Introduction

Identifying the action of natural selection from patterns of standing genetic variation has long been of interest to the population genetic community. The recent confluence of genomic data and the requisite computational power for large scale analyses now make this important goal a tractable problem. Of the many statistical methods developed to detect the presence and infer the strength of natural selection, those that make use of the frequency distribution of all derived mutations observed in a sample of chromosomes (the unfolded site-frequency spectrum, or SFS) appear to be the most powerful (Sawyer and Hartl 1992; Hartl et al. 1994; Akashi 1999; Bustamante et al. 2001; Nielsen et al. 2005b).

One technique that has been applied to make analysis of the SFS more robust to demographic and non-stationary evolutionary processes is to compare the SFS in a region of interest to the SFS in a genomic region putatively untouched by natural selection (Akashi 1999; Williamson et al. 2005). However, as large scale human polymorphism data has become available, a striking pattern has emerged. Figure 1 shows the normalized SFS expected under the standard neutral model (SNM) (Watterson 1975; Hudson 1990; Fu 1995) and the normalized SFS for the observed non-coding and synonymous mutations from 161 gene regions sequenced across a world-wide panel of 95 humans (NIEHS panel 2, as described in Hernandez et al. 2007). A visual inspection of figure 1 suggests that the observed data show an overrepresentation of both low and high frequency mutant alleles and a paucity of intermediate ones compared to the SNM.

An observed skew toward rare alleles is generally attributed to either the effect of natural selection restricting the spread of slightly deleterious mutations (Fu and Li 1993; Williamson and Orive 2002), or to the effect of a growing population where most mutations tend to be young (Slatkin and Hudson 1991; Griffiths and Tavaré 1994). A relative excess of mutant alleles at high frequency compared to the SNM is typically attributed to the presence of recurrent positive natural selection (Akashi and Schaeffer 1997; Hartl et al. 1994; Bustamante et al. 2001), linkage to the target of a recent selective sweep (Kim and Stephan 2000; Fay and Wu 2000), or the existence of hidden population substructure with low levels of migration (Wakeley and Aliacar 2001). Because the data used here includes several individuals from each population, hidden population structure is unlikely to have caused the excess high frequency tail, therefore an excess of both low and high frequency mutant alleles casts doubt upon the assumption of selective neutrality for both non-coding and synonymous mutations.

Another explanation for the relative excess of high frequency mutations in putatively neutral regions is based on violations of the parsimony assumption used to construct the SFS. Namely, if a substitution occurred during the divergence time of the 2 species being compared, and a subsequent polymorphism (back mutation) recently arose at the same site, then the nucleotide observed in the outgroup would not be the ancestral state of the SNP. Since most mutations are transitions, multiple mutations at a site (multiple hits) would often be back mutations, thus ancestral misidentification would often times go unnoticed. Because most derived mutations are expected to be rare (Watterson 1975; Fu 1995), ancestral misidentification would most often result in labeling low frequency mutations as high frequency ones (Baudry and Depaulis 2003).

Previous studies that have addressed multiple hits have either focused on correctly counting the number of substitutions between 2 diverged sequences (Jukes and Cantor 1969; Kimura 1980), or correctly counting the number of segregating mutations in a sample of sequences from hypervariable regions (Tajima 1996; Yang 1996). These studies typically assume that all sites are evolving homogeneously and at stationarity, but may allow for mutation rate
variation across sites (though constant over time). However, recent analysis of mammalian sequences have revealed mutational patterns that are much more complex than these models account for. In particular, the mutation rate at a site appears to depend on the source and target nucleotides, as well as the adjacent nucleotides 3′ and 5′ of the site (Hwang and Green 2004; Siepel and Haussler 2004). Under such a context-dependent model, site-specific mutation rates can vary in magnitude as much as 50-fold (Blake et al. 1992; Hess et al. 1994; Hwang and Green 2004; Siepel and Haussler 2004). The largest change in site-specific mutation rates is due to the production (or removal) of a methylated CpG dinucleotide (the so called “CpG-effect”). Surprisingly, non-CpG nucleotides can also vary as much 4-fold depending on their context (Hwang and Green, 2004). By taking the complexities of such a mutation process into account, we derive a mathematical model for the SFS that can correct for ancestral misidentification.

