

Site Frequency Spectra from Genomic SNP Surveys

Ganeshkumar Ganapathy¹

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

The unprecedented volume of data generated by genomic surveys permits detection of subtle deviations from analytical models widely used to represent canonical evolutionary processes, including genetic drift and mutation under the standard neutral model, 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 (SNPs). 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 segregating mutation from a genealogy with multiple mutations, engender distinct site frequency spectra that show highly significant departures from the predictions of the biallelic model. Some aspects of data acquisition may contribute to the prevalence of significant departures of site frequency spectra from expectation, even apart from any violations of the standard neutral model.

1 Introduction

1.1 Site frequency spectra

Since the advent of genomic surveys of variation, site frequency spectra (SFSs) have widely been used to summarize the pattern of variation at the single nucleotide polymorphism (SNP) sites that abound in the genomes of virtually all organisms. Fundamental population genetic analyses (Ewens 1972; Fu 1995) have characterized patterns of genetic variation expected under the standard neutral model. 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 classes of loci as 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, affect spectrum shape, and analytical expressions for the multiplicities of mutations at independent loci 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 actual spectra constructed from genomic SNP surveys conform to the scaled expected multiplicities. 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 in the spectra constructed from SNPs identified by direct sequencing through the NIEHS Environmental Genome Project. Ascertainment of SNPs through sequencing of a small panel (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 the underlying model that serves as the basis for interpretation of the data. Significant p -values may reflect departures from any aspect of the model, with some aspects fundamental to the key inferences under study and others merely incidental.

Bishop *et al.* (1975) have presented a lucid treatment of the effect on the Pearson chi square statistic of assessing the fit of 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 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). This expression implies that the Pearson chi square statistic has an expectation equal to the degrees of freedom only for a fit to the true model ($p_i^* = p_i$). Otherwise ($(p_i^* - p_i)^2 > 0$), the expectation grows with n for n large relative to the number of cells (k), as is virtually always the case for the analysis of genomic SNP data.

1.3 Comparison of sampling distributions

Here, we address the effect on interpretations of site frequency spectra of SNPs of closely related models of their origin. We show that restriction of consideration to sample genealogies that contain a single segregating site constitutes a form of ascertainment bias that introduces 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 first characterize the expected shape of site frequency spectra constructed from sample genealogies that contain exactly one segregating site (neutral SNP model). We show that

a folded version of the scaled multiplicity model follows directly from the Ewens sampling formula conditioned on the observation of exactly two alleles in the sample. Using simulated data from a non-recombining region generated by the `ms` program of Hudson (2002) under the infinite-sites model, we conduct a series of goodness of fit analyses to explore the differences between the neutral SNP model and the scaled multiplicity model. 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 describe the distributions of 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

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

Watterson (1975) observed that the number of mutations accumulated on level l has a geometric distribution with this parameter, and gave the pgf for the total number of segregating sites in a sample of size n :

$$g_S(a) = \prod_{l=2}^m \frac{l - 1}{l - 1 + \theta(1 - a)}. \quad (3)$$

In particular, the expected number of segregating sites corresponds to

$$\mathbb{E}[S] = g'_S(1) = \sum_{l=2}^m \frac{\theta}{l - 1}. \quad (4)$$

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

$$P(S = i|\theta) = \frac{m-1}{\theta} \sum_{l=2}^m (-1)^{l-2} \binom{m-2}{l-2} \left(\frac{\theta}{l-1+\theta} \right)^{i+1}. \quad (5)$$

