Published in final edited form as:

Theor Popul Biol. 2009 June ; 75(4): 346–354. doi:10.1016/j.tpb.2009.04.003.

Site Frequency Spectra from Genomic SNP Surveys

Ganeshkumar Ganapathy¹ and Marcy K. Uyenoyama²

¹ National Evolutionary Synthesis Center, 2024 W. Main Street, Suite A200, Durham, NC 27705-4667, USA gg28@duke.edu

² Department of Biology, Box 90338, Duke University, Durham, NC 27708-0338, USA, marcy@duke.edu

Abstract

Genomic survey data now permit an unprecedented level of sensitivity in the detection of departures from canonical evolutionary models, including expansions in population size and selective sweeps. Here, we examine the effects of seemingly subtle differences among sampling distributions on goodness of fit analyses of site frequency spectra constructed from single nucleotide polymorphisms. Conditioning on the observation of exactly two alleles in a random sample results in a site frequency spectrum that is independent of the scaled rate of neutral substitution (θ). Other sampling distributions, including conditioning on a single mutational event in the sample genealogy or randomly selecting a single mutation from a genealogy with multiple mutations, have distinct site frequency spectra that show highly significant departures from the predictions of the biallelic model. Some aspects of data filtering may contribute to significant departures of site frequency spectra from expectation, apart from any violation of the standard neutral model.

Keywords

site frequency spectrum; single nucleotide polymorphism; Ewens sampling formula; infinite-sites model; standard neutral model

1 Introduction

1.1 Site frequency spectra

Site frequency spectra (SFSs) are widely used to summarize patterns of genome-wide variation at the single nucleotide polymorphisms (SNPs) that abound in virtually all organisms. Fundamental population genetic analyses (Ewens 1972; Tajima 1989; Fu 1995; Griffiths and Tavaré 1998; Stephens 2000) have characterized patterns of genetic variation expected under the infinite-alleles and infinite-sites models of neutral substitution. A scaled version of those single-locus predictions now serve as the point of departure for the analysis of the SFSs comprising hundreds of thousands of independent SNP loci. Because the relative expected multiplicities depend only on sample size, departures from this expectation have been used to identify candidates for targets of selection or other locus-specific processes (for example, Kim *et al.* 2007). Numerical simulation studies (Braverman *et al.* 1995; Simonsen *et al.* 1995) have established that various phenomena, including hitchhiking and expansions in population size, affect spectrum shape, and analytical predictions now exist for a number of forms of departure

from the standard neutral model (Marth *et al.* 2004; Keightley and Eyre-Walker 2007; Živković and Wiehe 2008).

Few spectra constructed from actual genomic SNP surveys conform to expectation under the standard neutral model. For SNPs identified by direct sequencing through the NIEHS Environmental Genome Project, for example, Hernandez *et al.* (2007) noted a general excess of derived alleles in low and high multiplicities and a corresponding deficiency of alleles in intermediate frequencies. Ascertainment of SNPs through a small panel of individuals (Nielsen *et al.* 2004) introduces a different bias, toward an excess of SNPs in intermediate frequencies.

1.2 Fitting to an incorrect model

The sheer volume of information available from genomic databases confers unprecedented power to detect departures from models serving as the basis for interpretation of the data. Significant p -values may reflect departures from any aspect of a model, with some aspects fundamental to key inferences and others merely incidental.

Bishop *et al.* (1975) have presented a lucid treatment of the effect on the Pearson chi-square statistic of fitting data to an incorrect model. In a goodness of fit analysis of counts in k cells, the sample X^2 corresponds to

$$X^2 = \sum_{i=1}^k \frac{(n_i - np_i)^2}{np_i},$$

for n_i the observed count in cell i , $n (= \sum_i n_i)$ the total number of counts, and p_i the expected proportion in cell i . If the true proportions (p_i^*) of the multinomial distribution from which the observations (n_i) are sampled differ from those used to determine the expected counts (p_i), then the expectation of X^2 corresponds to

$$\mathbb{E}[X^2] = k - 1 + \sum_{i=1}^k \frac{(p_i^* - p_i)}{p_i} + (n - 1) \sum_{i=1}^k \frac{(p_i^* - p_i)^2}{p_i} \quad (1)$$

(Bishop *et al.* 1975, Section 9.6). For the number of counts (n) very large relative to the number of cells (k), as is the case for the analysis of genomic SNP data, even small departures between the true and fitted models ($(p_i^* - p_i)^2 \approx \varepsilon$) can cause the expected X^2 to exceed considerably the degrees of freedom ($df = k - 1$).

1.3 Comparison of sampling distributions

Here, we address the effect on the shape of expected site frequency spectra of SNPs of closely related models for their sampling distribution. Restriction of consideration to sample genealogies that contain a single mutational event induces a dependence of the SFS on the scaled mutation rate ($\theta = \lim 2Nu$, for N the effective number of genes and u the rate of neutral substitution). This dependence on θ implies that the SFS can provide a basis for the estimation of this fundamental parameter. Of particular significance for the detection of departures from the standard neutral model is that we expect classes of SNPs with distinct rates of neutral substitution to show distinct spectra, even in the absence of class-specific processes, including selection.

We characterize the expected shape of site frequency spectra constructed from sample genealogies that contain exactly one segregating site (neutral SNP model). A folded version of the model widely used to represent the standard neutral model (scaled multiplicity model) follows directly from the Ewens sampling formula (ESF, Ewens 1972) conditioned on the observation of exactly two alleles in the sample.

We used `ms` (Hudson 2002) to simulate 12×10^6 data sets from a non-recombining region for a range of values of θ (0.5 to 6.0). For each data set, we determined the number of segregating sites, the multiplicity of each mutation in the sample, the number of distinct haplotypes (alleles), the length of each branch, and the size of each branch (number of descendent tips). An `R` file with code for determining maximum likelihood estimates of θ and their confidence intervals is provided as Supplementary Data.

