Inferring the Probability of the Derived *vs.* the Ancestral Allelic State at a Polymorphic Site

Peter D. Keightley[1] and Benjamin C. Jackson
Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, EH9 3FL, United Kingdom

ABSTRACT It is known that the allele ancestral to the variation at a polymorphic site cannot be assigned with certainty, and that the most frequently used method to assign the ancestral state—maximum parsimony—is prone to misinference. Estimates of counts of sites that have a certain number of copies of the derived allele in a sample (the unfolded site frequency spectrum, uSFS) made by parsimony are therefore also biased. We previously developed a maximum likelihood method to estimate the uSFS for a focal species using information from two outgroups while assuming simple models of nucleotide substitution. Here, we extend this approach to allow multiple outgroups (implemented for three outgroups), potentially any phylogenetic tree topology, and more complex models of nucleotide substitution. We find, however, that two outgroups and the Kimura two-parameter model are adequate for uSFS inference in most cases. We show that using parsimony to infer the ancestral state at a specific site seriously breaks down in two situations. The first is where the outgroups provide no information about the ancestral state of variation in the focal species. In this case, nucleotide variation will be underestimated if such sites are excluded. The second is where the minor allele in the focal species agrees with the allelic state of the outgroups. In this situation, parsimony tends to overestimate the probability of the major allele being derived, because it fails to account for the fact that sites with a high frequency of the derived allele tend to be rare. We present a method that corrects this deficiency and is capable of providing nearly unbiased estimates of ancestral state probabilities on a site-by-site basis and the uSFS.

KEYWORDS nucleotide polymorphism; ancestral allele; derived allele; unfolded site frequency spectrum; parsimony; misinference

MANY population genetic and quantitative genetic analysis methods require the assignment of ancestral *vs.* derived states at polymorphic nucleotide sites. For example, Fay and Wu (2000) and Zeng *et al.* (2006) proposed statistics, *H* and *E*, that compare the numbers of high, intermediate, and low frequency derived variants, which can then be used to distinguish between different modes of natural selection and demographic change. A number of methods have also been developed to infer selection and demographic change based on the complete distribution of counts of derived alleles across sites, the unfolded site frequency spectrum (uSFS) (*e.g.*, Boyko *et al.* 2008; Schneider *et al.* 2011; Tataru *et al.* 2017).

The minor allele at a site or counts of numbers of minor alleles at a group of sites (the folded site frequency spectrum) can be observed directly from sequence polymorphism data. In contrast, the derived *vs.* the ancestral allele at a site cannot be known with certainty, because at least one outgroup is required for inference, and there is the possibility of more than one mutation separating the focal species from the outgroup. This also implies that the uSFS cannot be known precisely. For the purpose of ancestral state inference, rule-based maximum parsimony is the most frequently applied method in molecular evolutionary genetics (*e.g.*, Voight *et al.* 2006; Dreszer *et al.* 2007; Keinan *et al.* 2007; Sabeti *et al.* 2007; 1000 Genomes Project Consortium 2010, 2015; Lohse and Barton 2011; Langley *et al.* 2012; Schmidt *et al.* 2017). It has been recognized, however, that parsimony potentially produces misleading results (Felsenstein 1981; Collins *et al.* 1994; Eyre-Walker 1998). Of particular relevance here is that sites that have a low frequency of the derived allele are

doi: https://doi.org/10.1534/genetics.118.301120
Manuscript received March 23, 2018; accepted for publication May 14, 2018; published Early Online May 16, 2018.
Available freely online through the author-supported open access option.

Supplemental material available at Figshare: https://doi.org/10.25386/genetics.6275915.
[1]Corresponding author: Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Charlotte Auerbach Rd., Edinburgh EH9 3FL, United Kingdom. E-mail: peter.keightley@ed.ac.uk

usually more common than sites that have a high frequency of the derived allele. This implies that misinference tends to upwardly bias counts of high frequency derived alleles (Baudry and Depaulis 2003; Hernandez *et al.* 2007).

There is a related problem concerning the assignment of ancestral states, which does not seem to have been addressed. If ancestral states are assigned on site-by-site basis, potentially useful information is ignored. For example, consider the case of a single outgroup species that is uninformative about the ancestral allele of the variation in a focal species at a site. It is more likely, however, that the ancestral allele at the site is the low frequency allele, if sites with a high frequency of the derived allele are uncommon in the data set as a whole (as is usually the case).

Matsumoto *et al.* (2015) pointed out that ancestral states are not observable, that a single best ancestral reconstruction is not advisable, and that assuming one can bias molecular evolutionary inference. This was developed by Jackson *et al.* (2017), who assigned the ancestral state probability at a site as the inferred probability of the node for the common ancestor of the focal species and the closest outgroup, obtained using PAML (Yang 2007), while ignoring polymorphism data. However, this does not optimally weight information coming from the focal site itself and from the data as a whole.

Inference of ancestral states on a site-by-site basis has been problematic, but there has been progress in inferring the uSFS. Hernandez *et al.* (2007) developed a context-dependent substitution model using a single outgroup to infer the ancestral state at a polymorphic site in a focal species, and then implemented a step to correct for ancestral misidentification. Their simulations suggested, however, that the approach only partially corrects for ancestral misidentification, depending on the divergence between the focal species and the outgroup.

Schneider *et al.* (2011) developed a probabilistic method to infer the uSFS on a site-by-site basis, but did not use information from the frequencies of polymorphisms across all sites, so results from this method are biased. Keightley *et al.* (2016) developed a maximum likelihood (ML) method that addresses the deficiency in Schneider *et al.* (2011) and simulations suggested that it is capable of correctly inferring the uSFS. It uses a two-stage process in which the evolutionary rates are estimated by ML and then, assuming the rates, estimates the uSFS elements by ML, while correctly weighting information from informative and uninformative sites. However, the method is limited to two outgroups, assumes simple substitution models [for one outgroup, the Kimura two-parameter (K2) model; for two outgroups, the Jukes–Cantor (JC) model], and is not readily scalable to more than two outgroups or to more complex substitution models. It is unknown whether more realistic substitution models and/or additional outgroups significantly improves inference accuracy. Furthermore, it does not assign ancestral state probabilities on a site-by-site basis.

In this article, we develop the approach of Keightley *et al.* (2016), with the following objectives: (1) estimate the uSFS, allowing several outgroups, potentially any tree topology, and more realistic nucleotide substitution models; and (2) infer ancestral state probabilities for each polymorphic site in the data. We evaluate the performance of the new approach by simulations, apply it to data from the *Drosophila* Population Genomics Project (DPGP) as a test case, and reinfer the ancestral state probabilities for a population of the 1000 Genomes Project in humans, which were previously inferred by a parsimony-related approach.

Materials and Methods

Following Keightley *et al.* (2016), uSFS inference is carried out in two-steps. Evolutionary rate parameters are estimated from all sites in the data (including polymorphic and monomorphic sites) in step 1. In step 2 the uSFS is computed, conditional on the evolutionary rate parameter estimates. Information from steps 1 and 2 is then combined in a third step to infer the ancestral state probability for each polymorphic site.