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. Using simple and elegant arguments, Fu (1995) derived the mean and variance of the number of mutations in a sample of size m that have multiplicity i (ξ_i):

$$\mathbb{E}[\xi_i] = \theta/i, \quad (6a)$$

$$\text{Var}[\xi_i] = \theta/i + \sigma_{ii}\theta^2, \quad (6b)$$

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 in multiplicity i (6a) scaled to the total number of mutations (4),

$$f_s(i|m) = \frac{1/i}{\sum_{j=1}^{m-1} 1/j}, \quad (7)$$

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

2.3 Number of alleles

For the infinite alleles model of mutation, the Ewens sampling formula (ESF, Ewens 1972) provides the joint probability of the numbers in which distinct alleles appear in a sample of m genes:

$$p(\mathbf{a}) = \frac{m!}{\theta(\theta + 1) \dots (\theta + m - 1)} \prod_{i=1}^m \left(\frac{\theta}{i}\right)^{a_i} \frac{1}{a_i!}, \quad (8)$$

for $\theta = 2Nu$, u the per-generation rate of neutral mutation, and $\mathbf{a} = (a_1, a_2, \dots, a_m)$, 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 in the sample (K),

$$P(K = i|\theta) = \frac{l_i \theta^i}{L(\theta)}, \quad (9)$$

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

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

This distribution has pgf

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

The expectation and variance of the number of alleles are

$$\mathbb{E}[K] = \sum_{l=1}^m \frac{\theta}{\theta + l - 1} \quad (10a)$$

$$\text{Var}[K] = \sum_{l=1}^m \frac{\theta}{\theta + l - 1} - \sum_{l=1}^m \left(\frac{\theta}{\theta + l - 1}\right)^2. \quad (10b)$$

(Ewens 1972).

From the ESF (8), conditioned on the observation of a biallelic sample, we obtain a folded version of the scaled multiplicity model (7), in which the ancestral and derived alleles

are not distinguished. A random sample of m genes contains exactly two haplotypes with probability

$$P(K = 2) = g_K''(0)/2 = \frac{l_2\theta^2}{L(\theta)} = \sum_{l=2}^m \frac{\theta}{l-1} \prod_{j=2}^m \frac{j-1}{j-1+\theta}. \quad (11)$$

Conditioning on this event, we obtain from (8) the probability of a sample containing two alleles in multiplicities i and $m-i$:

$$\begin{aligned} P(a_i = 1, a_{m-i} = 1 | K = 2) &= \frac{1/i + 1/(m-i)}{\sum_{j=2}^m 1/(j-1)} \quad \text{for } i \neq m/2 \\ P(a_{m/2} = 2 | K = 2) &= \frac{2/m}{\sum_{j=2}^m 1/(j-1)}. \end{aligned} \quad (12)$$

That θ does not appear in these expressions reflects that the observed number of alleles K provides a sufficient statistic for the estimation of θ (Ewens 1972): the joint distribution of allele multiplicities (8) conditional on K is independent of θ .

2.4 Accounting for ascertainment as a SNP

2.4.1 Neutral SNP model

Here, we use SNP to describe a site at which a single mutational event has occurred in the genealogy of a sample of genes from a non-recombining locus. We describe sites at which two forms segregate in the sample as biallelic, recognizing SNPs as a subset of this group.

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

$$P(\text{SNP}) = g_S'(0) = \sum_{l=2}^m \frac{\theta}{l-1+\theta} \prod_{j=2}^m \frac{j-1}{j-1+\theta}, \quad (13)$$

for $g_S(a)$ the Watterson pgf (3) of the number of segregating sites. Comparison with (11) confirms the close relationship between conditioning on a single segregating site and conditioning on two segregating alleles: SNPs represent a subset of biallelic polymorphisms (see Table 1 for an example).

Conditional on having sampled 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

$$f_n(i|m, \theta) = \frac{\sum_{l=2}^{m-i+1} \frac{1}{\theta+l-1} \frac{\binom{m-i-1}{l-2}}{\binom{m-1}{l-1}}}{\sum_{j=2}^m \frac{1}{\theta+j-1}} = \frac{\frac{1}{i} \sum_{l=2}^{m-i+1} \frac{l-1}{\theta+l-1} \binom{m-l}{i-1}}{\binom{m-1}{i} \sum_{j=2}^m \frac{1}{\theta+j-1}}, \quad (14)$$

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

In contrast with (7) and (12), this expression depends on θ . In the limit as the rate of neutral substitution becomes small ($\theta \rightarrow 0$), (14) converges to (7). 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.