We conducted a series of goodness of fit analyses to explore the influence of various means of filtering large data sets to isolate single nucleotide polymorphisms. Our results indicate that seemingly subtle differences between the theoretical and actual sampling distributions can generate very highly significant X^2 values in tests involving the large number of observations typical of genomic SNP data, quite apart from departures from the standard neutral model.

2 Expected patterns of variation

We summarize basic descriptors of variation.

2.1 Number of segregating sites

On level l of the genealogy of a sample of genes (the segment comprising l lineages), the probability of the occurrence of a mutation more recently than a coalescence is

$$\frac{lu}{\binom{l}{2}/N+lu} = \frac{\theta}{l-1+\theta}$$

(see, for example, Ethier and Griffiths 1987). Watterson (1975) observed that the number of mutations accumulated on level l has a geometric distribution with this parameter, and gave the probability generating function (pgf) for the total number of segregating sites (S) in a sample of size n :

$$g_s(z) = \prod_{l=2}^m \frac{l-1}{l-1+\theta(1-z)}. \quad (2)$$

In particular, the expected number of segregating sites corresponds to

$$\mathbb{E}[S] = g'_s(1) = \sum_{l=2}^m \frac{\theta}{l-1}. \quad (3)$$

Tavaré (1984) has derived a simple expression for the probability mass function of S :

$$P(S=i|\theta)=\frac{m-1}{\theta}\sum_{l=2}^m(-1)^{l-2}\binom{m-2}{l-2}\left(\frac{\theta}{l-1+\theta}\right)^{i+1}. \quad (4)$$

2.2 Number of mutations with a given multiplicity

For the infinite-sites model, under which all mutations are detectable and distinguishable, Fu (1995) noted that the number of genes in a sample that bear a given mutation corresponds to the number of tips that descend from the branch of the sample gene genealogy on which the mutation arose. Fu (1995) derived the mean and variance of the number of mutations in a sample of size m that have multiplicity i (ξ_i):

$$\mathbb{E}[\xi_i]=\theta/i, \quad (5a)$$

$$\text{Var}[\xi_i]=\theta/i+\sigma_{ii}\theta^2, \quad (5b)$$

in which

$$\sigma_{ii}=\begin{cases} \beta_m(i+1) & \text{for } i < m/2 \\ 2(a_m - a_i)/(m-i) - 1/i^2 & \text{for } i = m/2 \\ \beta_m(i) - 1/i^2 & \text{for } i > m/2 \end{cases}$$

with

$$\beta_m(i)=\frac{2m(a_{m+1} - a_i)}{(m-i+1)(m-i)} - \frac{2}{m-i}.$$

The expected number of mutations present in the sample in multiplicity i (5a) scaled to the total number of mutations (3),

$$f_s(i|m)=\frac{1/i}{\sum_{j=1}^{m-1}1/j}, \quad (6)$$

is widely used to describe the expected SFS for a genomic sample of SNPs, each assumed to correspond to a mutation on an independent gene genealogy.

Expressions closely related or identical to (5) and (6) have been obtained in a variety of contexts (see especially Watterson 1974; Griffiths and Tavaré 1998). In particular, the frequency of a mutation in the sample provides information about its age (Kimura and Ohta 1973), and a number of elegant coalescence-based studies have elucidated the genealogical basis of this relationship (Griffiths and Tavaré 1998,2003; Wiuf and Donnelly 1999; Stephens 2000; Hobolth and Wiuf 2009). Tajima (1983,1989) and Griffiths and Tavaré (1998,2003) showed that (6) corresponds to the expected proportion (rather than number) of mutations that occur in multiplicity i in a sample of size m for low rates of mutation ($\theta \rightarrow 0$).

2.3 Number of alleles

For the infinite alleles model of mutation, the Ewens sampling formula provides the joint probability of the numbers in which distinct alleles appear in a sample of m genes:

$$p(\mathbf{a}) = \frac{m!}{\prod_{l=1}^m (\theta + l - 1)} \prod_{i=1}^m \binom{\theta}{i}^{a_i} \frac{1}{a_i!}, \quad (7)$$

for $\mathbf{a} = (a_1, a_2, \dots, a_m)$, with a_i the number of alleles observed exactly i times. Explicit reference to the genealogy of the sample under the standard neutral model has yielded elegant combinatorial derivations of the ESF (Kingman 1978; Donnelly 1986; Griffiths and Lessard 2005).

Ewens (1972) derived the probability mass function for the number of distinct alleles (K) observed in a sample of size m ,

$$P(K=i|\theta, m) = \frac{l_i \theta^i}{L_m(\theta)}, \quad (8)$$

for $L(\theta)$ providing the Stirling numbers of the first kind (l_i):

$$\begin{aligned} L_m(\theta) &= \theta(\theta+1) \cdots (\theta+m-1) \\ &= l_1 \theta + l_2 \theta^2 + \dots + l_m \theta^m. \end{aligned}$$

This distribution (8) has pgf

$$g_K(z) = \frac{L_m(\theta z)}{L_m(\theta)} = \prod_{l=1}^m \frac{\theta z + l - 1}{\theta + l - 1}.$$

Ewens (1972) gave the expectation and variance of the number of alleles:

$$\mathbb{E}[K] = \sum_{l=1}^m \frac{\theta}{\theta + l - 1} \quad (9a)$$

$$\text{Var}[K] = \sum_{l=1}^m \frac{\theta}{\theta + l - 1} - \sum_{l=1}^m \left(\frac{\theta}{\theta + l - 1} \right)^2. \quad (9b)$$

From the ESF (7), conditioned on the observation of a biallelic sample, we obtain a folded version of the scaled multiplicity model (6), in which the ancestral and derived alleles are not distinguished. A random sample of m genes contains exactly two haplotypes with probability

$$P(K=2) = \frac{l_2 \theta^2}{L_m(\theta)} = g''_K(0)/2 = \sum_{l=2}^m \frac{\theta}{l-1} \prod_{j=2}^m \frac{j-1}{j-1+\theta}. \quad (10)$$

Conditioning on this event, we obtain from the ESF (7) the probability of a sample containing two alleles in multiplicities i and $m-i$:

$$\begin{aligned} P(a_i=1, a_{m-i}=1|K=2) &= \frac{1/i+1/(m-i)}{\sum_{j=1}^{m-1} 1/j} \text{ for } i \neq m-i \\ P(a_{m/2}=2|K=2) &= \frac{2/m}{\sum_{j=1}^{m-1} 1/j} \text{ for } i=m-i. \end{aligned} \quad (11)$$

That θ does not appear in these expressions reflects Ewens's (1972) finding that the observed number of alleles K provides a sufficient statistic for the estimation of θ : the joint distribution of allele multiplicities (7) conditional on K is independent of θ .