Materials and Methods

Theory

In this study, we consider a context-dependent nucleotide mutation model. This model assumes that the rate of mutation at a nucleotide site depends on its flanking nucleotide context (i.e., its 5′ and 3′ neighboring nucleotides), and is composed of 16 4 × 4 nucleotide mutation rate matrices (one for each pair of flanking nucleotides). For ease of notation, we consider the 64 × 64 instantaneous trinucleotide mutation rate matrix that is restricted to only allow changes at the second position of each trinucleotide. Hwang and Green (2004) estimated the parameters of such a context-dependent mutation model assuming non-reversibility and strand-symmetry from an untranscribed DNA sequence of length 5.2 Mb across 19 mammalian species. In our applications, we will use their estimates obtained along the human lineage, but our derivations will be sufficiently general to allow for any model of mutation rate variation (subject to additional assumptions as necessary).

The probability of substituting 1 nucleotide for another over a time interval $t_i$ in a fixed context (i.e., when both 5′ and 3′ nucleotides do not change), can be obtained from the probability substitution matrix $P(t_i)$, where $P(t_i) = \exp(Q t_i) = \sum_{r=0}^{\infty} (Q t_i)^r / r!$. Note that time ($t_i$) is scaled in terms of the expected number of substitutions per site, and that the main diagonal of $Q$ was set so as to satisfy the mathematical requirement that each row sums to zero.

Probability of Ancestry

We will derive the probability of correctly identifying the ancestral state of a SNP when the orthologous site in the outgroup matches one of the segregating alleles. Let $M$ be the unknown ancestral state of the SNP, with $S$ the 2 segregating alleles (an unordered pair) and $U$ the allele observed in the outgroup. Assume that the divergence time between species is scaled in terms of the number of substitutions per site and known. For simplicity, and to accommodate the structure of our context-dependent mutation model, we assume that the nucleotides flanking the polymorphism have remained constant since the divergence of the 2 species being compared (henceforth referred to as the constant context assumption).

For the arbitrary case of observing an allele $U = u$ in the outgroup while the pair of alleles $S = \{u, x\}$ are segregating in the population, the probability of correctly identifying the ancestral state of the SNP, $p_{\text{nucl}}$, is

$$p_{\text{nucl}} = P(M = u | U = u, S = \{u, x\}, t_i) = P(M = u | U = u, S = \{u, x\}, t_i) / P(M = u | U = u, S = \{u, x\}, t_i).$$

This follows from conditional probability under the assumption that either $M = u$ or $M = x$ (equivalent to an infinite-sites assumption for within species polymorphism). Each of the terms in equation (1) can be simplified by reordering terms and applying conditional probability to sequences at stationarity for trinucleotide frequencies. For example, in the numerator we have

$$P(M = u, U = u, S = \{u, x\}, t_i) = P(U = u | t_i) P(M = u | U = u, t_i) P(S = \{u, x\} | M = u, U = u, t_i) = \sum_{u = x}^{u = x} P_{uuu}(t_i) \sum_{x = x}^{x = x} P_{xux}(t_i)$$

$$= \sum_{u = x}^{u = x} \sum_{x = x}^{x = x} P_{uuu}(t_i) P_{xux}(t_i)$$

where the second equality follows from the assumption of stationary trinucleotide frequencies since the time of divergence (first term), and that the type of SNP is independent of any outgroup information when conditioning on its ancestral state (third term). Note that by conditioning on the ancestral state of a polymorphic site to be $u$, we can treat the possible derived states as competing exponential processes. This model has been well studied, and it is known that the probability that the mutation will be of type $u \rightarrow x$ is just the ratio of its rate to the sum of all competing rates. In the final equality, time-reversibility refers to a model in which $p_{sv} P_{sv}(t_i) = p_{uv} P_{uv}(t_i)$ for all $u$ and $x$, and that in the case of non-reversibility we sum over $x$, the 4 possible allelic states that could have occupied the most.
recent common ancestor (MRCA) of both species while maintaining the constant context assumption. Also note that in the non time-reversible case, we have assumed that the branches leading to both species are equal, but this need not be true and is easily relaxed by conditioning on the 2 branch lengths. The probability of correctly identifying the ancestral state of a SNP using outgroup information is then obtained by substituting equation (2) and a similarly derived term into equation (1).