2.4.2 Estimating θ

Construction of the spectrum expected under the neutral SNP model (14) requires an estimate of θ . In our goodness of fit analyses to the expected counts under the neutral SNP model (14), we substituted the maximum-likelihood estimate (MLE) of θ and reduced the degrees of freedom by one.

For T the total number of nucleotide sites examined and n the number of SNP loci among the T sites (mutation number 1 in Table 1), the likelihood of θ corresponds to

$$P(D, n|T, \theta) = P(D|n, T, \theta)P(n|T, \theta), \quad (15)$$

for D the observed spectrum of allele multiplicities. 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 also has a Poisson distribution:

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

for λ the expected number of SNPs observed:

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

and $P(\text{SNP})$ given by (13). 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).

Figure 1 illustrates, for a sample of 19 genes, that the site frequency spectra expected under (7) and (14) show close correspondence for low rates of neutral substitution ($\theta = 0.01$) but that our correction for ascertainment as a SNP (14) predicts more rare and fewer common derived alleles for large mutation rates ($\theta = 10$). For samples of the size ($m = 19$) in our simulated data, (14) indicates that the expected numbers of singletons and doubletons increase monotonically with θ , decrease monotonically for multiplicities 4 through 18, and the expectation for multiplicity 3 shows non-monotonic changes over the range of θ values assigned in our simulations (0.5 through 6.0 in increments of 0.5).

3 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). 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 number of tips descendent from each branch.

3.1 Magnitude of segregating variation

Table 1 presents the number among the 10^6 genealogies comprising 19 genes generated 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 haplotypes and the proportion of those that represent 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 increases.

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

3.2 Estimates of θ

We compared aspects of the variation in the simulated data set to some basic predictions (Section 2).

3.2.1 Number of segregating sites

Figure 2 indicates excellent agreement between the total number of segregating sites and the Watterson distribution (5) among the 10^6 trees simulated under each value of θ .

In a Bayesian context, we also 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 (5) 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 3 gives the average posterior mode and the proportion of credible intervals that contained the actual value of θ .

3.2.2 Number of alleles

A comparison of the observed distribution of the number of distinct alleles (K) to (9) also indicated close agreement (Figure 3). Ewens (1972) showed that the higher moments of the distribution of K approach zero for large sample size (m), and Fig. 3 suggests a Gaussian-like shape at even $m = 19$ for the larger values of θ .

Table 4 indicates close corroboration of expressions (10) given by Ewens (1972) for the mean and variance of K . As in the case of the number of segregating sites (5), we examined the posterior distribution of θ based on the number of alleles (9), assuming as before a prior $P(\theta)$ with a uniform distribution over $[0.01, 100]$ and zero probability elsewhere. Table 5 gives the average posterior mode and the proportion of credible intervals that contained the actual value of θ .

3.2.3 Site frequency spectra from SNP data

For each row in Table 1, we used (15), with n equal to the row sum (number of SNP loci) and $T = 10^6$, to obtain an MLE of θ . Table 6 presents the MLEs and their approximate 95% confidence intervals, estimated as 2 log-likelihood units from the mode. The higher uncertainty of estimates for data sets generated under values of θ equal to 4.5 or higher likely reflect inadequacy of the Yates correction for continuity. For this range of θ values, only some hundred or fewer of the observed data sets (a hundredth percent or less of the 10^6 simulated trees) contained a single mutation, with the expected counts under (14) of samples with the derived allele in the highest multiplicities falling below 5.

4 Patterns of variation

4.1 Goodness of fit to SNP frequency spectra

For the subset of simulated sample genealogies that contained a single mutational event (Table 2), we computed the Pearson’s chi square value under the scaled (7) and neutral SNP (14) 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 3.2), we excluded from consideration data sets generated for values of θ equal to 4.5 or higher.