Representation of the data and some definitions

Suppose we have sampled m orthologous gene copies at a set of sites from a population of a focal species. The uSFS we require to estimate therefore contains $m - 1$ elements, excluding the elements where the ancestral or derived allele is fixed. We assume that we have randomly sampled a single gene copy at each site in one or more outgroup species. We assume that the tree topology relating the species is known and does not vary among sites (Figure 1). In the analysis we assume that the nucleotide variation within the focal species coalesces within the branch labeled b_1. The consequences of polymorphism in the outgroup species and violation of the assumptions of an invariant tree topology and coalescence within branch b_1 are investigated in simulations. The observed nucleotide configuration for a site is the count of each of the four nucleotides in the focal species (labeled X, Y for a biallelic site), along with the state for each outgroup (A, C, G, or T). Let the number of outgroups $= n$ (in Figure 1, $n = 3$), and denote the outgroups o_1, o_2...o_n. Assuming an unrooted tree (as in Figure 1), the number of branches in the tree is therefore $b = 2n - 1$.

Models of nucleotide substitution

The JC model, K2 model, and a model allowing six symmetrical rates (R6; Figure 2) are considered. All substitution models require the estimation of evolutionary rates (*i.e.*, mean number of nucleotide changes per site) for each branch, K_1...K_b. The rates are the only parameters for the JC model. For the K2 model, an additional parameter, κ, specifies the rate of transition mutations relative to the rate of transversions. For the R6 model, there are six symmetrical relative mutation rates, r_1...r_6, $\sum_{i=1}^{6} r_i = 1$(Figure 2), so five independent parameters, r_1...r_5, require to be estimated.

Estimation of rate parameters

Assuming the tree topology of Figure 1, there are b substitution rates and these, along with parameters of the substitution

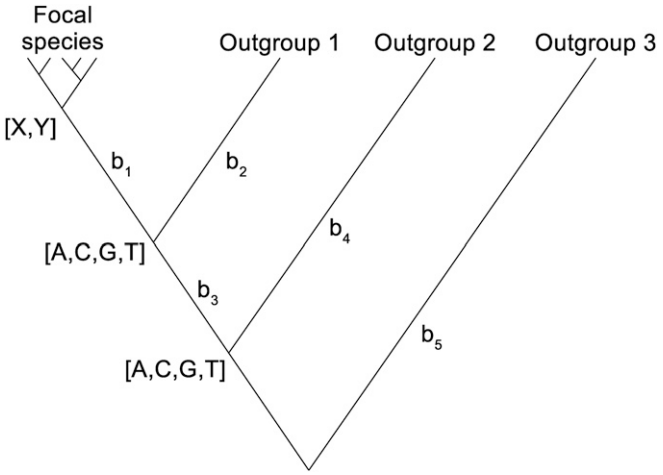

Figure 1 Representation of the data for uSFS and ancestral state inference. Polymorphism within the focal species (nucleotides X, Y) is assumed to coalesce within branch b_1. There are three outgroups, two unknown internal nodes, and five branches in this tree. The root of the tree is not identifiable, therefore branch b_5 extends from outgroup 3 to the node of b_3 and b_4.

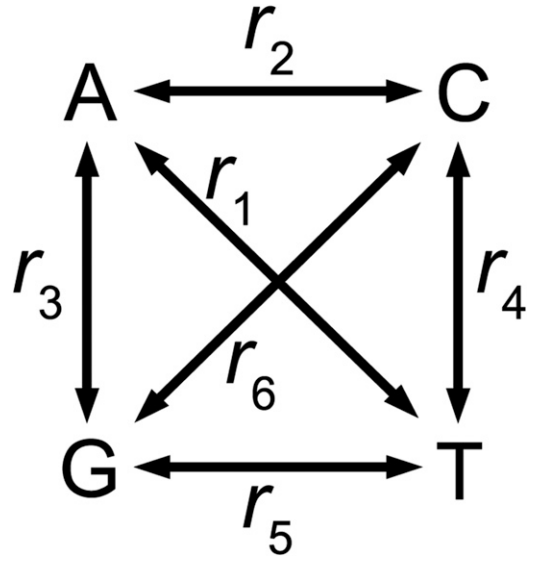

Figure 2 The R6 model.

model (*i.e.*, κ for the K2 model or $r_1 \ldots r_5$ for the R6 model), are estimated by ML using the simplex algorithm for likelihood maximization. We checked convergence by picking starting values for the parameters from wide distributions, restarting the algorithm when convergence had apparently been achieved, and checking that the same final maximum log likelihood was reached in multiple runs. Let $\boldsymbol{\varphi}$ be a vector specifying the model parameters, and let $\mathbf{y}_i$ be a vector specifying the observed nucleotide configuration for the focal species and the outgroups at site i. Sites are assumed to evolve independently, so the overall likelihood of the data is the product of probabilities of the observed nucleotide configuration for each site:

$$L = \prod_{i=1}^{\text{sites}} p(\mathbf{y_i}|\boldsymbol{\varphi}). \tag{1}$$

The probability of the nucleotide configuration for each site is evaluated by summing the probabilities for the $n_{\text{tree}} = 4^{n-1}$ possible unrooted trees, formed from all possible nucleotide combinations [A, T, G, C] at the unknown internal nodes along with the observed nucleotide configuration for the focal species and outgroups at the site.

$$p(\mathbf{y_i}|\boldsymbol{\varphi}) = \sum_{j=1}^{n_{\text{tree}}} p_{\text{tree}}(\mathbf{c_j}|\boldsymbol{\varphi}), \tag{2}$$

where $\mathbf{c_j}$ is a vector representing the observed nucleotide configuration for the focal species and the n outgroups along with the nucleotide states for the $b-1$ internal nodes for tree j. If the focal species is polymorphic at a site, the probability for that site is computed as the average probability for each observed nucleotide (X, Y in Figure 1).

The overall probability for a given tree is computed from the product of the probabilities of each branch ($k = 1 \ldots b$), conditional on the nucleotide states $x_{1,k}$ and $x_{2,k}$ representing the ancestral and derived nucleotides of that branch, given the nucleotide states specified in $\mathbf{c_j}$:

$$p_{\text{tree}}(\mathbf{c_j}|\boldsymbol{\varphi}) = \prod_{k=1}^{n_B} p_{\text{branch}}(x_{1,k}, x_{2,k}|\boldsymbol{\varphi}). \tag{3}$$

The probability for a branch depends on whether $x_{1,k}$ and $x_{2,k}$ differ from one another, the type of any difference (except in the case of the JC model), and the substitution rate parameters $\boldsymbol{\varphi}$.

Computation of p_{branch}

In computing the probability of observing nucleotides $x_{1,k}$ and $x_{2,k}$ on branch k, it is assumed that the number of nucleotide changes on the branch is Poisson distributed. Terms for more than two changes on a branch are disregarded. The method could be extended to allow more than two changes on a branch, but highly saturated sites would contribute little useful information. Let K_k be the evolutionary rate parameter for branch k, which is the mean number of changes for that branch.