Correcting the SFS

We consider the SFS constructed from a sample of \( n \) chromosomes with orthologous outgroup information. For all pairs of trinucleotides (generically denoted \( u \) and \( x \)) that meet the constant context assumption, let \( N_{ux}(i) \) be the number of diallelic sites at which \( i \) chromosomes carry allele \( x \), and the remaining \( n - i \) chromosomes as well as the outgroup carry the allele \( u \). \( N_{ux}(i) \) would then be the SFS for the inferred \( u \rightarrow x \) mutations. However, because of ancestral misidentification, \( N_{ux}(i) \) is not necessarily the SFS for all SNPs of type \( u \rightarrow x \). We therefore model \( N_{ux}(i) \) as a mixture of the number of true \( u \rightarrow x \) mutations at frequency \( i \) whose ancestral states were correctly identified (an unknown number denoted by \( R_{ux}(n-i) \)) and the number of true \( x \rightarrow u \) mutations at frequency \( n - i \) whose ancestral states were misidentified (another unknown number, denoted \( R_{ux}(n-i) \)).

The proportion of mutations from \( u \rightarrow x \) whose ancestral states were correctly identified, \( f_{ux} \), can then be written as the relative probability of observing \( u \) in the outgroup and a \( u \rightarrow x \) mutation segregating in the population (similar to equation (1)). Namely,

\[
 f_{ux} = \frac{P(M = u, U = u, S = \{u, x\}|i)}{P(M = u, U = u, S = \{u, x\}|i) + P(M = u, U = x, S = \{u, x\}|i)}. \tag{3}
\]

The number of \( u \rightarrow x \) mutations at frequency \( i \) whose ancestral states were correctly identified is then \( f_{ux}R_{ux}(i) \), and the number of \( x \rightarrow u \) mutations at frequency \( n - i \) whose ancestral states were misidentified is then \( (1 - f_{ux})R_{ux}(n-i) \). We can then write the observed SFS as a function of the true SFS as follows:

\[
 N_{ux}(i) = f_{ux}R_{ux}(i) + (1 - f_{ux})R_{ux}(n-i)
\]

\[
 N_{ux}(n-i) = f_{ux}R_{ux}(n-i) + (1 - f_{ux})R_{ux}(i). \tag{4}
\]

This is a system of 2 equations with 2 unknowns \((R_{ux}(i) \text{ and } R_{ux}(n-i))\) that can readily be solved to give the reconstituted SFS as a function of the observed quantities

\[
 R_{ux}(i) = \frac{f_{ux}N_{ux}(i) - (1-f_{ux})N_{ux}(n-i)}{f_{ux} + (1-f_{ux})} \tag{5}
\]

\[
 R_{ux}(n-i) = \frac{f_{ux}N_{ux}(n-i) - (1-f_{ux})N_{ux}(i)}{f_{ux} + (1-f_{ux})}. \tag{6}
\]

To obtain the \( i \)th entry of the corrected SFS, \( F_c(i) \), we then sum equations (5) and (6) over all pairs of trinucleotides \( u \) and \( x \) that meet the constant context assumption

\[
 F_c(i) = \sum_{\langle u,x \rangle} R_{ux}(i). \tag{7}
\]

There are 3 steps to implementing this correction:

1. Tally the number of SNPs in each trinucleotide mutation class at each frequency (e.g., \( N_{ACG \rightarrow ATG}(i) \) = the number of \( ATG \rightarrow ACG \) SNPs with \( i \) chromosomes carrying ATG and \( N_{ATG \rightarrow ACG}(n - i) \) = the number of \( ATG \rightarrow ACG \) SNPs with \( n - i \) chromosomes carrying ACG)

2. Calculate the probabilities of ancestral misidentification from equation (3) (e.g., \( f_{ACG \rightarrow ATG} \) and \( f_{ATG \rightarrow ACG} \))

3. Plug the results from steps 1 and 2 into equations (5) and (6), and sum across all pairs of trinucleotides.

Simulations

To test the effect of a context-dependent mutation model with and without natural selection on our ability to compare closely related sequences, we simulated population genetic data with an outgroup. Our simulation program was written in the C programming language, and can be described as generating 2 Wright-Fisher populations with a known divergence time under a context-dependent mutation process across finitely many sites. The structure of the simulation is as follows: a single diploid population of constant size \( N_e \) is evolved for a large number of generations (\( \geq 8N_e \)) to mix and introduce variation, at which point there is a speciation event that splits the population into 2 independent populations of constant size \( N_e \). After 2\( N_e \tau \) generations (the divergence time), a sample is taken from both populations. From these samples, several statistics are computed (e.g., Tajima’s \( D \), Fu and Li’s \( D \), Fay and Wu’s \( H \), etc. as described below).