Figure 4 indicates the highly significant chi square values obtained for the scaled multiplicity model (7). As the number of sample genealogies on which the SFS is based (n) increases, the X^2 values tend to increase, as expected for a fit to an incorrect model (1). In contrast, Figure 5 indicates no obvious relationship between the X^2 values obtained under the neutral SNP model (14) and the number of loci (n).

Figure 6 shows a comparison of the empirical cumulative distribution function of p -values associated with the X^2 values under the neutral SNP model (14) to the nominal significance level (for example, the proportion of spectra showing a p -value less than or equal to 0.01). For each value of θ , we partitioned the 10^6 simulated trees into 10^2 groups of 10^4 simulated trees. After eliminating non-SNP sites (trees showing a number of mutations different from 1), we determined the p -value associated with the spectrum constructed from each partition. Each partition contains a different number of trees: for $\theta = 0.5$, for example, the counts in Table 1 indicates that each partition contains on the order of 3,000 trees while for $\theta = 2.0$, the average partition contains about 200 trees.

In Figure 5, the low X^2 values for most of the range under $\theta = 2.5$ and high X^2 values under $\theta = 3.5$ appear unusual. Also, Figure 6 shows an apparent departure between the empirical and nominal cumulative distributions of p -values for $\theta = 0.5$. Even so, the simulated SNP data appear to lend greater support to the neutral SNP model (14) than to the scaled

multiplicity model (7).

4.2 Data showing poor fits to the scaled multiplicity model

To explore sampling distributions, we conducted goodness of fit tests on subsets of the simulated data. In spite of their apparent similarity to SNP spectra, spectra generated from a random mutation, a random branch, or all segregating sites showed poor fits to the scaled multiplicity model (7).

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. Because the `ms` program assigns the position of each new mutation from a uniform distribution, we selected the as the focal mutation the mutation with the lowest position number. This subset contains all trees in Table 1 except those with mutation number zero. For simulated data sets generated under the assignment of θ as 0.5, Figure 7 shows increasing X^2 values with amount of data (compare Fig. 5) and a strong departure of linearity between the nominal significance level and the sample p -value (compare Fig. 6). Both aspects indicate a very poor fit to the scaled multiplicity model (7). Analyses of data simulated under the other θ values studied show similar results.

Size of a random branch: For a given simulated sample genealogy, we sampled a branch at random, weighted by branch length relative to the total length of the tree, and determined the number of its descendants in the sample. The distribution of the number of descendants would not of course be observable in actual data, but this experiment permits us to examine branch size apart from the stochastic process of mutation. As the number of descendants of a branch does not depend on the scaled neutral substitution rate θ , our `ms` output provides a total of 12×10^6 simulated samples suitable for this analysis. As is the

case for spectra for a single random mutation (Fig. 7), a very highly significant departure of simulated spectra from the spectrum expected under the scaled multiplicity model (7) is evident from even small subsets of the data. For example, Table 7 compares the observed branch sizes and the expectations under the scaled multiplicity model (7) for a sample of 10^6 trees. Predictions from the scaled multiplicity model (7) show a deficiency of branches of size 1 through 4 and an excess of branches of all larger sizes. The scaled multiplicity model predicts a particularly large deficiency (18,371) of terminal branches (size 1), with this multiplicity class contributing almost 35% of the total X^2 value (3412, with $df=17$).

All segregating sites: Although the numbers of mutations of different multiplicities observed on a given sample genealogy are expected to be correlated (Fu 1995), we examined the shape of the spectrum of multiplicities of all mutations observed in many sample trees, expecting correlations between mutations on the same tree to be dwarfed by the large number of independent trees. For each spectrum, corresponding to 10,000 trees generated under a given value of θ , we determined the p -value obtained from testing goodness of fit to the scaled multiplicity model (7). Figure 8, showing the empirical cumulative distribution of p -values from 1,200 tests and a total of all 12×10^6 simulated trees, indicates a very poor fit.