2.4 Conditioning on a single segregating site

2.4.1 Neutral SNP model—Here, we use the term SNP to describe a non-recombining locus at which a single mutational event has occurred in the genealogy of a sample of genes. We describe sites at which two forms segregate in the sample as biallelic, recognizing SNPs as a subset of this group (see Table 1 for an example).

The probability that the genealogy contains a single mutation and that it lies on level l is

$$P(\text{SNP}, \delta_l=1) = \frac{\theta}{l-1+\theta} \prod_{j=2}^m \frac{j-1}{j-1+\theta},$$

for δ_l an indicator variable that takes the value 1 only if the mutation occurs on level l . Summing over levels, we confirm that the probability of a SNP is

$$P(\text{SNP}) = g'_S(0) = \sum_{l=2}^m \frac{\theta}{l-1+\theta} \prod_{j=2}^m \frac{j-1}{j-1+\theta}, \quad (12)$$

for $g_S(\cdot)$ the Watterson pgf (2) of the number of segregating sites. Comparison of (10) and (12) illustrates the close relationship between conditioning on a single segregating site and conditioning on two segregating alleles: SNPs represent a subset of biallelic polymorphisms.

Conditional on a genealogy containing a single mutation, the mutation arose on level l with probability

$$P(\delta_l=1|\text{SNP}) = \frac{\frac{1}{l-1+\theta}}{\sum_{j=2}^m \frac{1}{j-1+\theta}}.$$

Under our neutral SNP model, a SNP-defining mutation occurs in exactly i of the m sampled genes with probability

$$f_n(i|m, \theta) = \frac{\sum_{l=2}^{m-i+1} \frac{1}{\theta+l-1} \binom{m-i-1}{l-2}}{\sum_{j=2}^m \frac{1}{\theta+j-1}} = \frac{\frac{1}{i} \sum_{l=2}^{m-i+1} \frac{l-1}{\theta+l-1} \binom{m-l}{i-1}}{\binom{m-1}{i} \sum_{j=2}^m \frac{1}{\theta+j-1}}, \quad (13)$$

using Eq. (14) of Fu (1995).

In contrast with (6) and (11), this expression depends on θ . In the limit as the rate of neutral substitution becomes small ($\theta \rightarrow 0$), (13) converges to (6). Otherwise, we expect classes of SNPs that differ with respect to the rate of neutral substitution to show different site frequency spectra, even under the standard neutral model.

Figure 1 illustrates, for a sample of 19 genes, that the site frequency spectra expected under (6) and (13) show close correspondence for low rates of neutral substitution ($\theta = 0.01$) but that the neutral SNP model (13) predicts more rare and fewer common derived alleles for large mutation rates ($\theta = 10$). For samples of size $m = 19$ and the range of θ values in our simulations, the numbers of singletons and doubletons expected under the neutral SNP model (13) increase monotonically with θ and multiplicities 4 through 18 decrease monotonically, with the expectation for multiplicity 3 varying non-monotonically.

Like (5) and (6), the expressions shown here have been obtained previously in various contexts (e.g., Griffiths and Tavaré 1998,2003; Stephens 2000; Hobolth *et al.* 2008). The relationship between the frequency in the sample of a mutation and the level of the genealogy on which it arose has been exploited to address the distribution of the age of a mutation (Griffiths and Tavaré 1998; Wiuf and Donnelly 1999; Stephens 2000), and (13) can be obtained by rearranging equation (28) of Stephens (2000). Griffiths and Tavaré (2003) showed that the generalization of (13) to accommodate variable population size reduces to (6) under constant population size and low mutation rate ($\theta \rightarrow 0$).

2.4.2 Estimating θ —Construction of the spectrum expected under the neutral SNP model (13) requires an estimate of θ . In our goodness of fit analyses to the expected counts under the neutral SNP model (13), we substituted the maximum-likelihood estimate (MLE) of θ and reduced the degrees of freedom by one.

For n the number of SNP loci observed, D the observed spectrum of derived allele multiplicities, and T the total number of nucleotide sites, the likelihood of θ corresponds to

$$P(D, n|T, \theta) = P(D|n, T, \theta)P(n|T, \theta). \quad (14)$$

We model each derived allele count in Table 2 as the realization of an independent Poisson random variable, which implies that the total number of counts $n (= \sum_i n_i)$ also has a Poisson distribution:

$$P(n=k|T, \theta) = \frac{\lambda^k e^{-\lambda}}{k!},$$

for λ the expected number of SNPs observed,

$$\lambda = TP(\text{SNP}),$$

with $P(\text{SNP})$ given by (12). Conditional on n (sum of counts in a given row of Table 2), the joint distribution of multiplicities $P(D/n, T, \theta)$ is multinomial (see, for example, Bishop *et al.* 1975, Chap. 13).

3 Patterns of variation in simulated data

We used `ms` (Hudson 2002) to simulate 10^6 data sets of size $m = 19$ for each of 12 assignments of θ (0.5 to 6.0, in increments of 0.5).

3.1 Magnitude of segregating variation

Table 1 presents the number among the 10^6 19-gene sample genealogies, simulated under the indicated value of θ , that contained zero, exactly one, or more than one mutation. Also shown are the numbers of samples comprising exactly two alleles and the proportion of those that contained more than one mutation. As the rate of neutral substitution increases, the proportion of samples with a single mutation declines and the percentage of two-haplotype samples that contain more than one mutation tends to increase.

