{\theta}_{\text{major}}, \hat{\theta}_{\text{minor}})$ can be expressed as functions of expectation, variance, and covariance of L_{minor} and L_{major} (Equations 12–16). From Equation (16), the variance of $\hat{\theta}_{\text{minor}}$ is obtained as

$$\text{Var}(\hat{\theta}_{\text{minor}}) = \frac{\text{Var}(K_{\text{minor}})}{E^2(L_{\text{minor}}/2)} = \frac{2\theta E(L_{\text{minor}}) + \theta^2 \text{Var}(L_{\text{minor}})}{E^2(L_{\text{minor}})}, \quad (20)$$

where θ is estimated by Watterson's method (Watterson 1975), that is, $\hat{\theta}_w = K / \sum_{i=1}^{n-1} \frac{1}{i}$. Similarly, we have the variance of $\hat{\theta}_{\text{major}}$

$$\text{Var}(\hat{\theta}_{\text{major}}) = \frac{2\theta E(L_{\text{major}}) + \theta^2 \text{Var}(L_{\text{major}})}{E^2(L_{\text{major}})}. \quad (21)$$

Calculation of $\text{Cov}(\hat{\theta}_{\text{major}}, \hat{\theta}_{\text{minor}})$ is given in the Appendix. The result is

$$\text{Cov}(\hat{\theta}_{\text{major}}, \hat{\theta}_{\text{minor}}) = \theta^2 \frac{\text{Cov}(L_{\text{major}}, L_{\text{minor}})}{E(L_{\text{major}})E(L_{\text{minor}})}. \quad (22)$$

Following previous studies (Tajima 1989b; Fu and Li 1993), θ^2 was estimated by $K(K-1) / \left(\left(\sum_{i=1}^{n-1} \frac{1}{i} \right)^2 + \sum_{i=1}^{n-1} \frac{1}{i^2} \right)$. Under the standard neutral model, we have $E(\hat{\theta}_{\text{major}}) = E(\hat{\theta}_{\text{minor}}) = \theta$, thus $E(D_u) \approx 0$. However, it is expected that recent positive selection has stronger impact on L_{major} than L_{minor} (Figure 1, C and D), which reduces $E(L_{\text{major}})/E(L_{\text{minor}})$. This results in a stronger reduction of K_{major} than of K_{minor} . Thus $E(\hat{\theta}_{\text{major}}) < E(\hat{\theta}_{\text{minor}}) < \theta$, and $E(D_u) < 0$.

To get the inferred value of Γ for the locus, which is denoted by $\hat{\Gamma}$ (i.e., the branching pattern of the root), an unweighted pair group method with arithmetic mean (UPGMA) tree (Sneath and Sokal 1973) was reconstructed based on nucleotide differences. Here, only the information on the root of the UPGMA tree was collected. UPGMA was chosen in this study because it is an effective tree-construction algorithm, and can naturally determine the root. There are also other algorithms based on nucleotide differences that has similar efficiency, for instance neighbor-joining (Saitou and Nei 1987), but the latter is unable to infer the root without the information of outgroup. Moreover, in contrast to between-species phylogenetics, rate-variation of the mutation rate along branches is not a major concern in population genetic genealogies. A similar tree-bipartition algorithm was developed to estimate coalescent times (Tang *et al.* 2002; King and Wakeley 2016).

To calculate $\hat{D}_u$, let the sampled sequences be numbered from 1 to n , assuming that the sample consists of n sequences. Thus, $S = \{1, 2, \dots, n\}$ represents the whole sample. Without loss of generality, denote the descendants of the left branch originating from the root as $\{1, \dots, \hat{\Gamma}\}$, and the descendants of the right branch originating from the root as $\{\hat{\Gamma} + 1, \dots, n\}$. Then, under the infinite sites model,

$\hat{K}_{\text{major}}$ is the number of segregating sites in the subsample $\{1, \dots, \hat{\Gamma}\}$ (Supplemental Material, Figure S1 in File S1), and $\hat{K}_{\text{minor}} = \hat{K} - \hat{K}_{\text{major}}$, where $\hat{K}_{\text{minor}}$ and $\hat{K}_{\text{major}}$ are the observed number of mutations on minor branches and major branches, respectively. Therefore, we do not infer the ancestral and derived status of alleles to estimate K_{minor} and K_{major} , and the outgroup is only used to root the tree. Following Equations (18) and (19), $\hat{\theta}_{\text{minor}}$ and $\hat{\theta}_{\text{major}}$ can be calculated.

Finally, to test whether the data differ significantly from the expectation under the standard neutral model (one tail test), the p -value of the test is given by $P(\Gamma \geq \gamma_n, D_u \leq \hat{D}_u | \xi_1, \xi_2, \xi_{3+})$, where $\hat{D}_u$ is the estimated value of the statistic of the studied locus, and $\{\xi_1, \xi_2, \xi_{3+}\}$ is the observed compact site frequency spectrum of the locus (Li and Stephan 2005). When an outgroup is available, the compact site frequency spectrum is polarized, and ξ_1 is the number of singletons observed in the sample, ξ_2 the number of doubletons, and $\xi_{3+} = \sum_{i=3}^{n-1} \xi_i$ (Li and Stephan 2005). When an outgroup is not available, the compact site frequency spectrum is unpolarized.

The test can only be performed if the sample size is not extremely small (i.e., at least 10 diploid individuals, or the number of chromosomes $n \geq 20$) since it requires to infer an unbalanced tree (i.e., $\hat{\Gamma} \geq \gamma_n$). This restriction reduces the power because positive selection may not always produce an unbalanced tree. However, for the sake of keeping the false positive rate under control, we accept somewhat less sensitivity. We estimate the P -value of the test by generating 10^4 samples of polymorphism data under the standard neutral model, and choosing the significance level of 0.05.

Simulation data

To examine the performance of the D_u test, single-locus DNA polymorphism data were simulated in different demographic scenarios. Then, the root branching pattern of tree (Γ) was estimated by the UPGMA method to partition each simulated sample into major and minor parts. Simulations for neutrality were done according to the procedures described previously (Hudson 1991). Positive selection was assumed to be directional with codominant alleles: denoting b as the wild-type allele and B as the favored allele on the selected locus, the three genotypes, bb , bB , and BB , have the relative fitness 1, $1+s$, and $1+2s$, respectively. The dynamics of the selected locus follow a deterministic approximation for the selective stage, where the frequency of the beneficial allele increases from ε to $1 - \varepsilon$, where $\varepsilon = 1/2N$ (Kim and Stephan 2002; Li and Stephan 2006). We did not use time-forward simulations to generate the trajectory of the selected allele (Ewing and Hermisson 2010), and it is beyond the topic of this study to compare with the two trajectories.