JC model:

$$1.\ x_{1,k} = x_{2,k} : p_{\text{branch}} = \exp(-K_k) + \frac{1}{6}K_k^2 \exp(-K_k) \tag{4}$$

$$2.\ x_{1,k} \neq x_{2,k} : p_{\text{branch}} = \frac{1}{3}K_k \exp(-K_k) + \frac{1}{9}K_k^2 \exp(-K_k) \tag{5}$$

K2 model:

1. $x_{1,k} = x_{2,k}$: p_{branch}

$$= \exp(-K_k) + \frac{1}{2}K_k^2 \exp(-K_k)\frac{2+\kappa^2}{\kappa^2+4\kappa+4} \quad (6)$$

2. $x_{1,k} \neq x_{2,k}$, transition change:

$$p_{\text{branch}} = K_k \exp(-K_k)\frac{\kappa}{\kappa+2} + K_k^2 \exp(-K_k)\frac{1}{\kappa^2+4\kappa+4} \quad (7)$$

3. $x_{1,k} \neq x_{2,k}$, transversion change:

$$p_{\text{branch}} = K_k \exp(-K_k)\frac{1}{\kappa+2} + K_k^2 \exp(-K_k)\frac{\kappa}{\kappa^2+4\kappa+4} \quad (8)$$

R6 model (Figure 2):

1. $x_{1,k} = x_{2,k}$: $p_{\text{branch}} = p(0 \text{ changes}) + p(2 \text{ changes})$ (9)

Taking the example of $x_{1,k} = x_{2,k} = \text{A}$:

$$p(0 \text{ changes}) = \exp[-2K_k(r_1 + r_2 + r_3)] \quad (10)$$

Note that r_1, r_2, and r_3 are the relative rates for changes involving base A.

For $p(2 \text{ changes})$: The algorithm to compute the probability of observing the same ancestral and derived base when two changes have occurred on a branch is illustrated by a simplified example where all relative rates in the model apart from two (r_1 and r_4) are zero (Figure 2).

For the case of $x_{1,k} = x_{2,k} = \text{A}$, the sequence of events must therefore be an A → T change followed by a T → A change. The probability of these events is obtained from:

$$\int_0^1 p(\text{no mutation to time } y) \cdot p(\text{A}\rightarrow\text{T mutation}) \cdot p(\text{T}\rightarrow\text{A mutation between time } y \text{ and } 1)dy. \quad (11)$$

For the example where all relative rates in the model apart from r_1 and r_4 are zero, this is:

$$\frac{r_1^2}{(r_1+r_2+r_3)(r_1+r_4+r_5)} \int_0^1 \exp(-k_1 y)k_1(1-y)k_2 \times \exp[-k_2(1-y)]dy, \quad (12)$$

where $k_1 = 2K_k(r_1 + r_2 + r_3)$ and $k_2 = 2K_k(r_1 + r_4 + r_5)$. In this example, the relative rates r_2, r_3, and r_5 are all zero, but are included for completeness. Evaluation of the definite integral in (12) gives a closed form expression:

$$\frac{k_1 k_2 \exp(-k_2-k_1)[\exp(k_2) - \exp(k_1)k_2 + (k_1-1)\exp(k_1)]}{k_2^2 - 2k_1k_2 + k_1^2}. \quad (13)$$

The logic can be extended to allow all the relative rates to be nonzero.

2. $x_{1,k} \neq x_{2,k}$: $p_{\text{branch}} = p(1 \text{ change}) + p(2 \text{ changes})$

$p(1 \text{ change})$: Examine the example $x_{1,k} = \text{A}$, $x_{2,k} = \text{T}$.

$$p(1 \text{ change}) = K_k r_1 \{\exp[-2K_k(r_1+r_2+r_3)] + \exp[-2K_k(r_1+r_4+r_5)]\} \quad (14)$$

$p(2 \text{ changes})$: Examine the example $x_{1,k} = \text{A}$, $b_{2,k} = \text{C}$.

Assume that only r_1 and r_4 are nonzero (Figure 2), and that A is the ancestral base and C is the derived base. The sequence of events is therefore an A → T change followed by a T → C change. The probability of this event sequence is obtained from:

$$\int_0^1 p(\text{no mutation to time } y) \cdot p(\text{A}\rightarrow\text{T mutation}) \cdot p(\text{T}\rightarrow\text{C mutation between time } y \text{ and } 1)dy. \quad (15)$$

This is:

$$\frac{r_1 r_4}{(r_1+r_2+r_3)(r_1+r_4+r_5)} \int_0^1 \exp(-k_1 y)k_1(1-y)k_2 \times \exp[-k_2(1-y)]dy, \quad (16)$$

where k_1 and k_2 have the same meanings as above.

The algorithm can be extended to cases where the relative rates are all nonzero.

Computing uSFS elements

The ML approach described by Keightley *et al.* (2016) estimates the proportion of density, π_j, attributable to the major allele being the ancestral allele *vs.* the major allele being the derived allele for each uSFS element pair (indexed by j and $m-j$, where m is the number of gene copies sampled). We implemented this algorithm as follows, conditional on the ML estimate of the rate parameters, $\hat{\boldsymbol{\varphi}}$ (obtained by evaluating Equation 1), which are therefore assumed to be known without error. For a uSFS containing m elements, $m/2$ ML estimates require to be made. Assuming sites evolve independently (*cf.* Equation 1), the likelihood of π_j for the subset of sites (numbering sites_j) having j copies of the minor allele in the focal species is:

$$L(\pi_j) = \prod_{i=1}^{\text{sites}_j} [p(\mathbf{y_{i,1}}|\hat{\boldsymbol{\varphi}})\pi_j + p(\mathbf{y_{i,2}}|\hat{\boldsymbol{\varphi}})(1-\pi_j)], \quad (17)$$

where the probability of the observed nucleotide configuration for the focal species and the outgroups at the site is given by Equation 2, evaluated with the major allele $[p(\mathbf{y_{i,1}}|\hat{\boldsymbol{\varphi}})]$ and

the minor allele $[p(\mathbf{y}_{\mathbf{i},\mathbf{2}}|\hat{\boldsymbol{\varphi}})]$ as the state of the focal species at that site (see Figure 1).

Computing ancestral state probabilities on a site-by-site basis

The probability of allele X_i *vs.* allele Y_i being ancestral at site i could be computed from their relative probabilities, *i.e.*, $p_1 = p(\mathbf{y}_{\mathbf{i},\mathbf{1}}|\hat{\boldsymbol{\varphi}})$ and $p_2 = p(\mathbf{y}_{\mathbf{i},\mathbf{2}}|\hat{\boldsymbol{\varphi}})$,but this only uses information from the estimated rate parameters. It does not incorporate information from the number of major *vs.* minor copies at the site. For example, if the outgroup information were uninformative, we would assign $p_1 = p_2$. If there are few sites in the data set as a whole where the derived allele is at a high frequency, however, the estimated uSFS would tell us that A is more likely to be ancestral.

To infer the ancestral state probabilities for site i, information from the estimated rate parameters is augmented by the nearly independent information from the estimated uSFS (*cf.* Halligan *et al.* 2013). If there are j copies of the minor allele in the focal species at a site i, the probability of the major allele X_i being ancestral is:

$$p(\mathrm{X}_i = \text{ancestral}) = \frac{p_1 \widehat{\pi}_j}{p_1 \widehat{\pi}_j + p_2(1 - \widehat{\pi}_j)}. \qquad (18)$$

As a check on this equation, it can be shown that the sums of the ancestral state probabilities recovers the estimated uSFS.