Table 2 shows the multiplicities of mutations at loci for which the sample genealogy contained a single mutational event (mutation number equal to 1 in Table 1).

3.2 Number of mutations having a given multiplicity

Fu's (1995) analysis addressed the number (rather than proportion) of mutations that have a given multiplicity in a random sample genealogy. For each of the 10^6 sample genealogies simulated under a given assignment of θ , we distinguished among all mutations, in accordance with the infinite sites model assumed by Fu (1995), and determined the total number that occurred in each multiplicity. Table 3 indicates an excellent fit to the analytical mean and variance (5) of data simulated under $\theta = 1.0$, and other values of θ gave similar results.

Figure 2 shows the number of trees simulated under $\theta = 6.0$ which contained the number of mutations indicated on the abscissa in multiplicity 5. This distribution matches the analytical expressions (5) in mean and variance, but has a pronounced right skew. Fu (1996) developed a test based on a Hotelling-like statistic, with critical values determined by simulation, as a means of using the frequency spectrum to detect departures from the standard neutral model.

3.3 Estimates of θ

We compare estimates of θ inferred from the number of segregating sites (Watterson 1975), the number of alleles (Ewens 1972), and the site frequency spectrum (14).

3.3.1 Number of segregating sites—We found excellent agreement between the Watterson distribution (4) and the total number of segregating sites (S) observed among the 10^6 trees simulated under each value of θ (analyses not shown).

In a Bayesian context, we examined the posterior distribution of θ based on the number of segregating sites:

$$P(\theta|S) = \frac{P(S|\theta)P(\theta)}{P(S)}.$$

Assuming a prior $P(\theta)$ taking a uniform distribution over $[0.01, 100]$ and zero probability elsewhere, we rescaled the likelihood function (4) implied for each of the 10^6 sample genealogies generated under a given assignment of θ to obtain a posterior distribution of θ and determined its 95% credible interval. Table 4 compares the posterior mode and percentage of credible intervals that contained the actual value of θ . For values of θ greater than 0.5, the observed number of segregating sites appears to provide lower than expected coverage probabilities and overestimates of θ .

3.3.2 Number of alleles—Table 5 indicates excellent corroboration of the expressions (9) given by Ewens (1972) for the mean and variance of the number of distinct alleles (K). Figure 3 shows a close match of the entire simulated distribution of K to expectation (8). Ewens (1972) showed that the third and fourth moments of the distribution approach zero for large sample size (m), and Fig. 3 suggests a Gaussian-like shape for even $m = 19$ under the larger values of θ .

Table 4 gives the posterior mode and the proportion of credible intervals that contained the actual value of θ inferred from the number of alleles (8), assuming again a uniform prior distribution over $[0.01, 100]$ for θ . Both aspects appear to show trends similar to those for estimates based on the number of segregating sites, with the overestimation of θ more noticeable for large values of θ .

3.3.3 Site frequency spectra from sites with single mutations—For each row in Table 2, we used (14), with n equal to the row sum (number of SNP loci) and $T = 10^6$, to obtain a maximum likelihood estimate (MLE) of θ . Table 6 presents the MLEs, their approximate 95% confidence intervals (2 log-likelihood units around the mode), and the number of trees on which the estimates are based (from the single mutation column of Table 1). The higher uncertainty of estimates for data sets generated for values of θ equal to 4.5 or greater likely reflects inadequacy of the Yates continuity correction for the low number of loci.

To examine the coverage probabilities of Bayesian credible intervals based on the site frequency spectrum, we partitioned the 10^6 trees simulated under a given value of θ into 100 groups of 10^4 trees and conducted a separate analysis on each group, assuming as before a uniform prior distribution over $[0.01, 100]$. Table 4 presents the posterior modes, the proportion of credible intervals that included the actual value of θ , and the numbers of trees on which the estimates were based. We did not explore values of θ greater than 3.0 due to the low number of trees containing a single segregating site. As for the MLEs in Table 6, the width of the average credible interval increased with θ and the number of trees used in the estimation declined. For θ up to 3.0, basing the estimation on the site frequency spectrum appears to improve accuracy over using only the number of segregating sites (4) or only the number of alleles (8). For higher values of θ , the low number of trees per group appeared to have compromised the estimation of both the mode and the coverage probability.

4 Goodness of fit of site frequency spectra

4.1 Data restricted to sites with a single mutational event

For the subset of simulated sample genealogies that contained a single segregating site (Table 2), we computed the Pearson chi-square statistic under the scaled multiplicity (6) and neutral SNP (13) models, using the MLEs for θ given in Table 6 for the latter. To avoid large departures of the counts from approximate continuity (see Section 3.3.3), we excluded from consideration data sets generated for values of θ greater than 4.0, for which the expected counts in cells representing the highest multiplicities fell below 5.

Figure 4 indicates highly significant departures from the scaled multiplicity model (6). As expected for a fit to an incorrect model (1), the X^2 values tend to increase with number (n) of loci contributing to the SFS. In contrast, Figure 5, showing the fit to the correct model (single mutational event), indicates no obvious relationship between the total number of counts and the X^2 values obtained under the neutral SNP model (13). While the low X^2 values for most of the range under $\theta = 2.5$ and high X^2 values under $\theta = 3.5$ in Fig. 5 appear unusual, the simulated SNP data appear to lend much greater support to the correct neutral SNP model (13) than to the scaled multiplicity model (6).

Figure 6 shows the empirical cumulative distribution of p -values from goodness of fit tests to the neutral SNP model (13) for θ up to 2.0. For each value of θ , we partitioned the 10^6 simulated trees into 100 groups of 10^4 simulated trees. Elimination of non-SNP sites (trees showing a number of mutations different from 1) reduced the number of loci to the values indicated in the rightmost column of Table 4. For each partition, we determined the p -value associated with the spectrum constructed from those single-mutation trees. A perfect sample would show equality between the cumulative distribution of p -values and the nominal significance level (diagonal line). Figure 6 suggests a tendency toward false positives (Type I error) for $\theta = 0.5$.