Data availability

Maize HapMapV2 data were downloaded from the Panzea website (www.panzea.org), which consisted of 84 maize sequences (Chia *et al.* 2012). The core region (84.63–84.67 Mb) on chromosome 10 was further analyzed by the new method. This region was previously identified as a candidate of positive

Figure 3 False positive rate of the tests under (A) population size expansion, (B) bottleneck, (C) the finite island, and (D) the recent population divergence model. *DuOG* stands for the D_u test when an outgroup is available. *DuNoOG* represents the D_u test when no outgroup is available. $n = 100$, $K = 20$, and we assume no recombination within the locus. For population size expansions (E), $N_0/N_1 = 10$, where N_0 and N_1 are the current and ancestral effective population size, respectively. For bottlenecks (F), $N_0/N_1 = 100$ and $t_1 = 0.01$, where N_1 is the effective population size during the bottleneck, t_1 the duration of the bottleneck. For the finite island model (G), the number of demes is two. We assume that the sample is collected from one deme, and a MD from another. For the divergence model (H), the divergence time of the two demes is $0.5N$ generations ago, where N is the current effective population size of each deme. We assume two demes with same population size, and the sample is collected from one deme, and a MD from another.

selection on the maize genome (Tian *et al.* 2009; Lin *et al.* 2011). We also randomly selected and analyzed three chromosomal regions as neutral controls.

The software to perform the D_u test is platform independent. It requires a preinstalled standard Java environment (version 1.5 or higher). The software for calculating D_u can be freely downloaded from Zenodo (<http://doi.org/10.5281/zenodo.835226>) and our institutional website (<http://www.picb.ac.cn/evolgen/softwares/>).

Results

The distribution of D_u

We first examined the distribution of D_u under the standard neutral model without recombination (Figure S2 in File S1).

The shape of D_u distribution is slightly asymmetric, and, as expected, we have $E(D_u) \approx 0$ (Table S1 in File S1). With recombination, D_u distribution shifts to the right, while the variance becomes smaller, so D_u tends to be more positive (Figure S2 and Table S1 in File S1). This makes our test conservative. This is due to the fact that recombination breaks the region into several nonrecombined fragments. These fragments can have different genealogies. When there is no recombination, and under the infinite sites model, a genetic variant can either segregate in one of the major and minor subsample, or be monomorphic but distinct in the two subsamples. With recombination present, it could be segregating in both subsamples (Figure S1 in File S1). The genetic variant may then be counted as K_{major} . Thus, K_{major} tends to be overestimated and D_u tends to be positive

Figure 4 Powers of the tests under positive selection. *DuOG* stands for the D_u test with outgroup. *DuNoOG* for the D_u test without outgroup. (A) $n = 100$, $K = 20$. (B) $n = 100$, $K = 50$. (C) $n = 200$, $K = 20$. (D) $n = 200$, $K = 50$. We assumed that the population size is constant, $N = 10^5$. $s = 0.01$ (or $2Ns = 2000$) and $\tau = 0.0001$, where τ is the time back to the completion of the selective substitution (in units of $2N$ generations). The assumed selection coefficient represents strong selection in *Drosophila* (Li and Stephan 2006). The x-axis is the genetic distance from the selected allele to the examined neutral locus. We assume that there is no recombination within the locus.

when there is recombination within the locus. This agrees with the previous findings that recombination affects the distribution of summary statistics for detecting selection (Wall 1999).

Moreover, the infinite-sites model could be invalid due to multiple hits and homoplasy, although it is popularly used when detecting recent positive selection. We expect that the effects should be similar to recombination (Hudson 2001; McVean *et al.* 2002) and make the test more conservative, as discussed above.

False positive rates under neutrality

Population size expansion and bottleneck may result in the incorrect inference of positive selection when there is none (Nei *et al.* 1975; Watterson 1986; Tajima 1989a; Jensen *et al.* 2005). As many researchers pointed out (Akey 2009; Hermisson 2009), this problem cannot be underestimated, in particular in genome-wide scans. To investigate the robustness of the D_u test, the false positive rate is the probability that the null hypothesis (*i.e.*, the standard neutral model) is rejected when the neutral data are simulated under different demographic scenarios (Przeworski 2002; Jensen *et al.* 2005; Li 2011).

Population size expansion: We examined the robustness of the tests under recent and old population size expansions

(Figure 3E). The maximum false positive rate of Tajima's D test, Fu and Li's D test is $\sim 60\text{--}85\%$ in the cases examined (Figure 3A). It agrees with the previous findings that Tajima's D and Fu and Li's D test are sensitive to population size expansion (Tajima 1989a; Fu and Li 1993). In contrast to these tests, but similar to the MFDM test, the D_u test remains robust under all the examined expansion scenarios, no matter whether the outgroup is available or not. Fay and Wu's H test performs comparably to the D_u test in this model. We shall also point out that the false positive refers only to the case that the tests falsely identify the signal of positive selection in a neutral evolved sample. If we are interested with detecting population size expansion, the false positive rate would be indicated as power (Tajima 1989a; Fu 1997; Pluzhnikov *et al.* 2002).

Bottleneck: We examined the population bottleneck model (Figure 3F), including recent and old bottleneck scenarios. Note that the bottleneck model becomes the expansion model when the duration of the bottleneck is large. The D_u test remains robust under all the examined bottleneck scenarios, independent of the availability of an outgroup (Figure 3B). The robustness of Tajima's D test, Fu and Li's D test, and Fay and Wu's H test depends on the age of bottleneck. Generally, Fay and Wu's test is not robust for the most recent events and Tajima's D tests, Fu and Li's D for

Table 1 The presumptive signatures of positive selection on the chromosome 10 of maize

Win start	Win end	Γ	K	K_{minor}	D_u	P -value
84,631,154	84,651,154	81	185	94	-1.3858	0.0122
84,633,154	84,653,154	81	198	100	-1.3665	0.0078
84,635,154	84,655,154	81	210	108	-1.4343	0.0096
84,637,154	84,657,154	81	213	112	-1.5171	0.0070
84,639,154	84,659,154	81	234	123	-1.5190	0.0107
84,641,154	84,661,154	81	229	122	-1.5690	0.0095
84,643,154	84,663,154	81	244	131	-1.6007	0.0074
84,645,154	84,665,154	81	225	122	-1.6361	0.0080
84,647,154	84,667,154	81	220	118	-1.5934	0.0089
84,649,154	84,669,154	81	244	128	-1.5128	0.0082

$n = 84$, and the length of windows is 20 kb. Win Start: the start position of each window; Win End: the end position of each window; K : the number of segregating sites in each window; K_{minor} : the number of segregating sites on minor branches.

the intermediate age events. Additional simulations suggest that the robustness of the D_u test holds when the number of sampled chromosomes is very small (*i.e.*, 10 diploids or $n = 20$) (Figure S3 in File S1).