Simulations

We extended a simulation program described by Keightley *et al.* (2016) to simulate three outgroups for the topology illustrated in Figure 1. Briefly, unlinked sites with four nucleotide states were simulated in a diploid population of size $N = 100$. The mutation rate per site per generation was set to $\mu = \theta/N$, and the neutral genetic diversity, θ, was typically 0.01. The simulations allowed any variation within a population at a node of the phylogenetic tree to be passed to two ancestral subpopulations, which were formed by sampling chromosomes with replacement in one generation. To generate the data for uSFS inference, a single gene copy was randomly sampled from each outgroup species. We either simulated neutral sites, or a mixture of neutral and selectively constrained sites. If a mutation occurred at a selectively constrained site, its selection coefficient was $s/2$, where s is the difference in fitness between the homozygous mutant and the heterozygote. Fitness effects were multiplicative between and within loci.

DPGP data

We analyzed fourfold degenerate sites from the Rwandan sequences of the DPGP phase 2 data, comprising 17 haploid genomes (see Keightley *et al.* 2016 for details).

1000 Genomes data

We downloaded variant calls from the phase 3 release of the 1000 Genomes Project (from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/) and extracted the 99 unrelated individuals from the Luhya in Webuye, Kenya (henceforth LWK) population. First, we restricted our analyses to sites that were fourfold degenerate in all autosomal transcripts of protein-coding genes in humans according to Ensembl release 71. We used the six-way EPO multiple alignments of primate species (available from ftp://ftp.ensembl.org/pub/release-71/emf/ensembl-compara/epo_6_primate/) to determine the alleles in orangutans and macaques at each fourfold degenerate site, and to determine whether those sites were within a CpG in humans or either of the outgroup species. We used orangutan and macaque as outgroups in our analysis. Chimpanzee and gorilla are closer and potentially more informative, but they share a high proportion of polymorphism with human and this violates an assumption of our analysis. The EPO multiple alignments were first converted from .emf format to .maf format, and then specific regions were accessed using the WGAbed package (https://henryjuho.github.io/WGAbed/). The data for the human ancestral alleles, as used by the 1000 Genomes Project (1000 Genomes Project Consortium 2015), were downloaded from ftp://ftp.ensembl.org/pub/release-74/fasta/ancestral_alleles/.

Sites were retained for analysis if there was no missing data in humans or either outgroup species. Sites were further assigned to CpG and non-CpG categories. CpG sites were defined as sites that were CpG in their context in any of the three species: human (including both REF and ALT alleles), orangutan, or macaque. Non-CpG sites were defined as sites that were never CpG in their context in any of the same species, including both REF and ALT alleles in the human sample. Alleles at polymorphic sites were used to populate the uSFS following two methods: (1) using the ancestral allele provided by the 1000 Genomes Project to polarize derived and ancestral variants, and (2) using the ML method described in the present study.

Data availability statement

Software are available for download from https://sourceforge.net/projects/est-usfs/. Supplemental material available at Figshare: https://doi.org/10.25386/genetics.6275915.

Results

Simulation results

The uSFS inference method allows several outgroups to be included, but the extent of any benefit from additional outgroups has been unknown. To investigate this, we simulated unlinked sites according to the tree topology shown in Figure 1 with three outgroups, recorded the "true" uSFS, and compared it to uSFSs estimated using one, two, or three outgroups. High derived allele frequency uSFS elements are expected to be most affected by misinference (Baudry and Depaulis 2003; Keightley *et al.* 2016), so we focused on the last uSFS element (*e.g.*, element 19 of a 20-element uSFS). Our measures of bias and accuracy were the average deviation and root-mean-squared error (RMSE) for this element.

For the case of neutrally evolving sites, if data are simulated and analyzed under the JC model, there is a small amount of

negative bias affecting the last uSFS element (*i.e.*, high frequency uSFS elements tend to be slightly underestimated; Figure 3A). The cause of this negative bias is unknown, but it could be a consequence of violation of any of the assumptions described in the *Material and Methods*. RMSE is reduced somewhat if a second outgroup is added, but there is little benefit from adding a third outgroup (Figure 3B). If data are simulated including transition:transversion bias and the analysis is by the JC model, the last uSFS element is substantially overestimated (Supplemental Material, Figure S1). If the K2 or R6 models are used, however, only a small amount of bias is observed (Figure S1). As expected, parsimony-based inference seriously overestimates the frequency of high frequency derived alleles in all cases (Figure 3A and Figure S1). Parsimony does not provide ancestral state probabilities per site, because it assigns an allele as derived or ancestral with certainty. Parsimony will therefore be potentially seriously biased compared to computing ancestral state probabilities using Equation 18.

We then investigated accuracy and bias for the case of a fraction of sites subject to moderate purifying selection (scaled selection strength $Ns = 10$). This is relevant for inferring the uSFS for nonneutral sites, such as nonsynonymous sites of protein-coding genes, and for cases where there is variation in the mutation rate among sites, leading to variation in the rate of substitution. Such variation violates an assumption of the uSFS inference method and is therefore expected to cause the method to break down to some extent. As we previously observed (Keightley *et al.* 2016), the presence of variation in the rate of substitution leads to overestimation of high derived allele frequency uSFS elements (Figure 4A). The bias can be serious if there is only one outgroup, but is reduced if a second outgroup is included. However, there is only a small additional benefit from adding a third outgroup. Variation about the observed values is lower, on average, if additional outgroups are included (*i.e.*, RMSE is lower; Figure 4B), but again adding a third outgroup is of little benefit. As expected, parsimony performs poorly, overestimating the high frequency derived allele frequency.

Analysis of DPGP phase 2 data

To assess the performance of the uSFS inference procedure in a more realistic situation, we analyzed fourfold degenerate sites from the Rwandan sequences of the DPGP phase 2, which comprises 17 haploid genomes (provided by J. Campos). We compared the inferred uSFSs obtained using *Drosophila simulans* as the sole outgroup and using both *D. simulans* and *D. yakuba* as outgroups, and investigated the consequences of increasing the complexity of the substitution model. More complex substitution models fit the data much better (Table 1), largely driven by the approximately twofold transition: transversion mutation bias captured by the K2 model.

Although different nucleotide substitution models produce large differences in log likelihood, the estimated uSFS is appreciably different only between the JC and K2 models, and it is indistinguishable between the K2 and R6 models (Figure 5A).