Each simulated generation consisted of 3 main components: (i) random mating, whereby the chromosomes of each diploid individual were chosen from two individuals of the previous generation (with replacement) with probabilities given by their relative fitness (assuming fitnesses are multiplicative across sites), (ii) mutation, whereby a Poisson number of events occurred with mean \( \rho/2 = 2N_e \mu \) (the population scaled mutation rate per sequence per generation), and (iii) recombination, whereby a Poisson number of events occurred with mean \( \rho/2 = 2N_e \lambda \) (the population scaled recombination rate per sequence per generation).

Recombination events were distributed uniformly across individuals and sites (i.e., no hot spots). Mutation events were distributed uniformly among all chromosomes in the population, but sites within chromosomes were chosen according to each site’s “hit probability” (i.e., the site’s mutation rate relative to the rest of the sequence). Because the mutation rate at a site is context-dependent, a mutation event at a flanking nucleotide can change a site’s hit probability as well as the overall mutability of a sequence (though only nominally for sufficiently long sequences). After choosing a site to mutate, a new nucleotide was chosen according to the relative mutation rate from the current state into each of the other 3 states (conditional on the flanking nucleotides). If the chosen mutation event resulted in an
amino acid change and a selective effect was desired, then the derived state was also assigned a fitness coefficient from a specified distribution (e.g., discrete, Gamma, Normal, or a more complicated mixture model). Should the mutation event result in the genesis of a stop codon, a different nucleotide was chosen for this site using the same probabilities as before. Although our algorithm for avoiding stop codons reduces the state space for some mutations, and slightly alters the resulting stationary trinucleotide frequencies, this does not induce a systematic bias in terms of polymorphism or divergence since a mutation is still introduced into the population.

In order to accommodate our assumption of statistical stationarity, we seeded our founding population with a single sequence drawn from the stationary distribution induced by the context-dependent mutation model. To obtain such a sequence, we first determined the stationary distribution of nucleotides for each context, \( \pi \), which is the diagonal of \( \lim_{t \to \infty} \pi \exp(Q_t) \). By construction, each nucleotide context has 4 entries in \( \pi \) whose values correspond to the probability of each nucleotide in the given context. We then initialized each site of a sequence to correspond to the probability of each nucleotide in the given context. In order to avoid problems at the boundary of a sequence, we considered the sequence to be circular during the updating phase (i.e., the first and last site were considered to be neighbors). We continued to update the sequence until trinucleotide frequencies converged to stationarity. Convergence to stationarity was monitored using the \( \sqrt{R} \) statistic (Gelman et al. 2004). Convergence to \( \sqrt{R} \in (0.99, 1.01) \) for each of the 64 trinucleotide frequencies was ensured for 100 independent chains after 500 iterations (i.e., 500L nucleotide updates).

Unless otherwise noted, all simulation results discussed were obtained assuming 2 diploid populations with effective size of \( N_e = 250 \), where \( n = 50 \) chromosomes were sampled from 1 population and \( n = 2 \) chromosomes were sampled from the outgroup (both without replacement). In agreement with diffusion theory (Sawyer and Hartl 1992; Eywas 2004), neutral coalescent theory (Kingman 1982; Hudson 1990), and previous forward simulation studies (McVean and Charlesworth 2000; Williamson and Orive 2002), we found that the actual population size does not impact our results when the mutation rate, recombination rate, divergence time, and selection coefficients are scaled by the effective population size (tested using \( N_e = 250, 500, 1000, \text{and} 5000 \), results not shown).

Implementation of our forward simulation was done by storing a single consensus sequence for each population, and 2 Splay trees (efficient self-balancing binary search trees; Sleator and Tarjan 1985) for each individual. Each tree kept track of the mutations that an individual carried on a chromosome (including information regarding the generation in which they arose, whether or not they were synonymous, ancestral and derived codons, the states of the flanking nucleotides, the selective effect, as well as the true ancestral state). An individual inherited an entire tree from each parent, and the consensus sequence was updated after each fixation event, at which point the reference to that mutation was removed from the trees of each individual and stored in a separate Splay tree. Such a simulation allowed us to evaluate the effect of ancestral misidentification directly, since the true ancestral state of each SNP was known, and we had outgroup information.