Moreover, the examined bottleneck scenarios are particularly severe in so far as they produce particularly large false positive rate of Tajima's D test (Tajima 1989b) and the selective sweep based approach (Li and Stephan 2005), since they enlarge the variance of different summary statistics, for example π and Watterson's θ_w . As has been noted before (Wall *et al.* 2002; Wiehe *et al.* 2007), bottlenecks with an extreme effect on the false positive rate have a severity of $t_1 N_0 / N_1 \approx 1$, where t_1 is the duration of the bottleneck, N_0 the current effective population size, and N_1 the effective population size during the bottleneck. For such range of parameter values, in some loci, all the lineages coalesce rapidly during the bottleneck, while, in other loci, some lineages survive the bottleneck and have much older MRCA with other lineages (Depaulis *et al.* 2003). Bottlenecks with such an order of severity should be carefully examined whenever determining the false positive rate of a test for recent positive selection (Xiang-Yu *et al.* 2016).

Population structure: We also examined the effects of population subdivision under the symmetric island model (Figure 3G). Population subdivision has a profound effect on tree topology (Wakeley and Aliacar 2001). It was known that when the migration rate is intermediate, a lineage may migrate from one subpopulation to another, and the migrated lineage may not coalesce with any others before the most recent common ancestor. In such a case, the tree is unbalanced. In some cases, the migrated lineage causing unbalancedness can be detected by a simple phylogenetic method with sampling of an additional individual (known as migrant-detector, MD) (Li 2011) since the MD would coalesce with the migrated lineage first before coalescing with any others. Overall, the examined population structure scenario does not affect the D_u test while Tajima's D test and Fay and Wu's H test are sensitive to this confounding factor (Figure

3C). Fu and Li's D test is also robust to most migration parameters.

Then, we examined the population divergence model (Figure 3H), which is similar to the finite island model, but the subpopulations have a recent common ancestral population. All the tests are insensitive to the examined population divergence event for different migration parameters (Figure 3D). Moreover, an incomplete MD-sampling problem may occur in practice, since population structure often remains unknown. We investigated this by assuming that MDs are only sampled from several randomly chosen demes (Figure S4 in File S1). The D_u test is not completely robust to the effect of population structure, although the false positive rate of the D_u test is generally lower than that of Fay and Wu's H test (max 0.25 vs. 0.46). When there are no MDs available, the false positive rate of the D_u test varies with migration rate. It ranges between 0.05 and 0.25, and reaches the maximized value when $4Nm = 0.4$. The false positive rate of the D_u test decreases when the number of MDs increases. The false positive rate of the D_u test remains below 0.05 when MDs are available from about one-half of the demes. The hidden population structure may also lead to an unequal sampling problem. Following the previous studies (Przeworski 2002; Li 2011), we explored this problem with a wide range of migration parameters. We found that the D_u test performed better than Fay and Wu's H test, and is robust with respect to the unequal sampling problem in the most of cases (Figure S5 in File S1).

Power

We examined the sensitivity of the D_u test under a selective sweep. Our simulations suggested that the power of the D_u test can be above 60% when $n \geq 100$ (or the number of sampled diploid individuals is ≥ 50) (Figure 4A). As expected, the power increases when sample size and/or the number of observed polymorphic sites increase (Figure 4). Generally, the power (slightly) increases when an outgroup is available. Interestingly, when recombination occurs within the locus, the power of the D_u test remains similar to that without recombination within the locus (Figure 4 and Figure S6 in File S1). This suggests that recombination may not have an adverse effect on the power of D_u much, because selection reduces the total length of the coalescent tree, and, hence, reduces the opportunity for recombination (Sabeti *et al.* 2002).

We also investigated the effect of the population size expansion and bottleneck models on the power of D_u test, including old and recent demographic events. It has sufficiently high power to identify a selective sweep even when population size varied in the past (Figure S7 in File S1). We also surveyed the effect of different τ in detecting positive selection under the constant population size model and the bottleneck model, where τ is the time back to the completion of the selective substitution (in units of $2N$ generations, where N is the effective population size) (Figure S8 in File

Figure 5 Illustration of the effects of misinferred ancestral (upper cases) and derived (lower cases) status of single nucleotide polymorphism. (A) The true case. The open circles represent two mutations occurred on the minor and major branches. (B) The case that the ancestral and derived status of alleles is misinferred. The misinference does not affect the estimation of K_{minor} and K_{major} , since $\hat{K}_{\text{minor}} = 1$ and $\hat{K}_{\text{major}} = 1$ in both (A) and (B).

S1). It is expected that the power is lowered when τ increases. Simulations suggest that the D_u test is able to detect positive selection when $\tau < 0.1$.

Robustness to model misspecification

Next, we checked the dependence of the false positive rate on background selection (Charlesworth *et al.* 1993). Similar to the previous analysis (Fu 1997; Li 2011), we considered here a partially linked gene pair that, as a pair, is subject to purifying selection, with a neutral DNA segment placed in between. Background selection typically leads to an excess of singletons, but does not affect the high frequency derived mutations (Fu 1997). As expected, the false positive rate of the D_u test under background selection remains below the significance level of the test (Figure S9 in File S1). Among the three classical statistical tests, Fay and Wu's H test is also robust to background selection.