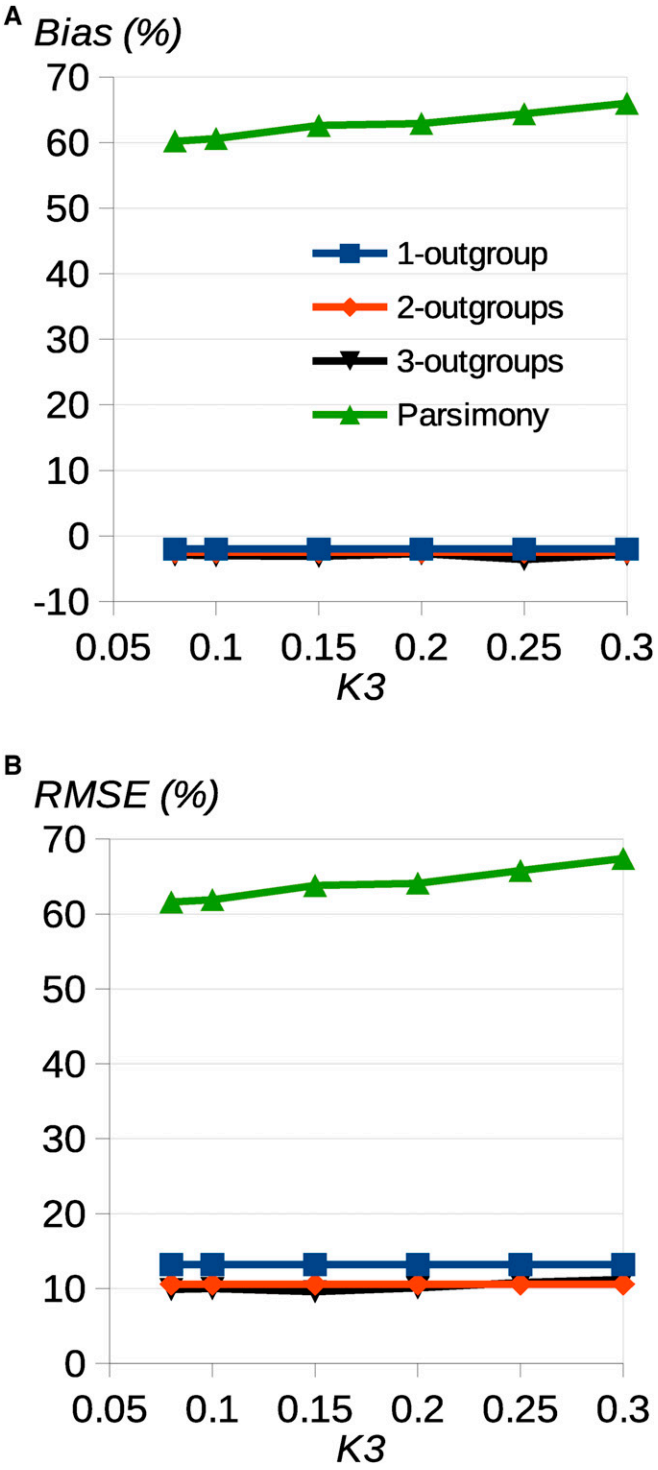

Figure 3 Effect of adding additional outgroups. Simulation results showing (A) the percentage bias = average deviation from the true uSFS, and (B) RMSE for uSFS element 19, as a function of divergence, K_3, from a third outgroup. There were 100,000 sites simulated in 360 replicates under the JC model, and $K_1 = 0.1$ and $K_2 = 0.1$. There were 20 gene copies sampled at each site in the focal species. Blue, red, yellow, green = results from uSFS inference with one, two, and three outgroups and parsimony, respectively. Note that estimates for one and two outgroups are invariant as a function of K_3.

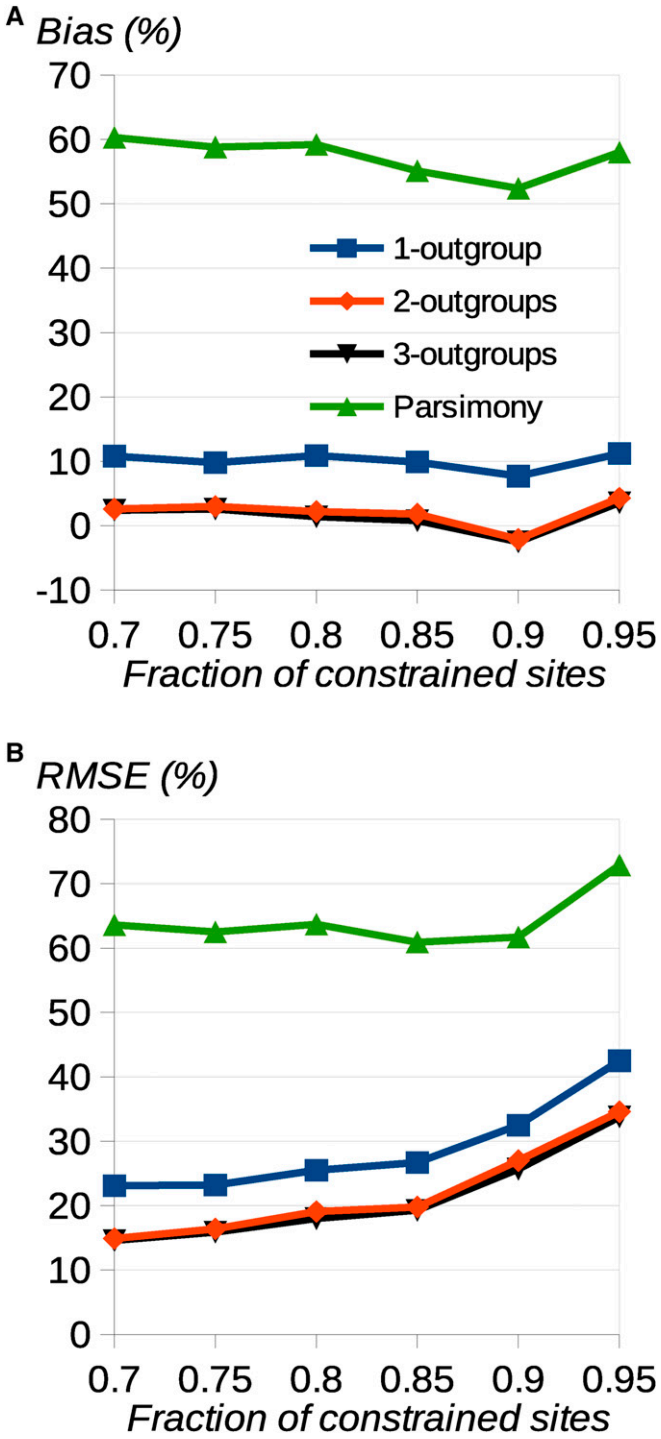

Figure 4 Effect of presence of selectively constrained sites on uSFS inference. Simulation results showing (A) the percentage bias and (B) RMSE for uSFS element 19 as a function of the fraction of constrained sites. There were 10,000 sites simulated in 3600 replicates under the JC model with three outgroups, and $K_1 = 0.1$, $K_2 = 0.15$, and $K_3 = 0.15$. There were 20 gene copies sampled at each site in the focal species. Blue, red, yellow, green = results from uSFS inference with one, two, and three outgroups and parsimony, respectively.

Consistent with the simulation results, the inclusion of a second outgroup (*D. yakuba*) perceptibly reduced the high derived allele frequency uSFS elements, compared to using a single outgroup (*D. simulans*) (Figure 5B). There is an uptick at the right-hand side of the inferred uSFS, but it is unknown whether this is a consequence of misinference, ongoing positive selection on fourfold sites, or positive selection on linked sites. Consistent with the simulations, parsimony infers a substantially higher frequency of high frequency derived allele classes.