4.2 Data restricted to biallelic sites

As noted in Section 2.3, the ESF (7) links the allele frequency spectrum to the expected number of mutations that occur in a sample in a given multiplicity (5a): the ESF, conditioned on the observation of exactly two alleles in the sample (11), corresponds to the folded version of the scaled multiplicity model (6).

Figure 7 shows for assignments of θ up to 2.5, the X^2 values obtained in goodness of fit tests to the folded scaled multiplicity model (11) as a function of the number of sample genealogies considered (n). We refrained from analyzing data simulated for assignments for larger values of θ , for which the expected number of counts in at least one cell fell below 5. These plots suggest no obvious increase in the X^2 values with total count number as would be expected under fitting to an incorrect model (1). Of possible concern in a number of cases is an unusually low X^2 value, indicating too good a fit.

Application of goodness of fit tests to the neutral SNP model (13), incorrectly regarding the two alleles segregating in each sample as defined by a single mutational event, underestimates θ and generates highly significant X^2 values (Table 7). As Fig. 1 would suggest, the neutral SNP model predicts too many samples containing a rare allele and a corresponding deficiency of samples with the two alleles in comparable frequencies. Although the expectations under the neutral SNP model (13) converge to those under the scaled multiplicity model (6) as θ becomes small, Table 1 indicates that even for $\theta = 0.5$, a substantial fraction (nearly 17%) of the simulated biallelic data sets contained more than a single mutation.

4.3 Other data sets showing a poor fit to the scaled multiplicity model

We also explored the effects of other kinds of data filtering, including construction of the spectrum of all mutations (no filtering) and restriction of consideration to a single random mutation in each tree or to a single random branch in each tree. These data partitions all showed a significantly poor fit to the scaled multiplicity model (6).

4.3.1 Random mutation—For each simulated sample genealogy that contained at least one mutation, we determined the multiplicity in the sample of a mutation chosen uniformly at random, without weighting by frequency in the sample. This subset contains all trees in Table 1 except those with zero segregating sites. We found increasing sample X^2 values with the number of trees, as in Fig. 4, and a marked departure between the nominal significance level

and the cumulative distribution of p -values. Both aspects indicate a very poor fit to the scaled multiplicity model (6).

4.3.2 Size of a random branch—For a given simulated genealogy, we sampled a branch at random, weighting by length relative to the total length of the tree, and determined the number of tips descendent from that branch. While the number of descendants is not generally observable, this experiment permits us to examine branch size apart from the additional stochastic process of mutation. As the scaled neutral substitution rate θ does not affect the size of a branch, our *ms* output provides a total of 12×10^6 simulated samples suitable for this analysis. A very highly significant departure of simulated spectra from the spectrum expected under the scaled multiplicity model (6) is evident even from subsets of the data. For example, Table 8 compares the observed branch sizes and the expectations under the scaled multiplicity model (6) for a subsample of 10^6 trees. Relative to predictions from the scaled multiplicity model (6), the observed trees showed an excess of branches of size 1 through 4 and a deficiency of branches of all larger sizes. A fit to the scaled multiplicity model would indicate a particularly large excess (18,371) of singletons (terminal branches 1), with this multiplicity class contributing almost 35% of the total X^2 value (3412, with $df=17$), even in the absence of population expansions or selective sweeps.

4.3.3 All segregating sites—Although the relative numbers of alleles or mutations in a given sample genealogy are of course correlated (7), we constructed the spectrum of all mutations observed in many sample trees generated under a given value of θ , expecting correlations between mutations on the same tree to be dwarfed by the large number of independent trees. We partitioned the 10^6 samples simulated under each value of θ into 100 groups of 10^4 trees and determined the p -value associated with a fit to the scaled multiplicity model (6). The empirical cumulative distribution of p -values for the 1,200 tests using all 12×10^6 simulated trees indicated a very poor fit, with 80% of the tests giving p -values less than 0.04.

5 Discussion

5.1 Dependence of SNP site frequency spectra on θ

Fundamental to the interpretation of site frequency spectra constructed from surveys of genomic variation are the analyses of Ewens (1972) and Fu (1995) for the infinite-alleles and infinite-sites model of mutation, respectively. We note that the folded version of the scaled multiplicity model (6) can be obtained directly from the ESF (7), conditioned on biallelic samples (11). A striking property of the ESF (7) is that the allele frequency spectrum conditioned on the number of observed alleles (K) is independent of θ (Ewens 1972). This property is shared by the scaled multiplicity model (6), which is widely used to represent the expected SFS under the standard neutral model for genome-wide SNP data.

In contrast, the observed SFS for sample genealogies that contain a single mutational event (13) does in fact provide information about θ , implying that site frequency spectra can provide a basis for the estimation of this fundamental parameter. Tables 4 and 6 indicate high accuracy of both Bayesian and maximum likelihood estimates of θ for data sets comprising 10^3 or more SNPs. Liu *et al.* (2009) have recently developed a generalized least squares method for estimating θ using Fu's (1995) expressions for the means and covariances (5) of the counts (rather than proportions) of mutations across multiplicities.

Our analysis suggests that differences between the sampling distributions imposed by restriction to biallelic variation (11) and by restriction to single mutational events (13) are readily detectable in analyses incorporating the volume of data typical of genomic SNP surveys. SNPs constitute a subset of biallelic polymorphisms, and the probability of a single segregating

site (12) is nearly identical to the probability of a biallelic polymorphism (10). However, our simulated data (Table 1) illustrate that for large θ a substantial proportion of biallelic polymorphisms may comprise multiple segregating mutations in the genealogy of the sample. Incorrect application of the neutral SNP model (13), which assumes a single segregating site, results in substantial underestimation of θ and highly significant X^2 values in goodness of fit analyses (Table 7). Similarly, data restricted to sample genealogies containing a single segregating site show highly significant departures (Fig. 4) from the scaled multiplicity model (6) or the biallelic model (11), while giving strong support to the correct neutral SNP model (Figs. 5 and 6).