Application to experimental data from maize

Previous studies (Tian *et al.* 2009; Lin *et al.* 2011) identified a selective sweep in a core region 83.5–86 Mb on chromosome 10 of the maize genome in a data set containing 28 sequences (Gore *et al.* 2009). It is a subset of data that we used in this study ($n = 84$) (Chia *et al.* 2012). Following these studies we conducted an analysis on the core region in the latest maize data set using the D_u test. Ten windows were analyzed, and the P -values of the test were obtained by analyzing 10^4 simulated data (Table 1). After false discovery rate (FDR) correction for multiple testing (Benjamini and Hochberg 1995), test results in all windows were significant (FDR < 0.05). The examined consecutive windows are largely overlapped, thus, the FDR correction is rather conservative. It indicates that a putative beneficial allele may have occurred near, or within, the core region, which agrees with previous studies (Tian *et al.* 2009; Lin *et al.* 2011). We also randomly selected five 20-kb regions as controls and found no inferred unbalanced trees (Table S2 in File S1). This indicates that the significant results of the core region may not be due to hidden population structure.

Discussion

Unlike the demography-adjusted tests of neutrality (Li and Stephan 2006; Rafajlović *et al.* 2014), the D_u test does not need prior information on the demography of the species. The demography can be inferred (Li and Stephan 2006; Gutenkunst *et al.* 2009; Li and Durbin 2011; Liu and Fu 2015), and used as the null hypothesis when detecting recent positive selection (Li and Stephan 2006). Or an empirical background frequency spectrum can be used for identifying sites in the genome that were affected by recent positive selection (Nielsen *et al.* 2005; DeGiorgio *et al.* 2016). This is indeed an efficient means to reduce the false positive rate, due to the confounding effect of demography, and to increase the power (Pavlidis *et al.* 2008); however, these methods usually require genome-wide polymorphism data. Thus, we have developed the single-locus approach, which is robust against the most confounding effects of demography.

Employing a tree-based approach (Li 2011; Disanto *et al.* 2013; Li and Wiehe 2013; Hunter-Zinck and Clark 2015; Ronen *et al.* 2015), we propose the D_u test for recent positive selection at a single locus. It requires DNA polymorphism data obtained from a single locus (~ 1 –50 kb chromosomal region) when randomly selected sequences (at least 10 diploid individuals) are available. It may not be suitable for genome-wide scans for positive selection because of the multiple-testing problem. Our analysis has demonstrated that it is robust to varying population size models, and most population structure scenarios examined. To further reduce the false positive rate of the population structure, we recommend to implement the MD approach (but it requires more sampling) (Li 2011), and to analyze a few randomly selected control loci, since population structure is detectable on a genome-wide scale.

Moreover, the varying read mapping coverage among lineages may affect the test. If some lineages have higher coverage than others, those lineages may falsely appear to be longer branches than others. Thus, we recommend further confirming the DNA polymorphism data within candidate regions, or that randomly chosen controls could be applied to examine this confounding issue.

Tree unbalancedness

Previously, tree unbalancedness has usually been examined for species tree (Uyenoyama 1997; Purvis and Agapow 2002; Holman 2005; Ford *et al.* 2009). Here, we focus on tree unbalancedness within a species. The main forces to produce unbalanced trees are random drift (Tajima 1983), selection (Kaplan *et al.* 1989; Fay and Wu 2000; Sibert *et al.* 2002; Kim and Nielsen 2004; Blum *et al.* 2006b), and population structure (Wakeley and Aliacar 2001; Przeworski 2002). Population structure affects the whole genome, whereas the effects of selection are always restricted to individual loci located near the selected allele.

To measure tree unbalancedness, we used the maximum value of the left and right subtree sizes of the inferred UPGMA tree, denoted by Γ . Here, Γ represents the branching pattern of the root, thus, no complete tree needs to be inferred in this study. Γ is also different with other summary statistics of tree unbalancedness, such as Colless' index C (Colless 1982), the total cophenetic index (Mir *et al.* 2013), and Matsen's index (Matsen 2006), mainly because Γ contains only limited information of the tree.

Overview of the D_u test

Our method aims at comparing the summary statistics ($\hat{\theta}_{\text{major}}$ and $\hat{\theta}_{\text{minor}}$) calculated for the two subtrees of the unbalanced tree. It is expected that the mutation rate on the two subtrees is the same, and that $\hat{\theta}_{\text{major}}$ and $\hat{\theta}_{\text{minor}}$ are comparable. We have shown that positive selection typically imposes a strong impact on branch length in one of the two subtrees, while population size change affects the branch lengths of both subtrees. Thus, D_u tends to be more negative under positive selection, while $E(D_u) \approx 0$ under neutral scenarios. Especially, an unbalanced tree is expected to occur when a neutral locus is partially linked to a beneficial allele. Therefore, a significant P of the D_u test indicates that the locus is at close genetic distance from the beneficial allele (Figure 3). Moreover, our results show that the D_u test has high power to detect positive selection in an expanded or bottlenecked population (Figure S8 in File S1).

In this study, $\hat{\theta}_{\text{major}}$ and $\hat{\theta}_{\text{minor}}$ are estimated based on the average tree, similar with Watterson's $\hat{\theta}_w$. It has been shown that $\hat{\theta}_{\text{UPBLUE}}$ is more precise than $\hat{\theta}_w$ since the former integrates the phylogenetic information (Fu 1994). Thus, the estimates of $\hat{\theta}_{\text{major}}$ and $\hat{\theta}_{\text{minor}}$ could be improved further in the future.

Moreover, to infer the branching pattern of the root (Γ), the UPGMA method may a poor estimator under certain conditions. We recommend to root the tree by an outgroup whenever it is possible. However, we also note that the performance of the D_u test depends only slightly on the availability of an outgroup in all the examined cases. This is of practical importance since an outgroup may not easily be available.

It is known that tests such as Fay and Wu's H , are sensitive to misinference of derived and ancestral variants of segregating

sites by using outgroup (Baudry and Depaulis 2003; Hernandez *et al.* 2007a), and such misinference is an annoying problem when searching for positive selection (Hernandez *et al.* 2007b). In contrast, the misinference of derived and ancestral alleles does not affect the D_u test, since $\hat{K}_{\text{minor}}$ and $\hat{K}_{\text{major}}$ remain unchanged (Figure 5).

A previous study has shown that, based on simulated data across 400 kb, the selective sweep (multiple-loci) approach has power to detect positive selection when $\tau \leq 0.2$ (Li and Stephan 2005). Another independent study demonstrated that long-haplotype based methods can detect much younger positive selection than do selective sweep approaches (Sabeti *et al.* 2006). Therefore, it is a useful and reasonable time scale ($\tau < 0.1$) that is suitable for our approach.