Analysis of 1000 Genomes data

SNP ancestral states inferred by the 1000 Genomes Project Consortium (2010, 2015) have been widely used (*e.g.*, Mondal *et al.* 2015; Yang and Slatkin 2016; Harris and Pritchard 2017). In their 2015 article, a heuristic approach was used to assign the ancestral state based on the inferred human–chimpanzee common ancestor and the human–chimpanzee–orangutan common ancestor. Allele frequency information was not incorporated. We reinferred the ancestral state at fourfold degenerate and zerofold degenerate sites in the LWK population, using the ML method presented here, and compared the resulting uSFSs (Figure 6 and Figure S2). Because uSFSs from the full data set of 99 individuals (198 chromosomes) were difficult to visualize, we downsampled the LWK population to 25 randomly chosen individuals. The results from the full data set are qualitatively similar to those from the downsampled data and are presented in Figure S3. uSFSs inferred using one or two outgroups show only minor differences (Figure S4).

For non-CpG sites, the uSFS produced using the 1000 Genomes Project's ancestral states and the uSFSs produced by our ML method broadly agreed (Figure 6A). In contrast, for CpG sites, the results under the 1000 Genomes method and the JC model depart from the K2 model and the R6 model at the right-hand side of the inferred uSFS (Figure 6B). Under 1000 Genomes and JC, there is a pronounced uptick at high frequency derived variants, which is not present in the two more complex substitution models. In the case of the last uSFS element, for example, the 1000 Genomes and JC differ from the more complex models by about a factor of two.

CpG sites have an ~10-fold higher mutation rate than non-CpG sites in humans, due to an elevation in the number of C → T and G → A transitions (Nachman and Crowell 2000). This was borne out in the inferred branch lengths and the ratio of transition rate to transversion rate (κ) at the two classes of site. Under the R6 model, which is the best-fitting model for both classes of site, the length of the branch between the human–orangutan common ancestor and humans was 0.0083 for non-CpG sites and 0.092 for CpG sites. Estimates of κ under the K2 model were 4.2 and 8.3 for non-CpG and CpG sites, respectively, which are broadly in agreement with previous studies (*e.g.*, Keightley *et al.* 2011).

Discussion

This article generalizes a method we previously developed for inferring the uSFS (Keightley *et al.* 2016) by allowing the

Table 1 Differences between log likelihood of simpler models and the R6 model

	Change in log likelihood	
Model	1000 Genomes	DPGP
JC	−35,000	−13,000
K2	−55	−1,400
R6	0	0

The log likelihoods are obtained from stage 1 of the analysis (estimation of rate parameters, see text). The data analyzed are fourfold degenerate sites from the1000 Genomes Project and Rwandan sequences of DPGP phase 2.

inclusion of multiple outgroup species and potentially any phylogenetic tree topology (although only topologies of the type illustrated in Figure 1 have been implemented in the software). The new method gives nearly identical results to the previous method if the same outgroups are analyzed and the same substitution model is assumed. The new method implements three substitution models: the JC, K2, and R6 models. These models are nested. The K2 model gives the same likelihood as the JC model if the transition:transversion ratio parameter κ is fixed at 1. If the R6 parameters are constrained such that $r_3 = r_4$ (transition mutations) and $r_1 = r_2 = r_5 = r_6$ (transversion mutations) (see Figure 2), the same ML is obtained as the K2 model. Consistent with our previous results (Keightley *et al.* 2016), simulations suggest that the inclusion of a second outgroup generally increases the accuracy of uSFS inference, especially in the presence of variation in the rate of substitution among sites. The inclusion of a third outgroup did not, however, lead to a further improvement in uSFS inference accuracy. In the real data sets we have analyzed from *Drosophila* and humans, more complex substitution models gave higher log likelihoods in stage 1 of the analysis (evolutionary rate parameter estimation; Table 1), but this did not translate into a benefit in stage 2 (uSFS element inference) beyond the K2 model. The nucleotide substitutions models implemented are somewhat simplified in the sense that rates of change between pairs of nucleotides are symmetrical and these parameters do not vary between branches. It is possible that more complex models allowing these complications would lead to a further improvement, given that such effects are common in real data. A further weakness we hope to address in the future is its noncontext dependence of a substitution model (so we cannot deal with hypermutable CpGs), and further development along the lines of, for example, Arndt *et al.* (2003) will be needed.

We investigated whether our new method and parsimony produce different results when applied to real data. In the case of DPGP phase 2, parsimony estimates a much higher proportion of high frequency derived alleles (Figure 5). This has consequences for population genetic analysis. For example, if a three-epoch demographic model is fitted to the fourfold SFSs estimated by parsimony and by our present method (Schneider *et al.* 2011), the inferred population size changes and timings differ substantially (Table S1). In the case of the 1000 Genomes Project, we divided the data into CpG and non-CpG sites and inferred uSFSs separately for each class. At non-CpG sites there was a close agreement between the