As has been shown in various contexts (Griffiths and Tavaré 1998, 2003; Stephens 2000), the neutral SNP model (13) reduces to the scaled multiplicity model (6) in the limit of low rates of neutral substitution ($\theta \rightarrow 0$). As θ increases, samples conditioned on a single segregating site show more rare mutations and fewer common mutations (Fig. 1). Because an excess of rare variants is an iconic signature of selective sweeps or expansions in effective population size (Braverman *et al.* 1995; Simonsen *et al.* 1995), our results suggest that careful consideration of the sampling distribution of genomic variation may help to avoid unwarranted inferences about the operation of locus-specific evolutionary processes.

5.2 Departures from the scaled multiplicity model

We found that sampling distributions that show only subtle numerical and conceptual departures from canonical models can be very strongly rejected on the basis of sufficiently large numbers of observations. Describing kinds of neutral variation that show poor fits to the scaled multiplicity model (6) may be as valuable as confirming the fit to expectation of data generated under the correct model.

Single nucleotide polymorphisms may be considered similar to polymorphisms due to a single mutational event in a sample genealogy, regardless of the total number of events in the tree. However, we observed very highly significant deviations from the scaled multiplicity model (6) of site frequency spectra generated by extracting a single random mutation from simulated trees. Similarly, the distributions of the size (number of descendent tips) of a randomly chosen branch and the multiplicities of all segregating sites do not fit the scaled multiplicity model (Section 4.3).

5.3 Modelling actual SNP data

We found that the allele frequency spectra generated by restricting simulated data to biallelic samples fit the folded scaled multiplicity model (11) very well (Figure 7) and the neutral SNP model (13) very poorly (Table 7). Conversely, simulated data restricted to sample genealogies that contained a single mutational event gave strong support to the neutral SNP model and strongly rejected the scaled multiplicity model (Section 4.1).

We suggest that neither model describes actual SNP data. Segregation of exactly two nucleotide bases in a sample may reflect multiple independent substitutions of the same derived base for the ancestral base or back mutations, in violation of both the infinite-sites (5) and infinite-alleles models (7) as well as the neutral SNP model (13). Of the standard models of mutation of population genetics, SNPs may conform most closely to a finite-sites, K -allele model, in which new mutations assume one of four states (A, C, G, or T). Whether actual SNPs show site frequency spectra similar to those expected for the standard neutral model under this mutation process awaits further analytical and statistical development.

Acknowledgments

In a lifetime of work, Sam Karlin transformed several entire fields. It continues to be an honor to learn from him, to draw upon the part of his work that extended into evolutionary biology, and to contribute to this memorial volume. We are grateful to the anonymous reviewers for valuable comments and references to key works, Asger Hobolth for important insights, and Benjamin D. Redelings for questioning the interpretation of SNP spectra. Support from the National Evolutionary Synthesis Center (NESCent), Durham, NC, for a NESCent Postdoctoral Fellowship to GG and for the Genomic Introgression working group is gratefully acknowledged. Public Health Service grant GM 37841 (MKU) provided partial support for this research.

References

- Bishop, YMM.; Fienberg, SE.; Holland, PW. Discrete multivariate analysis: Theory and practice. The MIT Press; 1975.
- Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W. The hitchhiking effect on the site frequency spectrum of DNA polymorphism. *Genetics* 1995;140:783–796. [PubMed: 7498754]
- Donnelly P. Partition structures, Polya urns, the Ewens sampling formula, and the ages of alleles. *Theor Pop Biol* 1986;30:271–288. [PubMed: 3787504]
- Ethier SN, Griffiths RC. The infinitely-many-sites model as a measure-valued diffusion. *Ann Probab* 1987;15:515–545.
- Ewens WJ. The sampling theory of selectively neutral alleles. *Theor Pop Biol* 1972;3:87–112. [PubMed: 4667078]
- Fu YX. Statistical properties of segregating sites. *Theor Pop Biol* 1995;48:172–197. [PubMed: 7482370]
- Fu YX. New statistical tests of neutrality for DNA samples from a population. *Genetics* 1996;143:557–570. [PubMed: 8722804]
- Griffiths RC, Lessard S. Ewens' sampling formula and related formulae: combinatorial proofs, extensions to variable population size and applications to ages of alleles. *Theor Pop Biol* 2005;68:167–177. [PubMed: 15913688]
- Griffiths RC, Tavaré S. The age of a mutation in a general coalescent tree. *Commun Statist – Stochastic Models* 1998;14:273–295.
- Griffiths, RC.; Tavaré, S. The genealogy of a neutral mutation. In: Green, PJ.; Hjort, NL.; Richardson, S., editors. *Highly structured stochastic systems*. Oxford Univ. Press; Oxford: 2003. p. 393–412. chapter 13
- Hernandez RD, Williamson S, Bustamante CD. Context dependence, ancestral misidentification, and spurious signatures of natural selection. *Mol Biol Evol* 2007;24:1792–1800. [PubMed: 17545186]
- Hobolth A, Uyenoyama MK, Wiuf C. Importance sampling for the infinite sites model. *Statistical Applications in Genetics and Molecular Biology* 2008;7:Article 32
- Hobolth A, Wiuf C. The genealogy, site frequency spectrum and ages of two nested mutant alleles. *Theor Pop Biol*. 2009;this volume
- Hudson RR. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 2002;18:337–338. [PubMed: 11847089]
- Keightley PD, Eyre-Walker A. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics* 2007;177:2251–2261. [PubMed: 18073430]
- Kim S, Plagnol V, Hu TT, Toomajian C, Clark RM, Ossowski S, Ecker JR, Weigel D, Nordborg M. Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nat Genet* 2007;39:1151–1155. [PubMed: 17676040]
- Kimura M, Ohta T. The age of a neutral mutant persisting in a finite population. *Genetics* 1973;75:199–212. [PubMed: 4762875]
- Kingman JFC. Random partitions in population genetics. *Proc R Soc Lond A* 1978;361:1–20.
- Liu X, Maxwell TJ, Boerwinkle E, Fu YX. Inferring population mutation rate and sequencing error rate using the SNP frequency spectrum in a sample of DNA sequences. *Mol Biol Evol*. 2009;in press
- Marth GT, Czabarka E, Murvai J, Sherry ST. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics* 2004;166:351–372. [PubMed: 15020430]