In order to evaluate the effect of natural selection, we implemented 3 “shift models” of selection. In a shift model, the selection coefficient at a site returns to zero when a mutation fixes in the population (i.e., fitness is relative). In our simulations, the population genetic selection coefficient \( \gamma = 2N_e s \) was drawn for each nonsynonymous mutation. The first selective scheme \( (\gamma_3) \) is a Normal-shift model where \( \gamma_3 \sim N(0, 2) \). Normal-shift models are generally considered to represent positive selection, as a new mutation will increase the individual’s fitness half of the time, and positively selected mutations are much more likely to be maintained/ixed in the population (Cutler 2000). The second selective scheme \( (\gamma_2) \) refers to a gamma-shift model, where all nonsynonymous mutations are deleterious with \( \gamma_2 \sim -\Gamma(1, 0.25) \) (this is a model of negative selection, where \( \Gamma \) is the common gamma distribution with mean 4, but reflected across the y-axis). The third selection scheme \( (\gamma_1) \) refers to a mixture of normals model with a point mass at zero, where most mutations are neutral (54.5% of nonsynonymous mutations have \( \gamma_3 = 0 \)) or strongly deleterious (45% of nonsynonymous mutations have \( \gamma_3 \sim N(-10, 5) \)), but a small fraction are strongly advantageous (0.5% of nonsynonymous mutations have \( \gamma_3 \sim N(50, 5) \)). Note that in each of these models, synonymous sites are completely neutral (i.e., \( \gamma_{syn} = 0 \)).

In order to contrast our simulations under the context-dependent mutation model with a nucleotide mutation model having no context effects, we simulated evolution under a generalized Kimura 2-parameter model (Kimura 1980). In this model, each nucleotide has its own transition and transversion mutation rates, but there were no context effects such as the hypermutability of CpG dinucleotides. The parameters of this model were estimated by Zhang and Gerstein (2003) from 1,726 human ribosomal protein pseudogenes.

We verified the neutral component of our simulation study by comparing the number of segregating sites, fixed differences, and the SFS to those generated under the coalescent simulation program ms (Hudson 2002). In supplementary table S1, we show that the distribution of the observed number of segregating sites and fixed differences (with short divergence times) in our simulation matched the expected distribution very closely. However, as the divergence time increases, the number of unobserved substitutions grows substantially. Because we store information regarding all mutations during the simulation, we can see that the total number of substitutions that occurred matches the number we expect to see very closely, but due to multiple substitutions per site, the observed number of fixed differences is much less than expected.

**Simulations for Poisson Random Field Analysis**

A statistical test for positive selection using the SFS under the Poisson random field (PRF) framework of Sawyer...
In the case of positive selection all synonymous mutations are neutral (scaled mutation rate shows the average over 5,000 simulations, with an average population SNP is inferred from the outgroup, and dotted lines show the result of SNP is known, dashed lines show the SFS when the ancestry for each human-chimpanzee. Solid lines show the SFS when the ancestry for each (gray and black, respectively) when using an outgroup as diverged as identification on the SFS with and without rampant positive selection neutral (with population scaled selection coefficient and Hartl (1992) was recently proposed (Nielsen et al. 2005a). This test assumes that each nonlethal mutation en-

genic recombination, each of length 1 kb with datasets consisting of 200 independent loci with no intra-
lection to ancestral misidentification, we simulated 1,000 site

The Effect of Ancestral Misidentification

Results

The Effect of Ancestral Misidentification

Using forward simulations with a context-dependent mutation process, we compared the effect of ancestral misidentification on the SFS under neutrality to the effect of recurrent positive selection. Figure 2 shows that while the expected frequency distribution of neutral mutations (solid black line) differs from the frequency distribution of mutations with recurrent positive selection (solid gray line), ancestral misidentification can cause the frequency distribution of neutral mutations (dashed black line) to look substantially like positive selection. This qualitative similarity suggests that complex mutation patterns across sequences of finite length can compromise statistical tests of the standard neutral model, even when comparing species as closely related as human and chimpanzee.