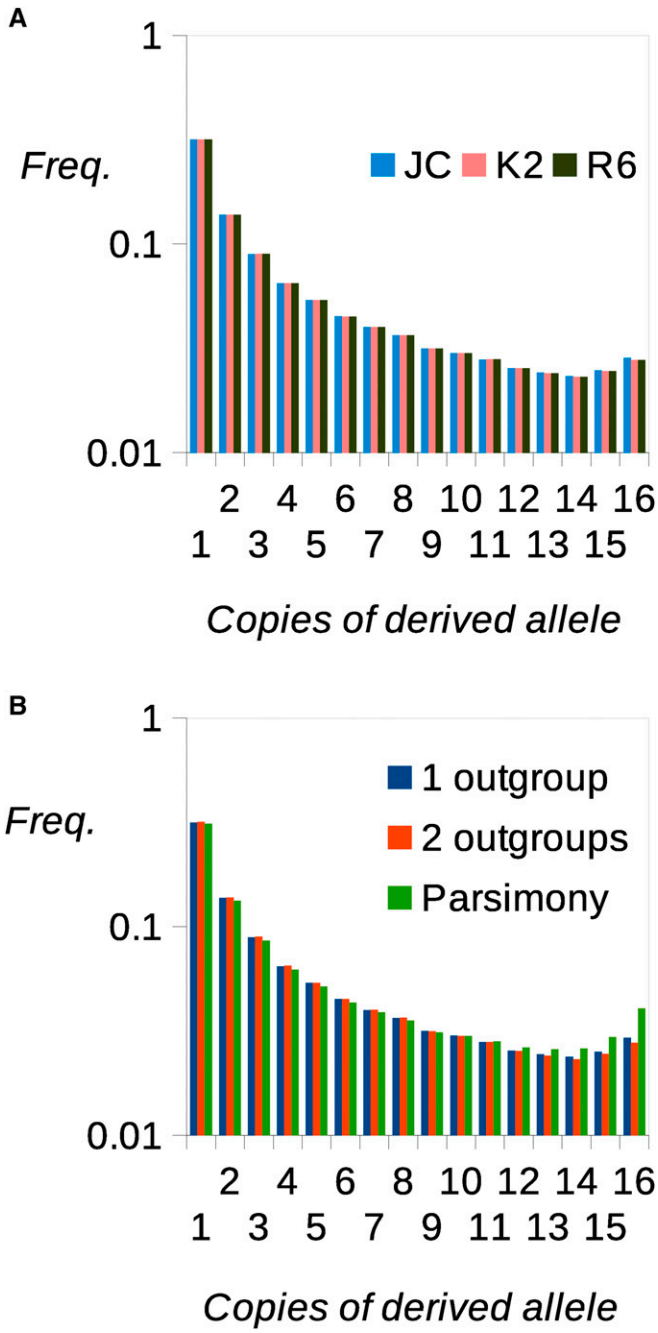

Figure 5 Analysis of fourfold degenerate sites of DPGP phase 2. (A) uSFSs estimated assuming three different substitution models. (B) uSFSs estimated using the method described in this article based on one outgroup (*D. simulans*) or two outgroups (*D. simulans* and *D. yakuba*) along with the uSFS inferred using parsimony. Freq., frequency.

uSFS generated using the 1000 Genomes Project's ancestral alleles to polarize variants and the uSFSs generated using the ML method. Parsimony is a more justifiable method of reconstructing ancestral states when the amount of change is small over the evolutionary time being considered, because it assumes *a priori* that change is unlikely (Felsenstein 1981). In contrast, parsimony is likely to be less accurate at CpG sites, which have an ~10-fold higher rate of evolution. Our results

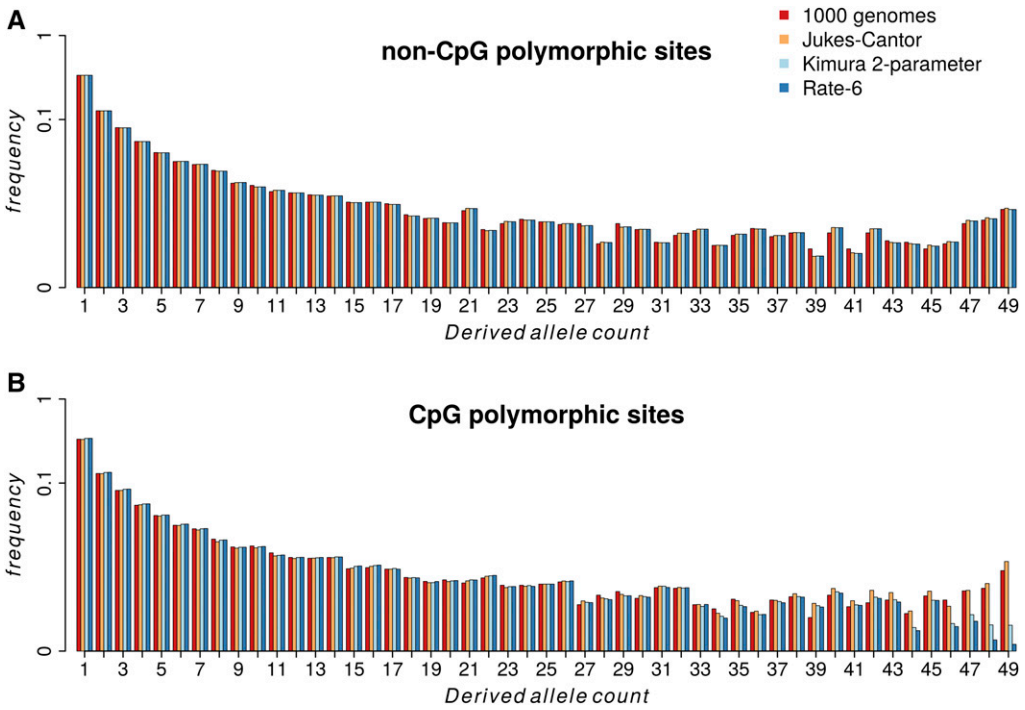

Figure 6 uSFSs for fourfold degenerate sites inferred by the 1000 Genomes Project and by the methods described in this article for three nucleotide substitution models. (A) non-CpG sites. (B) CpG sites.

bear this out. The uSFSs for CpG sites differed in the frequency of high frequency derived variants between the 1000 Genomes and the K2 and R6 models by up to a factor of eight. These are the class of variants where the greatest probability of misinference is expected. The JC model more closely mirrored the 1000 Genomes Project uSFS, presumably because it was unable to capture the ratio between transition rate and transversion rate at CpG sites, which is around twofold more extreme compared to non-CpG sites.

We have also addressed the problem of calculating ancestral state probabilities for polymorphic sites on a site-by-site basis. In doing so, we take into account both the nucleotide substitution parameter estimates (which determine the frequencies of multiple hits) and the frequencies of derived *vs.* ancestral alleles at other sites in the data. There are two main situations where this can make a significant difference compared to using parsimony. The first concerns sites where the outgroups are different in state from the focal species. These sites are frequently removed from the analysis (*e.g.*, Keinan *et al.* 2007; Sabeti *et al.* 2007; 1000 Genomes Project Consortium 2010, 2015; Langley *et al.* 2012), leading to underrepresentation of polymorphic sites, especially sites that have a low frequency of the derived allele, which tend to be the most common. The second situation concerns the tendency of parsimony to overestimate the frequency of sites with a high frequency of the derived allele. Consider the two configurations of nucleotides at a site of focal species and two outgroups shown in Figure 7. Assume that this is one of a large number of sites generated by simulation. At the site in question, there are 19 As and 1 C in the 20 gene copies sampled. In Figure 7A, the two outgroups are state A. By parsimony, the ancestral allele of the variation in the focal species would therefore be assigned as A. If the branch length b_1 (Figure 1) is 0.05, and using only information from the inferred substitution rates (*i.e.*, using the relative values of p_1 and p_2 calculated using Equation 2), $p(\text{A} = \text{ancestral}) = 0.98$. Taking into account the fact that high frequency derived allele sites are rare in the data set as a whole, and applying Equation 18, base A is even more strongly supported as the ancestral allele, *i.e.*, $p(\text{A} = \text{ancestral}) > 0.99$. This illustrates that parsimony is a good approximation for sites likely to have a low number of derived gene copies. The outcome is different for Figure 7B, where the two outgroups have the same state as the minor allele of the focal species. By parsimony, the ancestral allele would be assigned C, implying that we are certain the site has 19 copies of the derived allele. Using only information from substitution rate parameters and applying Equation 2, $p(\text{A} = \text{ancestral}) = 0.016$. Taking into account other sites in the data, which tell us that sites having 19 derived allele copies are uncommon, and applying Equation 18, $p(\text{A} = \text{ancestral}) = 0.14$. Thus, we

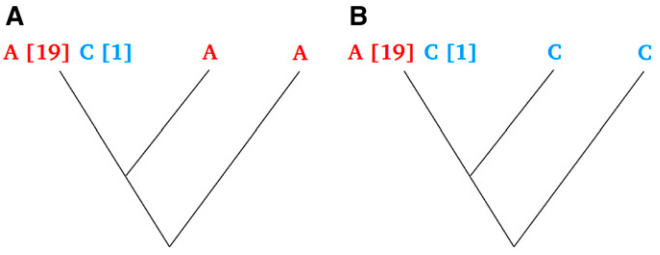

Figure 7 Example of a polymorphic site where 20 gene copies are sampled in a focal species and two outgroups have different nucleotide states. (A) Major allele agrees with outgroups. (B) Minor allele agrees with outgroups.

are much less certain that the derived allele is A at this site. This probability increases (decreases) if the outgroups are more distant (closer) to the focal species.