- Nielsen R, Hubisz MJ, Clark AG. Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. *Genetics* 2004;168:2372–2382.
- Simonsen KL, Churchill GA, Aquadro CF. Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics* 1995;141:413–429. [PubMed: 8536987]
- Stephens M. Times on trees, and the age of an allele. *Theor Pop Biol* 2000;57:109–119. [PubMed: 10792976]
- Tajima F. Evolutionary relationship of DNA sequences in finite populations. *Genetics* 1983;105:437–460. [PubMed: 6628982]
- Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 1989;123:585–595. [PubMed: 2513255]
- Tavaré S. Line-of-descent and genealogical processes, and their applications in population genetics models. *Theor Pop Biol* 1984;26:119–164. [PubMed: 6505980]
- Živković D, Wiehe T. Second-order moments of segregating sites under variable population size. *Genetics* 2008;180:341–357. [PubMed: 18716326]
- Watterson GA. The sampling theory of selectively neutral alleles. *Adv Appl Prob* 1974;6:463–488.
- Watterson GA. On the number of segregating sites in genetical models without recombination. *Theor Pop Biol* 1975;7:256–276. [PubMed: 1145509]
- Wiuf C, Donnelly P. Conditional genealogies and the age of a neutral mutant. *Theor Pop Biol* 1999;56:183–201. [PubMed: 10544068]

Figure 1.

Histogram (bars) of the multiplicities of derived SNP alleles expected in trees with a single segregating site (13) compared to the expectation (curve) under (6), for low (left, $\theta = 0.01$) and high (right, $\theta = 10$) scaled rates of neutral substitution.

Figure 2.

Histogram of the number, among 10^6 samples of size $m = 19$ simulated under $\theta = 6.0$, that contain the number of mutations on the abscissa in multiplicity 5.

Figure 3.

Histograms (bars) of the observed number of trees that contain the number of alleles indicated on the abscissa compared to the analytical distribution (curves) from (8).

Figure 4.

Sample Pearson chi-square values for the fit to the scaled multiplicity model (6) of SFSs generated from simulated SNP data (single mutational event) for samples of $m = 19$ genes. Shown on the abscissa are the numbers of sets of 10,000 loci in the group contributing to the spectrum, with 100 corresponding to all 10^6 simulated genealogies. The solid line indicates the average X^2 value and the dashed line corresponds to the expectation for a fit to the correct model (df=17, for multiplicities of the derived allele of 1 through 18).

Figure 5.

Sample Pearson chi-square values for the fit to the neutral SNP model (13) of SFSs generated from simulated SNP data (single mutational event) for samples of $m = 19$ genes. The dashed horizontal line indicates the expectation ($df=16$, after estimation of θ), with other features as described for Fig. 4.

Figure 6.

Empirical cumulative distribution of p -values obtained for 100 site frequency spectra for the neutral SNP model (13) applied to SNP data (single segregating site) simulated under the four indicated values of θ . The diagonal line represents the nominal significance level.

Figure 7.

Sample Pearson chi-square values for the fit to the folded scaled multiplicity model (11) of biallelic frequency spectra generated under the indicated θ value. The solid horizontal line represents the average X^2 value and the dashed line the expectation ($df=8$).

Table 1

Genetic variation in samples of size 19

θ	Mutation number			Biallelic		
	0	1 ^a	> 1	Counts ^b	% > 1 mut ^c	
0.5	204,420	297,831	497,749	357,998	16.8	
1.0	52,737	133,862	813,401	183,449	27.0	
1.5	15,867	54,129	930,004	82,534	34.4	
2.0	5,254	21,975	972,771	36,763	40.2	
2.5	1,866	9,121	989,013	16,555	44.9	
3.0	776	4,190	995,034	8,029	47.8	
3.5	306	1,845	997,849	3,908	52.8	
4.0	128	890	998,982	1,980	55.1	
4.5	66	415	999,519	1,020	59.3	
5.0	29	208	999,763	519	60.0	
5.5	7	99	999,894	273	63.7	
6.0	3	66	999,931	155	57.4	

^aTrees containing a single mutational event

^bSamples containing exactly two haplotypes

^c% of trees with more than 1 mutation

Table 2

Derived allele counts in sample genealogies containing a single mutation

θ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
0.5	94,296	45,003	29,118	21,294	16,587	13,523	11,475	9,755	8,457	7,582	6,770	6,159	5,570	5,162	4,746	4,348	4,150	3,836
1.0	44,968	21,130	13,334	9,467	7,409	5,908	4,761	4,173	3,572	3,117	2,720	2,483	2,290	1,932	1,844	1,712	1,610	1,432
1.5	18,913	8,857	5,487	3,755	2,914	2,326	1,831	1,535	1,361	1,230	1,016	958	803	774	671	644	553	501
2.0	8,060	3,638	2,180	1,526	1,122	865	727	608	518	442	399	380	343	303	259	227	192	186
2.5	3,460	1,509	914	602	484	362	273	243	229	185	160	142	118	110	104	83	76	67
3.0	1,561	727	424	267	209	179	132	124	104	70	69	58	58	47	57	36	46	22
3.5	719	310	184	147	91	71	58	43	44	25	28	27	20	20	11	21	19	7
4.0	327	179	96	49	46	34	29	27	14	15	16	4	12	12	6	10	9	5
4.5	168	67	45	31	24	13	13	8	6	7	5	6	3	4	5	1	6	3
5.0	87	38	23	9	8	10	7	3	4	4	6	5	1	1	2	0	0	0
5.5	36	16	14	9	5	3	3	0	1	1	2	0	2	2	2	2	0	1
6.0	22	15	9	4	2	2	2	1	1	1	2	2	0	1	1	1	0	0