Some of the most common methods for identifying departures from the standard neutral model (SNM) are based on comparing summary statistics of the SFS (Tajima 1989; Fu and Li 1993; Fay and Wu 2000). Among the most common summaries are those that use the SFS to estimate the population scaled mutation rate, \( \theta \). One of the first estimates of \( \theta \) was proposed by Watterson (1975), and uses the total number of segregating sites (\( \theta_W \)). Another common estimate is based on the average pairwise heterozygosity (\( \theta_\pi \); Tajima (1983)). Under the assumptions of the SNM, these 2 estimates of \( \theta \) are unbiased; however, departures from the SNM will affect these 2 estimates differently. Tajima (1989) capitalized on this characteristic and proposed a statistical test for the equality of \( \theta_W \) and \( \theta_\pi \), and defined the test statistic \( D \). However, since neither \( \theta_W \) nor \( \theta_\pi \) require ancestral information (i.e., both can be written as a function of the folded SFS), ancestral misidentification has no effect on Tajima’s \( D \). Indeed, the distribution of the \( D \) statistic for 5,000 forward simulations with ancestral misidentification significantly overlaps the distribution of \( D \) from 20,000 coalescent simulations for a range of parameters (MWU (Mann-Whitney U test) \( P = 0.35, 0.24, 0.39 \) for \( \theta = 2.15, 4.3, 10.75 \), respectively).

The second test statistic we considered was the statistic proposed by Fu and Li (1993) that uses outgroup information (denoted here as \( D_\beta \)). The motivation for this test was based on the notion that purifying selection or a recent selective sweep would lead to a relative excess of young/rare mutations in the population, while balancing selection would lead to a relative deficiency. To capture this effect, \( D_\beta \) statistically tests for the equality of \( \theta_\beta \) and \( \theta_\pi \) (the estimate of \( \theta \) based on the number of mutations that are carried by only a single chromosome in the sample). However, one of the effects of ancestral misidentification is that some rare mutations are mislabeled as being at high frequency. This causes Fu and Li’s test to reject the SNM more frequently than expected at the 5% level due to a deficiency of rare mutants (i.e., an excess of observations greater than the 95% critical value, shown in fig. 3a), especially when divergence times are long.

The third test statistic that we consider is Fay and Wu’s \( H \) (Fay and Wu 2000). This test was motivated by the desire to identify regions that may have been the target of a recent selective sweep. One characteristic of a selective sweep in a recombinating locus is that some neutral alleles will have hitchhiked to very high frequency in the population. To identify regions with an overrepresentation of high frequency mutants, \( H \) statistically tests for the equality of \( \theta_\beta \) and \( \theta_\pi \) (the estimate of \( \theta \) based on the homozygosity of derived mutations; Fu 1995). However, one of the primary effects of ancestral misidentification is to increase the inferred proportion of derived mutations at very high frequency. As with the test of Fu and Li, ancestral
Correcting the SFS

To account for ancestral misidentification in the presence of context-dependent mutation rate variation, we model each component of the SFS as a mixture of SNPs whose ancestral states were correctly identified and SNPs whose ancestral states were misidentified (2 unknown quantities; see Materials and Methods). In our model, the mixture components account for variation in substitution probabilities, stationary frequencies, and relative rates of mutation for each type of polymorphism in a given nucleotide context. The resulting model is a linear system of equations with unknown quantities that can readily be solved to reconstitute the true SFS.

Figure 2 shows examples of simulated SFS with and without positive natural selection (in gray and black, respectively). The solid lines show the SFS when the ancestral state is known (or the expected SFS in the case of neutrality; Watterson 1975; Fu 1995). The dashed lines show the SFS when the ancestral state of each SNP is identified using an outgroup with divergence similar to human and chimpanzee. The dotted lines show the SFS after correcting for ancestral misidentification. If the correction were perfect, then the dotted lines would be directly on top of the solid lines. This is nearly the case for our neutral simulations, but a slight deviation remains in the case of positive selection. However, this result is promising, given that our correction does not explicitly model the effect of natural selection.

For a range of mutation, selection, and divergence parameters, 5,000 polymorphism datasets were simulated with an outgroup. In each dataset, the ancestral state of each SNP was identified using the outgroup, and the average proportion of SNPs whose ancestral states were misidentified was calculated (fig. 4). We then applied our correction to the observed data using both the observed divergence and the true divergence (measured in terms of the average number of substitutions per site), and calculated the number of SNPs whose ancestral states were still misidentified (fig. 4). Though our correction can eliminate a majority of the ancestral misidentification events for all the parameters tested (with and without selection), underestimating the level of divergence can have a strong impact (particularly for highly diverged outgroups). In our simulations of selection (fig. 2 and 4c and d), our correction tends to underestimate the effect of ancestral misidentification. This is primarily because selection provides an unaccounted for source of variance in the substitution probabilities across sites.