We applied the D_u test to a well characterized genomic region on the maize genome. The data set we used contains different populations from domesticated improved maize lines and traditional landraces (Chia *et al.* 2012; Hufford *et al.* 2012). It was proposed that maize experienced a bottleneck (Wright *et al.* 2005; Tian *et al.* 2009). Despite the complex history and population structure of the considered sample, the D_u test identified the selective sweep in the core region, but not in the randomly selected control regions.

To summarize, we have demonstrated that the D_u test not only remains robust to a wide range of parameters of demography, but also has high power to detect positive selection. It possesses improvements compared to the existing single locus methods. Nicely, all the components in the summary statistics (D_u) can be computed analytically.

Acknowledgments

We would thank Martin Lascoux for his comments. J.L., Z.Y., and H.L. were supported financially by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB13040800) and the National Natural Science Foundation of China (nos. 91531306, 91731304). T.W. acknowledges financial support by the German Research Foundation (DFG-SFB680).

Literature Cited

- Achaz, G., 2009 Frequency spectrum neutrality tests: one for all and all for one. *Genetics* 183: 249–258.
- Adron, J. R., M. W. Hahn, and B. S. Cooper, 2015 Revisiting classic clines in *Drosophila melanogaster* in the age of genomics. *Trends Genet.* 31: 434–444.
- 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., G. Zhang, K. Zhang, L. Jin, and M. D. Shriver, 2002 Interrogating a high-density SNP map for signatures of natural selection. *Genome Res.* 12: 1805–1814.
- Barton, N. H., 1998 The effect of hitchhiking on neutral genealogies. *Genet. Res.* 72: 123–133.
- Baudry, E., and F. Depaulis, 2003 Effect of misoriented sites on neutrality tests with outgroup. *Genetics* 165: 1619–1622.

- Benjamini, Y., and Y. Hochberg, 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 57: 289–300.
- Blum, M. G. B., O. Francois, and S. Janson, 2006a The mean, variance and limiting distribution of two statistics sensitive to phylogenetic tree balance. *Ann. Appl. Probab.* 16: 2195–2214.
- Blum, M. G. B., E. Heyer, O. Francois, and F. Austerlitz, 2006b Matrilineal fertility inheritance detected in hunter-gatherer populations using the imbalance of gene genealogies. *PLoS Genet.* 2: e122.
- Charlesworth, B., M. T. Morgan, and D. Charlesworth, 1993 The effect of deleterious mutations on neutral molecular variation. *Genetics* 134: 1289–1303.
- Chia, J. M., C. Song, P. J. Bradbury, D. Costich, N. de Leon *et al.*, 2012 Maize HapMap2 identifies extant variation from a genome in flux. *Nat. Genet.* 44: 803–807.
- Colless, D. H., 1982 Review of “phylogenetics: the theory and practice of phylogenetic systematics”. *Syst. Zool.* 31: 100–104.
- DeGiorgio, M., C. D. Huber, M. J. Hubisz, I. Hellmann, and R. Nielsen, 2016 SWEEPfinder2: increased sensitivity, robustness and flexibility. *Bioinformatics* 32: 1895–1897.
- Depaulis, F., S. Mousset, and M. Veuille, 2003 Power of neutrality tests to detect bottlenecks and hitchhiking. *J. Mol. Evol.* 57: S190–S200.
- Disanto, F., A. Schliz, and T. Wiehe, 2013 Yule-generated trees constrained by node imbalance. *Math. Biosci.* 246: 139–147.
- Ewing, G., and J. Hermisson, 2010 MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics* 26: 2064–2065.
- Fay, J. C., and C.-I. Wu, 2000 Hitchhiking under positive Darwinian selection. *Genetics* 155: 1405–1413.
- Ferretti, L., M. Perez-Enciso, and S. Ramos-Onsins, 2010 Optimal neutrality tests based on the frequency spectrum. *Genetics* 186: 353–365.
- Ferretti, L., A. Ledda, T. Wiehe, G. Achaz, and S. E. Ramos-Onsins, 2017 Decomposing the site frequency spectrum: the impact of tree topology on neutrality tests. *Genetics* 207: 229–240.
- Ford, D., F. A. Matsen, and T. Stadler, 2009 A method for investigating relative timing information on phylogenetic trees. *Syst. Biol.* 58: 167–183.
- Fu, W., and J. M. Akey, 2013 Selection and adaptation in the human genome. *Annu. Rev. Genomics Hum. Genet.* 14: 467–489.
- Fu, Y. X., 1994 A phylogenetic estimator of effective population size or mutation rate. *Genetics* 136: 685–692.
- Fu, Y.-X., 1997 Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* 147: 915–925.
- Fu, Y.-X., and W.-H. Li, 1993 Statistical tests of neutrality of mutations. *Genetics* 133: 693–709.
- Galtier, N., F. Depaulis, and N. H. Barton, 2000 Detecting bottlenecks and selective sweeps from DNA sequence polymorphism. *Genetics* 155: 981–987.
- Gore, M. A., J. M. Chia, R. J. Elshire, Q. Sun, E. S. Ersoz *et al.*, 2009 A first-generation haplotype map of maize. *Science* 326: 1115–1117.
- Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson, and C. D. Bustamante, 2009 Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* 5: e1000695.
- Hermisson, J., 2009 Who believes in whole genome scans for selection? *Heredity* 103: 283–284.
- Hernandez, R. D., S. H. Williamson, and C. D. Bustamante, 2007a Context dependence, ancestral misidentification, and spurious signatures of natural selection. *Mol. Biol. Evol.* 24: 1792–1800.
- Hernandez, R. D., S. H. Williamson, L. Zhu, and C. D. Bustamante, 2007b Context-dependent mutation rates may cause spurious signatures of a fixation bias favoring higher GC-content in humans. *Mol. Biol. Evol.* 24: 2196–2202.
- Holman, E. W., 2005 Nodes in phylogenetic trees: the relation between imbalance and number of descendent species. *Syst. Biol.* 54: 895–899.
- Hudson, R. R., 1991 Gene genealogies and the coalescent process, pp. 1–44 in *Oxford Surveys in Evolutionary Biology*, Vol. 7, edited by D. Futuyma, and J. Antonovics. Oxford University Press, New York.
- Hudson, R. R., 2001 Two-locus sampling distributions and their application. *Genetics* 159: 1805–1817.
- Hufford, M. B., P. Bilinski, T. Pyhajarvi, and J. Ross-Ibarra, 2012 Teosinte as a model system for population and ecological genomics. *Trends Genet.* 28: 606–615.
- Hunter-Zinck, H., and A. G. Clark, 2015 Aberrant time to most recent common ancestor as a signature of natural selection. *Mol. Biol. Evol.* 32: 2784–2797.
- Jensen, J. D., Y. Kim, V. Bauer DuMont, C. F. Aquadro, and C. D. Bustamante, 2005 Distinguishing between selective sweeps and demography using DNA polymorphism data. *Genetics* 170: 1401–1410.
- Jensen, J. D., A. Wong, and C. F. Aquadro, 2007 Approaches for identifying targets of positive selection. *Trends Genet.* 23: 568–577.
- Kaplan, N. L., R. R. Hudson, and C. H. Langley, 1989 The “hitchhiking effect” revisited. *Genetics* 123: 887–899.
- Kim, Y., and R. Nielsen, 2004 Linkage disequilibrium as a signature of selective sweeps. *Genetics* 167: 1513–1524.
- Kim, Y., and W. Stephan, 2002 Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* 160: 765–777.
- King, L., and J. Wakeley, 2016 Empirical Bayes estimation of coalescence times from nucleotide sequence data. *Genetics* 204: 249–257.
- Kirkpatrick, M., and M. Slatkin, 1993 Searching for evolutionary patterns in the shape of a phylogenetic tree. *Evolution* 47: 1171–1181.
- Li, H., and R. Durbin, 2011 Inference of human population history from individual whole-genome sequences. *Nature* 475: 493–496.
- Li, H.-P., 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, H.-P., and W. Stephan, 2005 Maximum likelihood methods for detecting recent positive selection and localizing the selected site in the genome. *Genetics* 171: 377–384.
- Li, H.-P., and W. Stephan, 2006 Inferring the demographic history and rate of adaptive substitution in *Drosophila*. *PLoS Genet.* 2: e166.
- Li, H.-P., and T. Wiehe, 2013 Coalescent tree imbalance and a simple test for selective sweeps based on microsatellite variation. *PLoS Comput. Biol.* 9: e1003060.
- Lin, K., H.-P. Li, C. Schlötterer, and A. Futschik, 2011 Distinguishing positive selection from neutral evolution: boosting the performance of summary statistics. *Genetics* 187: 229–244.
- Liu, X., and Y.-X. Fu, 2015 Exploring population size changes using SNP frequency spectra. *Nat. Genet.* 47: 555–559.
- MacCallum, C., and E. Hill, 2006 Being positive about selection. *PLoS Biol.* 4: e87.
- Matsen, F. A., 2006 A geometric approach to tree shape statistics. *Syst. Biol.* 55: 652–661.
- Maynard Smith, J., and J. Haigh, 1974 The hitch-hiking effect of a favourable gene. *Genet. Res.* 23: 23–35.
- McVean, G., P. Awadalla, and P. Fearnhead, 2002 A coalescent-based method for detecting and estimating recombination rates from gene sequences. *Genetics* 160: 1231–1241.