4.3 Number of mutations having a given multiplicity

Although the scaled expectation (7) is widely used as the expected site frequency spectrum for SNPs, the main results (6) of Fu (1995) in fact correspond to the number of mutations that have a given multiplicity in a random sample genealogy.

Moments: To confirm (6), we determined, for each of the 10^6 sample genealogies simulated under a given assignment of θ , whether at least one mutation in the tree occurred in a specified multiplicity. Figure 9 shows a histogram of the number of trees simulated under $\theta = 6.0$ which contained the number of mutations indicated on the abscissa in multiplicity

5. The strong right skew of the distribution makes challenging hypothesis testing using a moments approach, although Fu (1996) has suggested using a Hotelling-like statistic as a means of detecting departures from the expected numbers of mutations across multiplicities.

Table 8 indicates an excellent fit of the observed number of mutations in a given multiplicity to the analytical mean and variance (6).

Goodness of fit of two-allele spectra: As noted in Section 2.3, the ESF (8) links the allele frequency spectrum to the expected number of mutations that occur in a sample in a given multiplicity (6a): the ESF, conditioned on the observation of exactly two alleles in the sample (12), corresponds to the folded version of the scaled multiplicity model (7).

Figure 10 shows for a range of assignments of θ values, the X^2 values obtained for goodness of fit analyses to the folded scaled multiplicity model (12) as a function of the number of sample genealogies considered (n). We refrained from analyzing data simulated under values of θ of 3 and larger, for which the expected number of counts in at least one cell falls below 5. These plots suggest no obvious increase in the X^2 values with n as would be expected under fitting to an incorrect model (1). Of possible concern is the unusually low X^2 values obtained, indicating a fit too close to expected.

Although the expectations under the neutral SNP model (14) converge to those under the scaled multiplicity model (7) 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. Accordingly, goodness of fit analyses to the neutral SNP model (14), incorrectly regarding the two haplotypes segregating in each sample as defined by a single mutational event, gives highly significant X^2 values (Table 9) and underestimates θ . 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.

5 Discussion

5.1 Dependence of SNP site frequency spectra on θ

Fundamental to the interpretation of 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. Of relevance to analyses of biallelic SNP data, we find that the folded version of the scaled multiplicity model (7) studied by Fu (1995) can be obtained directly from the ESF (8), conditioned on biallelic samples (12). A striking property of the ESF (8) 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 (7), which is widely used to represent the standard neutral model.

In contrast, the site frequency spectrum of variants observed in samples containing a single segregating site (14) does in fact provide information about θ beyond that contained in the number of polymorphisms. Examination of our simulated data indicates that the number of polymorphisms and the total number of sites (n and T) alone provide an excellent estimate of θ , with an improvement on the order of 0.1% contributed by the additional information provided by the full spectrum (15). Even so, the dependence on θ of site frequency spectra under the neutral SNP model (14) represents a new property of SNP sampling distributions. Our analysis suggests that differences between the sampling distributions imposed by restriction to biallelic variation (12) and by restriction to single mutational events (14) are readily detectable in analyses incorporating the volume of data typical of genomic SNP surveys.

Although SNPs constitute a subset of biallelic polymorphisms and the probability of a SNP (13) is nearly identical to the probability of a biallelic polymorphism (11), our simulated data (Table 1) indicate that a substantial proportion of biallelic polymorphisms may comprise multiple segregating mutations in the genealogy of the sample. Incorrect application of the neutral SNP model (14), which assumes a single segregating site, results in substantial

underestimation of θ and highly significant X^2 values in goodness of fit analyses (Table 9). Similarly, data restricted to sample genealogies containing a single segregating site show highly significant departures (Fig. 4) from the scaled multiplicity model 7 or the biallelic model (12), while giving strong support to the neutral SNP model (Figs. 5 and 6).

In the limit of low rates of neutral substitution ($\theta \rightarrow 0$), the SFS for the neutral SNP model (14) reduces to the scaled multiplicity model (7). 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 numbers among the signatures of selective sweeps or expansions in effective population size (Braverman *et al.* 1995; Simonsen *et al.* 1995), our results suggest that a careful consideration of the sampling distribution of genomic variation may reduce the incidence of unwarranted inferences about the operation of locus-specific evolutionary processes.