Table 3

Mean and variance of the number of mutations with the indicated multiplicity under the infinite-sites model

Multiplicity	Mean		Variance	
	Observed ^a	Expected ^b	Observed ^a	Expected ^c
1	1.0025	1.0000	1.200	1.199
2	0.5001	0.5000	0.661	0.661
3	0.3332	0.3333	0.471	0.471
4	0.2492	0.2500	0.371	0.372
5	0.2009	0.2000	0.311	0.310
6	0.1664	0.1667	0.267	0.267
7	0.1429	0.1429	0.236	0.236
8	0.1244	0.1250	0.210	0.212
9	0.1102	0.1111	0.190	0.192
10	0.1007	0.1000	0.173	0.171
11	0.0908	0.0909	0.159	0.159
12	0.0833	0.0833	0.148	0.149
13	0.0774	0.0769	0.142	0.140
14	0.0718	0.0714	0.134	0.132
15	0.0658	0.0667	0.123	0.125
16	0.0623	0.0625	0.118	0.119
17	0.0590	0.0588	0.114	0.113
18	0.0556	0.0556	0.108	0.108

^a Among 10^6 simulated trees with $\theta = 1.0$ ^b From (5a)^c From (5b)

Table 4

Bayesian estimates of θ and coverage probabilities

Actual θ	Segregating sites ^a		Allele number ^b		SNP spectrum ^c		Number of loci ^f
	Mode	Coverage ^d	Mode	Coverage ^d	Mode	Coverage ^e	
0.5	0.511	95.2	0.559	95.6	0.489	97	2,978.3
1.0	1.027	94.5	1.109	91.3	1.001	99	1,338.6
1.5	1.548	92.9	1.657	88.7	1.501	99	541.3
2.0	2.075	94.1	2.211	90.7	2.004	93	219.8
2.5	2.604	93.6	2.769	92.8	2.516	95	91.2
3.0	3.129	94.3	3.324	94.2	2.990	97	41.9
3.5	3.664	94.1	3.889	90.8			
4.0	4.196	93.8	4.461	93.1			
4.5	4.730	94.4	5.028	90.6			
5.0	5.259	93.3	5.604	94.7			
5.5	5.796	93.1	6.185	91.4			
6.0	6.332	93.1	6.781	89.2			

^aFrom (4), based on 10^6 loci
^bFrom (8), based on 10^6 loci
^cFrom (13), based on 100 groups of 10^4 loci
^d% of 10^6 95% credible intervals that contained the true value
^eNumber out of 100 95% credible intervals that contained the true value
^fAverage number of loci (of 10^5 examined) per credible interval

Table 5Observed^a and expected^b moments of allele number

θ	Mean		Variance	
	Observed	Expected	Observed	Expected
0.5	2.454	2.454	1.231	1.233
1.0	3.549	3.548	1.952	1.954
1.5	4.438	4.439	2.447	2.448
2.0	5.196	5.195	2.807	2.811
2.5	5.857	5.854	3.081	3.086
3.0	6.433	6.436	3.301	3.300
3.5	6.957	6.958	3.469	3.468
4.0	7.431	7.430	3.608	3.600
4.5	7.858	7.860	3.705	3.704
5.0	8.252	8.255	3.789	3.785
5.5	8.617	8.619	3.848	3.849
6.0	8.959	8.956	3.900	3.897

^a Among 10^6 simulated trees^b From (9)

Table 6Maximum likelihood estimate of θ and approximate confidence interval

Actual θ	MLE ^a	95% CI ^b	Tree number ^c
0.5	0.499	(0.496, 0.502)	297,831
1.0	1.001	(0.998, 1.004)	133,862
1.5	1.500	(1.495, 1.505)	54,129
2.0	2.003	(1.995, 2.010)	21,975
2.5	2.513	(2.500, 2.525)	9,121
3.0	2.984	(2.965, 3.003)	4,190
3.5	3.506	(3.476, 3.537)	1,845
4.0	3.991	(3.946, 4.037)	890
4.5	4.523	(4.455, 4.595)	415
5.0	5.027	(4.927, 5.133)	208
5.5	5.591	(5.440, 5.753)	99

^aBased on (14)^bSpanning 2 log likelihood units^cSample trees with a single segregating site

Table 7

Biallelic data fit to the neutral SNP model

Actual θ	MLE ^a	χ^2 of fit ^b
0.5	0.272	634.7
1.0	0.817	2098.8
1.5	1.265	1809.5
2.0	1.711	1159.6
2.5	2.161	725.8
3.0	2.585	433.0
3.5	3.024	251.8
4.0	3.456	171.8
4.5	3.895	82.5
5.0	4.361	54.8
5.5	4.821	51.8
6.0	5.243	22.8

^aFrom (14)^bTo (13) with df=16

Table 8

Observed and expected branch size distributions

Branch size ^a	Observed ^b	Expected ^c	Δ^d
1	304,485	286,114	18,371
2	148,412	143,057	5,355
3	96,429	95,371	1,058
4	71,645	71,529	116
5	56,205	57,223	-1,018
6	46,430	47,686	-1,256
7	39,769	40,873	-1,104
8	33,833	35,764	-1,931
9	29,937	31,790	-1,853
10	26,607	28,611	-2,004
11	23,834	26,010	-2,176
12	22,045	23,843	-1,798
13	20,317	22,009	-1,692
14	18,495	20,437	-1,942
15	17,004	19,074	-2,070
16	15,778	17,882	-2,104
17	15,005	16,830	-1,825
18	13,770	15,895	-2,125

^aNumber of descendant tips of a random branch^bAmong 10⁶ simulated trees^cFrom the scaled multiplicity model (6)^d Δ = Observed – Expected