- Mir, A., F. Rossello, and L. Rotger, 2013 A new balance index for phylogenetic trees. *Math. Biosci.* 241: 125–136.
- Nei, M., T. Maruyama, and R. Chakraborty, 1975 The bottleneck effect and genetic variability in populations. *Evolution* 29: 1–10.
- Nielsen, R., S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark *et al.*, 2005 Genomic scans for selective sweeps using SNP data. *Genome Res.* 15: 1566–1575.
- Pavlidis, P., S. Hutter, and W. Stephan, 2008 A population genomic approach to map recent positive selection in model species. *Mol. Ecol.* 17: 3585–3598.
- Pavlidis, P., J. D. Jensen, and W. Stephan, 2010 Searching for footprints of positive selection in whole-genome SNP data from nonequilibrium populations. *Genetics* 185: 907–922.
- Pluzhnikov, A., A. Di Rienzo, and R. R. Hudson, 2002 Inferences about human demography based on multilocus analyses of Non-coding sequences. *Genetics* 161: 1209–1218.
- Przeworski, M., 2002 The signature of positive selection at randomly chosen loci. *Genetics* 160: 1179–1189.
- Purvis, A., and P.-M. Agapow, 2002 Phylogeny imbalance: taxonomic level matters. *Syst. Biol.* 51: 844–854.
- Pybus, M., P. Luisi, G. Dall'Olio, M. Uzundun, H. Laayouni *et al.*, 2015 Hierarchical boosting: a machine-learning framework to detect and classify hard selective sweeps in human populations. *Bioinformatics* 31: 3946–3952.
- Rafajlović, M., A. Klassmann, A. Eriksson, T. Wiehe, and B. Mehlig, 2014 Demography-adjusted tests of neutrality based on genome-wide SNP data. *Theor. Popul. Biol.* 95: 1–12.
- Ronen, R., N. Udpa, E. Halperin, and V. Bafna, 2013 Learning natural selection from the site frequency spectrum. *Genetics* 195: 181–193.
- Ronen, R., G. Tesler, A. Akbari, S. Zakov, N. A. Rosenberg *et al.*, 2015 Predicting carriers of ongoing selective sweeps without knowledge of the favored allele. *PLoS Genet.* 11: 27.
- Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. P. Levine, D. J. Richter *et al.*, 2002 Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419: 832–837.
- Sabeti, P. C., S. F. Schaffner, B. Fry, J. Lohmueller, P. Varilly *et al.*, 2006 Positive natural selection in the human lineage. *Science* 312: 1614–1620.
- Sabeti, P. C., P. Varilly, B. Fry, J. Lohmueller, E. Hostetter *et al.*, 2007 Genome-wide detection and characterization of positive selection in human populations. *Nature* 449: 913–919.
- Saitou, N., and M. Nei, 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4: 406–425.
- Scheinfeldt, L. B., and S. A. Tishkoff, 2013 Recent human adaptation: genomic approaches, interpretation and insights. *Nat. Rev. Genet.* 14: 692–702.
- Schrider, D. R., and A. D. Kern, 2016 S/HIC: robust identification of soft and hard sweeps using machine learning. *PLoS Genet.* 12: e1005928.
- Sibert, A., F. Austerlitz, and E. Heyer, 2002 Wright-Fisher revisited: the case of fertility correlation. *Theor. Popul. Biol.* 62: 181–197.
- Sneath, P. H., and R. R. Sokal, 1973 *Numerical Taxonomy*. W. H. Freeman and Co., San Francisco.
- Tajima, F., 1983 Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105: 437–460.
- Tajima, F., 1989a The effect of change in population size on DNA polymorphism. *Genetics* 123: 597–601.
- Tajima, F., 1989b Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123: 585–595.
- Tang, H., D. O. Siegmund, P. D. Shen, P. J. Oefner, and M. W. Feldman, 2002 Frequentist estimation of coalescence times from nucleotide sequence data using a tree-based partition. *Genetics* 161: 447–459.
- Teshima, K. M., G. Coop, and M. Przeworski, 2006 How reliable are empirical genomic scans for selective sweeps? *Genome Res.* 16: 702–712.
- Tian, F., N. M. Stevens, and E. S. Buckler, 2009 Tracking footprints of maize domestication and evidence for a massive selective sweep on chromosome 10. *Proc. Natl. Acad. Sci. USA* 106: 9979–9986.
- Uyenoyama, M. K., 1997 Genealogical structure among alleles regulating self-incompatibility in natural populations of flowering plants. *Genetics* 147: 1389–1400.
- Vitti, J. J., S. R. Grossman, and P. C. Sabeti, 2013 Detecting natural selection in genomic data. *Annu. Rev. Genet.* 47: 97–120.
- Wakeley, J., and N. Aliacar, 2001 Gene genealogies in a metapopulation. *Genetics* 159: 893–905.
- Wall, J. D., 1999 Recombination and the power of statistical tests of neutrality. *Genet. Res.* 74: 65–79.
- Wall, J. D., P. Andolfatto, and M. Przeworski, 2002 Testing models of selection and demography in *Drosophila simulans*. *Genetics* 162: 203–216.
- Watterson, G. A., 1975 On the number of segregating sites in genetic models without recombination. *Theor. Popul. Biol.* 7: 256–276.
- Watterson, G. A., 1986 The homozygosity test after a change in population size. *Genetics* 112: 899–907.
- Wiehe, T., V. Nolte, D. Živković, and C. Schlötterer, 2007 Identification of selective sweeps using a dynamically adjusted number of linked microsatellites. *Genetics* 175: 207–218.
- Wright, S. I., I. V. Bi, S. G. Schroeder, M. Yamasaki, J. F. Doebley *et al.*, 2005 The effects of artificial selection of the maize genome. *Science* 308: 1310–1314.
- Xiang-Yu, J., Z. Yang, K. Tang, and H.-P. Li, 2016 Revisiting the false positive rate in detecting recent positive selection. *Quant. Biol.* 4: 207–216.
- Zeng, K., Y.-X. Fu, S. Shi, and C.-I. Wu, 2006 Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics* 174: 1430–1439.