5.2 Departures from the scaled multiplicity model

Because even subtle departures from canonical sampling distributions can be detected using the magnitude of data typical of genomic studies, we suggest that a description of kinds of neutral variation that show poor fits to the scaled multiplicity model (7) may be valuable.

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, Fig. 7 indicates that site frequency spectra generated by extracting a single random mutation from each simulated tree containing at least one segregating site show very significant deviations from the scaled multiplicity model (7). Similarly, the distributions of the size (number of descendent tips) of a randomly chosen branch (weighted by length) and the multiplicities of all segregating sites depart from the scaled multiplicity model (Section 4.2).

While these distributions show subtle numerical and conceptual departures from the frequency distribution of single mutations in sample genealogies that contain no other mutations, they can be very strongly rejected on the basis of sufficiently large numbers of observations.

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 (12) very well (Figure 10) and the neutral SNP model (14) very poorly (Table 9). 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 in a sample of exactly two nucleotide bases may reflect multiple independent substitutions of the same derived base for the ancestral base, in violation of the infinite-alleles assumption of the ESF (8) and the folded scaled multiplicity model (12). Single nucleotide polymorphisms may also reflect multiple substitutions at the same site in a single line of descent, in violation of the infinite-sites assumption of the neutral SNP model (14). Of the standard population genetic models of mutation, 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 under this mutation process the standard neutral model awaits further analytical and statistical development.

Acknowledgments

In a lifetime of work, Sam Karlin influenced several entire fields. It was an honor to learn from him and to marvel over the part of his work that extended into evolutionary biology.

We are indebted to the editors for the opportunity to contribute to this memorial volume. We thank 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, Y. M. M., Fienberg, S. E., and Holland, P. W., 1975. Discrete multivariate analysis: Theory and practice. The MIT Press.
- Braverman, J. M., Hudson, R. R., Kaplan, N. L., Langley, C. H., and Stephan, W., 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphism. *Genetics* **140**, 783–796.
- Donnelly, P., 1986. Partition structures, Polya urns, the Ewens sampling formula, and the ages of alleles. *Theor. Pop. Biol.* **30**, 271–288.
- Ewens, W. J., 1972. The sampling theory of selectively neutral alleles. *Theor. Pop. Biol.* **3**, 87–112.
- Fu, Y.-X., 1995. Statistical properties of segregating sites. *Theor. Pop. Biol.* **48**, 172–197.
- Fu, Y.-X., 1996. New statistical tests of neutrality for DNA samples from a population. *Genetics* **143**, 557–570.
- Griffiths, R. C. and Lessard, S., 2005. Ewens’ sampling formula and related formulae: combinatorial proofs, extensions to variable population size and applications to ages of alleles. *Theor. Pop. Biol.* **68**, 167–177.

- Hernandez, R. D., Williamson, S., and Bustamante, C. D., 2007. Context dependence, ancestral misidentification, and spurious signatures of natural selection. *Mol. Biol. Evol.* **24**, 1792–1800.
- Hudson, R. R., 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337–338.
- Keightley, P. D. and Eyre-Walker, A., 2007. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics* **177**, 2251–2261.
- Kim, S., Plagnol, V., Hu, T. T., Toomajian, C., Clark, R. M., Ossowski, S., Ecker, J. R., Weigel, D., and Nordborg, M., 2007. Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nat. Genet.* **39**, 1151–1155.
- Kingman, J. F. C., 1978. Random partitions in population genetics. *Proc. R. Soc. Lond. A* **361**, 1–20.
- Marth, G. T., Czabarka, E., Murvai, J., and Sherry, S. T., 2004. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics* **166**, 351–372.
- Nielsen, R., Hubisz, M. J., and Clark, A. G., 2004. Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. *Genetics* **168**, 2372–2382.
- Simonsen, K. L., Churchill, G. A., and Aquadro, C. F., 1995. Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics* **141**, 413–429.
- Tavaré, S., 1984. Line-of-descent and genealogical processes, and their applications in population genetics models. *Theor. Pop. Biol.* **26**, 119–164.