We determined the extent to which ancestral misidentification can produce statistical evidence for positive selection using a recently proposed statistical test based on the Poisson random field (Nielsen et al. 2005a). This test assumes that each nonlethal mutation enters the population and is assigned to one of three categories: neutral, positively selected, or negatively selected with probabilities $p_0$, $p_+$, and $p_-$ (respectively). We found that the LRT based on the PRF framework is not robust to ancestral misidentification, but the false-positive rate can be controlled by correcting for ancestral misidentification. As shown in figure 3c,
the LRT for the true SFS (based on saving the ancestral state of each polymorphism during the simulation) was conservative at the 5\% level, rejecting $H_0$ only 2.4\% of the time. Applying the LRT to the uncorrected (observed) SFS resulted in rejecting $H_0$ 60.7\% of the time at the 5\% critical value. After performing the correction proposed in equation (7) using the observed divergence, the LRT rejected $H_0$ 4.6\% of the time at the 5\% critical value.

An alternative approach to dealing with ancestral misidentification might be to restrict the analysis to the subset of SNPs that have outgroup support from multiple species (e.g., the SNPs that have an allele matching both chimpanzee and rhesus macaque). To evaluate this strategy, we used a non-coding data set from the African-American population (described in our concurrent article Hernandez et al. 2007). Orthologous regions were identified in both chimpanzee and macaque using BLAT (Kent 2002). As shown in figure 5, using macaque as an outgroup to identify the ancestral states of human SNPs would lead to a very large excess of high frequency derived mutations (a result of \(~6\%\) divergence over 25 million years). Using chimpanzee alone results in a lower proportion of high frequency derived mutations, but still an excess. Requiring the chimpanzee allele to match the macaque allele at each human SNP reduces the proportion of high-frequency derived mutations further, but there is still a slight excess. The excess high frequency tail completely vanishes when we apply the correction proposed above. It is also important to note that while the ancestral state of 10,179 SNPs could be identified using the chimpanzee, requiring the chimpanzee to equal the macaque results in just 7,082 SNPs (a reduction of 30\%). However, by using the correction proposed in this paper when inferring ancestral states using chimpanzee, only 404 SNPs (3.96\%) must be removed from the analysis (due to violations of the constant-context assumption).

To see the effect that the various strategies for identifying ancestral states of SNPs would have on population genetic analyses, we asked whether the SFS produced from each strategy could be explained by a simple demographic history with no selection. We used the method developed by Williamson et al. (2005), which fits a 2-epoch demographic model to the SFS in the absence of selection. The model assumes that at some time $\tau$ in the past (scaled in terms of the current effective population size), the effective population size instantaneously changed from $N_A$ to $N_C$ (i.e., from the ancestral to the current population size, with $\omega = N_A/N_C$ the magnitude of the change). We then performed a goodness-of-fit test, to see whether the inferred demographic model could sufficiently explain the data. Using the chimpanzee alone, we inferred $\tau = 0.11$ and $\omega = 0.49$

![Fig. 4.—Simulation results showing the proportion of SNPs whose ancestral states have been misidentified for each value of the average population scaled mutation rate per site $\theta$ (a–b) or selection scheme (c–d). Each bar represents the average over 5,000 simulations. “Uncorrected” refers to inferring the ancestral state of each SNP using an outgroup sequence with population scaled divergence time $\tau = 10$ (a and c, representing human–chimpanzee divergence) or $\tau = 50$ (b and d, representing human–macaque divergence), and “corrected” refers to the data after performing our proposed correction with either the observed number of substitutions per site or the true number of substitutions per site (as determined by saving information during the simulation). Simulations with selection (c–d) were performed with $\theta = 0.0043$ per site and no intergenic recombination.](https://academic.oup.com/mbe/article-abstract/24/8/1792/2925629/1798-Hernandez-et-al)

![Fig. 5.—The SFS using 4 strategies to identify the ancestral states of each SNP. The solid gray (dashed black) line shows the result of using the rhesus macaque (chimpanzee) genome, the dashed gray line shows the result of using the chimpanzee alone, the dashed line shows the result of using the chimpanzee and macaque, and the solid line shows the result of applying our correction to the observed data using only the chimpanzee outgroup. The legend indicates the resulting number of SNPs (S) for each strategy.](https://academic.oup.com/mbe/article-abstract/24/8/1792/2925629/1798-Hernandez-et-al)
Ancestral misidentification of a SNP occurs when the ancestral state is not the allele observed at the orthologous site in the outgroup. Because most derived mutations tend to be rare (Watterson 1975; Fu 1995), ancestral misidentification most often leads to mislabeling low frequency derived mutations as very high frequency ones when constructing the site-frequency spectrum (SFS) using an outgroup. This results in an increase in the proportion of SNPs found at very high frequency, a pattern similar to positive selection. We have shown that ancestral misidentification causes many statistical tests to reject selective neutrality more frequently than expected.