Communicating editor: N. Rosenberg

Appendix

Statistical Properties of L_{major} , L_{minor} and other Components of D_u .

For a general tree topology, where $\lfloor n/2 \rfloor < \Gamma \leq n-1$, the expectations and variances of major and minor branch lengths can be analytically computed. Let $x = n - \Gamma$; thus, there are $C = \binom{n-2}{n-x-1}$ possible binary vectors B_m , $m = 1, 2, \dots, C$. Each B_m represents an array of coalescent events of the tree. The expectation of L_{major} is as follows:

$$\begin{aligned} E(L_{\text{major}}) &= E(E(L_{\text{major}}|m)) \\ &= \frac{1}{C} \sum_{e_2=3}^{x+2} \sum_{e_3=e_2+1}^{x+3} \dots \sum_{e_{n-x}=e_{n-x-1}+1}^n E\left(2 \sum_{i=e_2}^{e_3-1} T_i + 3 \sum_{i=e_3}^{e_4-1} T_i + \dots + (n-x) \sum_{i=e_{n-x}}^n T_i\right) \end{aligned}$$

Following this logic, $E(L_{\text{major}})$ can then be rewritten as follows:

$$\begin{aligned} E(L_{\text{major}}) &= \frac{1}{C} \sum_{k=3}^n \sum_{i=3}^{x+2} \sum_{j=1, k-x-1 \leq j \leq k-i+1}^{n-x-1} \binom{k-i}{j-1} \binom{n-k}{n-x-j-1} (j+1) E(T_k) \\ &= \frac{1}{C} \sum_{k=3}^n d_k E(T_k) = \frac{1}{C} \sum_{k=3}^n \frac{2d_k}{k(k-1)} \end{aligned}$$

where $d_k = \sum_{i=3}^{x+2} \sum_{j=1, k-x-1 \leq j \leq k-i+1}^{n-x-1} \binom{k-i}{j-1} \binom{n-k}{n-x-j-1} (j+1)$, and k denotes the number of lineages (counting both the major and the minor subtrees) during the coalescent time period T_k , and i the number of lineages (counting both the major and the minor subtrees) at a time right before the most recent common ancestor for the major subsample, $(j+1)$ the number of lineages in the major subtree during the coalescent time period T_k (Figure 2). As one of components in d_k , $\binom{n-k}{n-x-j-1}$ means that there are $(n-k)$ coalescent events before the time period T_k , among which $(n-x-j-1)$ events occur in the major subtree, and $\binom{k-i}{j-1}$ indicates that there are $(k-i)$ coalescent events between the time period T_k and the time of the most recent common ancestor for the major subsample, among which $(j-1)$ events occur in the major subtree.

Similar to the derivation of $E(L_{\text{major}})$, we can derive the variance of L_{major} and L_{minor} as follows:

$$\begin{aligned}
\text{Var}(L_{\text{major}}) &= E(L_{\text{major}} - E(L_{\text{major}}))^2 = E\left(E\left((L_{\text{major}} - E(L_{\text{major}}))^2 \middle| m\right)\right) \\
&= \frac{1}{C} \sum_{m=1}^{n_T} E\left((L_{\text{major}} - E(L_{\text{major}}))^2 \middle| m\right) \\
&= \frac{1}{C} \sum_{m=1}^{n_T} (E(L_{\text{major}}^2 | m) - 2E(L_{\text{major}} | m)E(L_{\text{major}}) + E^2(L_{\text{major}})) \\
&= \frac{1}{C} \sum_{m=1}^{n_T} E(L_{\text{major}}^2 | m) - 2E(L_{\text{major}}) \frac{1}{C} \sum_{m=1}^{n_T} E(L_{\text{major}} | m) + \frac{1}{C} \sum_{m=1}^{n_T} E^2(L_{\text{major}}) \\
&= \frac{1}{C} \sum_{m=1}^{n_T} E(L_{\text{major}}^2 | m) - E^2(L_{\text{major}}) \\
&= \frac{1}{C} \sum_{e_2=3}^{x+2} \sum_{e_3=e_2+1}^{x+3} \cdots \sum_{e_{n-x}=e_{n-x-1}+1}^n E\left(2 \sum_{i=e_2}^{e_3-1} T_i + 3 \sum_{i=e_3}^{e_4-1} T_i + \cdots + (n-x) \sum_{i=e_{n-x}}^n T_i\right)^2 - E^2(L_{\text{major}}) \\
&= \frac{1}{C} \sum_{e_2=3}^{x+2} \sum_{e_3=e_2+1}^{x+3} \cdots \sum_{e_{n-x}=e_{n-x-1}+1}^n \left(\sum_{i=e_2}^n a_i^2 E(T_i^2) + \sum_{i \neq j, e_2 < i, j \leq n} a_i a_j E(T_i) E(T_j) \right) - E^2(L_{\text{major}}) \\
&= \frac{1}{C} \sum_{e_2=3}^{x+2} \sum_{e_3=e_2+1}^{x+3} \cdots \sum_{e_{n-x}=e_{n-x-1}+1}^n \left(\sum_{i=e_2}^n \frac{8a_i^2}{i^2(i-1)^2} + \sum_{i \neq j, e_2 < i, j \leq n} \frac{4a_i a_j}{ij(i-1)(j-1)} \right) - E^2(L_{\text{major}}) \\
&= \frac{1}{C} \sum_{e_2=3}^{x+2} \sum_{e_3=e_2+1}^{x+3} \cdots \sum_{e_{n-x}=e_{n-x-1}+1}^n \left(\sum_{i=e_2}^n \sum_{j=i}^n \frac{8a_i a_j}{ij(i-1)(j-1)} \right) - E^2(L_{\text{major}})
\end{aligned}$$

where $a_i = \begin{cases} k, & e_k \leq i < e_{k+1}, k \in \{2, 3, \dots, n-x\} \text{ and } e_{n-x+1} = n+1. \\ 0, & \text{otherwise} \end{cases}$

$$\begin{aligned}
\text{Var}(L_{\text{minor}}) &= \frac{1}{C} \sum_{m=1}^{n_T} E(L_{\text{minor}}^2 | m) - E^2(L_{\text{minor}}) \\
&= \frac{1}{C} \sum_{e_2=3}^{x+2} \sum_{e_3=e_2+1}^{x+3} \cdots \sum_{e_{n-x}=e_{n-x-1}+1}^n E\left(\sum_{i=2}^{e_2-1} iT_i + \sum_{i=e_2}^{e_3-1} (i-2)T_i + \sum_{i=e_3}^{e_4-1} (i-3)T_i + \cdots + \sum_{i=e_{n-x}}^n (i-n+x)T_i\right)^2 - E^2(L_{\text{minor}}) \\
&= \frac{1}{C} \sum_{e_2=3}^{x+2} \sum_{e_3=e_2+1}^{x+3} \cdots \sum_{e_{n-x}=e_{n-x-1}+1}^n \left(\sum_{i=2}^n b_i^2 E(T_i^2) + \sum_{i \neq j, 2 < i, j \leq n} b_i b_j E(T_i) E(T_j) \right) - E^2(L_{\text{minor}}) \\
&= \frac{1}{C} \sum_{e_2=3}^{x+2} \sum_{e_3=e_2+1}^{x+3} \cdots \sum_{e_{n-x}=e_{n-x-1}+1}^n \left(\sum_{i=2}^n \frac{8b_i^2}{i^2(i-1)^2} + \sum_{i \neq j, 2 < i, j \leq n} \frac{4b_i b_j}{ij(i-1)(j-1)} \right) - E^2(L_{\text{minor}}) \\
&= \frac{1}{C} \sum_{e_2=3}^{x+2} \sum_{e_3=e_2+1}^{x+3} \cdots \sum_{e_{n-x}=e_{n-x-1}+1}^n \left(\sum_{i=2}^n \sum_{j=i}^n \frac{8b_i b_j}{ij(i-1)(j-1)} \right) - E^2(L_{\text{minor}})
\end{aligned}$$

where b_i is determined by:

$$b_i = \begin{cases} i, & 2 \leq i < e_2 \\ i-k, & e_k \leq i < e_{k+1} \end{cases},$$

where $2 \leq k \leq n-x$, $e_{n-x+1} = n+1$.

Since

$$\begin{aligned}
E(K_{\text{major}}K_{\text{minor}}) &= E\left[E(K_{\text{major}}K_{\text{minor}}|L_{\text{major}}L_{\text{minor}})\right] \\
&= E\left[E(K_{\text{major}}|L_{\text{major}})E(K_{\text{minor}}|L_{\text{minor}})\right] \\
&= E\left(\frac{\theta}{2}L_{\text{major}}\frac{\theta}{2}L_{\text{minor}}\right) = \frac{\theta^2}{4}E(L_{\text{major}}L_{\text{minor}}),
\end{aligned}$$

we have

$$\begin{aligned}
\text{Cov}(\hat{\theta}_{\text{major}}, \hat{\theta}_{\text{minor}}) &= \text{Cov}\left(\frac{2K_{\text{major}}}{E(L_{\text{major}})}, \frac{2K_{\text{minor}}}{E(L_{\text{minor}})}\right) = \frac{4\text{Cov}(K_{\text{major}}, K_{\text{minor}})}{E(L_{\text{major}})E(L_{\text{minor}})} \\
&= \frac{4\left(E(K_{\text{major}}K_{\text{minor}}) - E(K_{\text{major}})E(K_{\text{minor}})\right)}{E(L_{\text{major}})E(L_{\text{minor}})} \\
&= \frac{4\left(\frac{\theta^2}{4}E(L_{\text{major}}L_{\text{minor}}) - \frac{\theta}{2}E(L_{\text{major}})\frac{\theta}{2}E(L_{\text{minor}})\right)}{E(L_{\text{major}})E(L_{\text{minor}})} \\
&= \theta^2 \frac{\text{Cov}(L_{\text{major}}L_{\text{minor}})}{E(L_{\text{major}})E(L_{\text{minor}})}
\end{aligned}$$