Acknowledgments

We thank Tom Booker and Dan Halligan for helpful discussions. This project has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 694212).

Literature Cited

1000 Genomes Project Consortium, 2010 A map of human genome variation from population-scale sequencing. Nature 467: 1061–1073 [corrigenda: Nature 473: 544 (2011)]. DOI: 10.1038/nature09534https://doi.org/10.1038/nature09534

1000 Genomes Project Consortium, 2015 A global reference for human genetic variation. Nature 526: 68–74. https://doi.org/10.1038/nature15393

Arndt, P. F., C. B. Burge, and T. Hwa, 2003 DNA sequence evolution with neighbor-dependent mutation. J. Comput. Biol. 10: 313–322. https://doi.org/10.1089/10665270360688039

Baudry, E., and F. Depaulis, 2003 Effect of misoriented sites on neutrality tests with outgroup. Genetics 165: 1619–1622.

Boyko, A. R., S. H. Williamson, A. R. Indap, J. D. Degenhardt, R. D. Hernandez *et al.*, 2008 Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet. 4: e1000083. https://doi.org/10.1371/journal.pgen.1000083

Collins, T. M., P. H. Wimberger, and G. J. P. Naylor, 1994 Compositional bias, character state bias and character state reconstruction using parsimony. Syst. Biol. 43: 482–496. https://doi.org/10.1093/sysbio/43.4.482

Dreszer, T. R., G. D. Wall, D. Haussler, and K. S. Pollard, 2007 Biased clustered substitutions in the human genome: the footprints of male-driven biased gene conversion. Genome Res. 17: 1420–1430. https://doi.org/10.1101/gr.6395807

Eyre-Walker, A., 1998 Problems with parsimony in sequences of biased base composition. J. Mol. Evol. 47: 686–690. https://doi.org/10.1007/PL00006427

Fay, J. C., and C.-I. Wu, 2000 Hitchhiking under positive Darwinian selection. Genetics 155: 1405–1413.

Felsenstein, J., 1981 Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17: 368–376. https://doi.org/10.1007/BF01734359

Halligan, D. L., A. Kousathanas, R. W. Ness, B. Harr, L. Eöry *et al.*, 2013 Contributions of protein-coding and regulatory change to adaptive molecular evolution in murid rodents. PLoS Genet. 9: e1003995. https://doi.org/10.1371/journal.pgen.1003995

Harris, K., and J. K. Pritchard, 2017 Rapid evolution of the human mutation spectrum. eLife 6: e24284. https://doi.org/10.7554/eLife.24284

Hernandez, R. D., S. H. Williamson, and C. D. Bustamante, 2007 Context dependence, ancestral misidentification, and spurious signatures of natural selection. Mol. Biol. Evol. 24: 1792–1800. https://doi.org/10.1093/molbev/msm108

Jackson, B. C., J. L. Campos, P. R. Haddrill, B. Charlesworth, and K. Zeng, 2017 Variation in the intensity of selection on codon bias over time causes contrasting patterns of base composition evolution in *Drosophila*. Genome Biol. Evol. 9: 102–123. https://doi.org/10.1093/gbe/evw291

Keightley, P. D., L. Eöry, D. L. Halligan, and M. Kirkpatrick, 2011 Inference of mutation parameters and selective constraint in mammalian coding sequences by approximate Bayesian computation. Genetics 187: 1153–1161. https://doi.org/10.1534/genetics.110.124073

Keightley, P. D., J. L. Campos, T. R. Booker, and B. Charlesworth, 2016 Inferring the frequency spectrum of derived variants to quantify adaptive molecular evolution in protein-coding genes of *Drosophila melanogaster*. Genetics 203: 975–984. https://doi.org/10.1534/genetics.116.188102

Keinan, A., J. C. Mullikin, N. Patterson, and D. Reich, 2007 Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. Nat. Genet. 39: 1251–1255. https://doi.org/10.1038/ng2116

Langley, C. H., K. Stevens, C. Cardeno, Y. C. G. Lee, D. R. Schrider *et al.*, 2012 Genomic variation in natural populations of *Drosophila melanogaster*. Genetics 192: 533–598. https://doi.org/10.1534/genetics.112.142018

Lohse, R. J. H., and N. H. Barton, 2011 A general method for calculating likelihoods under the coalescent process. Genetics 189: 977–987. https://doi.org/10.1534/genetics.111.129569

Matsumoto, T., H. Akashi, and Z. Yang, 2015 Evaluation of ancestral sequence reconstruction methods to infer nonstationary patterns of nucleotide substitution. Genetics 200: 873–890. https://doi.org/10.1534/genetics.115.177386

Mondal, M., F. Casals, T. Xu, G. M. Dall'Olio, M. Pybus *et al.*, 2016 Genomic analysis of Andamanese provides insights into ancient human migration into Asia and adaptation. Nat. Genet. 48: 1066–1070. https://doi.org/10.1038/ng.3621

Nachman, M. W., and S. L. Crowell, 2000 Estimate of the mutation rate per nucleotide in humans. Genetics 156: 297–304.

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: 913–918. https://doi.org/10.1038/nature06250

Schmidt, J. M., P. Battlay, R. S. Gledhill-Smith, R. T. Good, C. Lumb *et al.*, 2017 Insights into DDT resistance from the *Drosophila melanogaster* genetic reference panel. Genetics 207: 1181–1193. https://doi.org/10.1534/genetics.117.300310

Schneider, A., B. Charlesworth, A. Eyre-Walker, and P. D. Keightley, 2011 A method for inferring the rate of occurrence and fitness effects of advantageous mutations. Genetics 189: 1427–1437. https://doi.org/10.1534/genetics.111.131730

Tataru, P., M. Mollion, S. Glémin, and T. Bataillon, 2017 Inference of distribution of fitness effects and proportion of adaptive substitutions from polymorphism data. Genetics 207: 1103–1119. https://doi.org/10.1534/genetics.117.300323

Voight, B. F., S. Kudaravalli, X. Wen, and J. K. Pritchard, 2006 A map of recent positive selection in the human genome. PLoS Biol. 4: e72 [erratum: PLoS Biol. 4: e154; corrigenda: PLoS Biol. 5: e147 (2007)]. https://doi.org/10.1371/journal.pbio.0040072

Yang, M. A., and M. Slatkin, 2016 Using ancient samples in projection analysis. G3 (Bethesda) 6: 99–105. https://doi.org/10.1534/g3.115.023788

Yang, Z., 2007 PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24: 1586–1591. https://doi.org/10.1093/molbev/msm088

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: 1431–1439. https://doi.org/10.1534/genetics.106.061432

Communicating editor: S. Wright