Ignoring outgroup information and using the frequency distribution of minor alleles (the folded SFS) leads to a lack of power for some statistical tests (Bustamante et al. 2001). Moreover, the inclusion of another outgroup that is more diverged seems to cause a large fraction of the data (30%) to be removed while not significantly reducing the problem. We therefore propose a correction for the effect of ancestral misidentification on the SFS that takes advantage of our growing knowledge regarding the underlying neutral mutation process. The novelty of this method is in our use of a finite sites model for between species divergence, while maintaining an infinite sites model for within species polymorphism in the context of a realistic mutational process. Our technique is independent of population demography and ambivalent regarding the presence/absence of natural selection. Though it has long been suggested that frequency information is informative about the ancestral state of a polymorphism (Watterson and Guess 1977; Griffiths and Tavaré 1998), we have modeled ancestral misidentification strictly as a function of the mutational process for the sake of being robust to unknown demographic histories, and to make inference regarding natural selection (or the distribution thereof) more conservative.

Our analysis shows that ancestral misidentification occurs frequently between human and chimpanzee, and we argue that this is primarily due to the fact that neutral mutation rates can vary both across sites as well as over time. An effect of ancestral misidentification could also be caused by sequencing errors in the outgroup. However, if this were the sole cause of the observed excess high frequency derived SNPs, then we would not expect to see a more dramatic effect when ancestral states were determined using rhesus instead of chimpanzee (fig. 5). Moreover, it may be the case that no species pair exists such that one can reliably polarize polymorphism data. For species more closely related than human and chimpanzee, further complications may arise due to lineage sorting and shared polymorphisms. It is therefore necessary to consider probabilistic models that can account for ancestral misidentification. However, our approach makes several simplifying assumptions that may be violated in some situations. For example, we assume that the flanking nucleotides of each polymorphism are constant between species, which may not always be the case. Using the polymorphism dataset consisting of 11,626 genes from 39 humans and a chimpanzee presented in Bustamante et al. (2005), we estimated that roughly 2.6% of SNPs in the coding regions of the human genome violate this constant context assumption (1.2% due to adjacent fixed differences and 1.4% due to adjacent SNPs). In the current non-coding dataset, 5.1% of SNPs were in violation of the constant context assumption (1.9% due to adjacent fixed differences and 3.2% due to adjacent polymorphisms). However, the inclusion and exclusion of polymorphic sites in violation of the constant context assumption has no effect on the qualitative results presented here.

The model presented here also assumes that all SNPs are independent, an assumption that is clearly not true for many SNPs occurring within the same gene. However, the effect of this assumption is minimized when pooling SNPs from across the genome. Moreover, because mutation rates and patterns are not necessarily constant across different regions of the genome (e.g., it is unknown whether different isochores are following different mutational patterns), it is of interest to develop new techniques that will allow us to estimate context-dependent effects and account for ancestral misidentification on a local scale.

The effect we have observed will likely hold for any species pair in which there is sufficient mutation rate variation. In mammals, context-dependence contributes significantly to the amount of mutation rate variation. However, in some species, such as Drosophila, context-dependence has not been found (Andolfatto 2005). In these species, other sources of mutation rate variation must be accounted for.

Though the context-dependent mutation pattern we have conditioned on may not be constant across all sequences, and though there may be other sources of mutation rate variation in addition to context-dependence, our analysis and simulations suggest a consistent pattern: without taking ancestral misidentification into account during the analysis of human polymorphism data, spurious signs of positive selection will be observed.

Acknowledgments

We are grateful to Jarosław Pillardy for computational assistance, Dara G. Torgerson for helpful comments, and Rasmus Nielsen for insightful suggestions. This work was funded by a National Science Foundation grant (0516310) to CDB.
Literature Cited


John H. McDonald, Associate Editor

Accepted May 11, 2007