Živković, D. and Wiehe, T., 2008. Second-order moments of segregating sites under variable population size. *Genetics* **180**, 341–357.

Watterson, G. A., 1975. On the number of segregating sites in genetical models without recombination. *Theor. Pop. Biol.* **7**, 256–276.

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

^a SNP loci

^b samples 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: Bayesian estimates of θ
from the number of segregating sites**

Actual θ	Posterior mode	Coverage proportion (%) ^a
0.5	0.511	95.16
1.0	1.027	94.52
1.5	1.548	92.93
2.0	2.075	94.06
2.5	2.604	93.61
3.0	3.129	94.31
3.5	3.664	94.07
4.0	4.196	93.80
4.5	4.730	94.35
5.0	5.259	93.28
5.5	5.796	93.12
6.0	6.332	93.05

^a percentage of 95% credible intervals that contain θ

**Table 4: Observed and expected
moments of allele number**

θ	Mean		Variance	
	observed	expected	observed	expected
0.5	2.453534	2.454032	1.231106	1.233486
1.0	3.549408	3.547740	1.951957	1.954076
1.5	4.438375	4.439020	2.446917	2.448191
2.0	5.195786	5.195479	2.807311	2.810826
2.5	5.856663	5.853651	3.080711	3.086477
3.0	6.433185	6.436076	3.301061	3.300199
3.5	6.957187	6.957902	3.469166	3.467741
4.0	7.431186	7.429920	3.608442	3.599747
4.5	7.858271	7.860159	3.705250	3.703772
5.0	8.252192	8.254791	3.788577	3.785386
5.5	8.616990	8.618681	3.847789	3.848809
6.0	8.958781	8.955749	3.900116	3.897307

**Table 5: Bayesian estimates of θ
from the number of alleles**

Actual θ	Posterior mode	Coverage proportion (%) ^a
0.5	0.559	95.6
1.0	1.109	91.3
1.5	1.657	88.7
2.0	2.211	90.7
2.5	2.769	92.8
3.0	3.324	94.2
3.5	3.889	90.8
4.0	4.461	93.1
4.5	5.028	90.6
5.0	5.604	94.7
5.5	6.185	91.4
6.0	6.781	89.2

^a percentage of 95% credible intervals that contain θ

**Table 6: Estimate of θ
and approximate confidence interval**

Actual θ	MLE ^a of θ	95% CI ^b
0.5	0.499	(0.496, 0.502)
1.0	1.001	(0.998, 1.004)
1.5	1.500	(1.495, 1.505)
2.0	2.003	(1.995, 2.010)
2.5	2.513	(2.500, 2.525)
3.0	2.984	(2.965, 3.003)
3.5	3.506	(3.476, 3.537)
4.0	3.991	(3.946, 4.037)
4.5	4.523	(4.455, 4.595)
5.0	5.027	(4.927, 5.133)
5.5	5.591	(5.440, 5.753)

^a based on (15)

^b spanning 2 log likelihood units

**Table 7: Observed and expected
branch size distributions**

Branch size ^a	Observed ^b	Expected ^c	Δ ^d
1	304,485	28,6114	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

^a number of descendant tips of a random branch

^b among 10^6 simulated samples of size 19

^c from the scaled multiplicity model (7)

^d $\Delta = \text{Observed} - \text{Expected}$

**Table 8: Mean and variance of the
number of mutations with the indicated multiplicity**

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 in 10^6 simulated samples of size 19 with $\theta = 1.0$

^b from (6a)

^c from (6b)

**Table 9: Fit of biallelic data
to the neutral SNP model**

Actual θ	MLE ^a of θ	X^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

^a from (15)

^b to neutral SNP model (14)

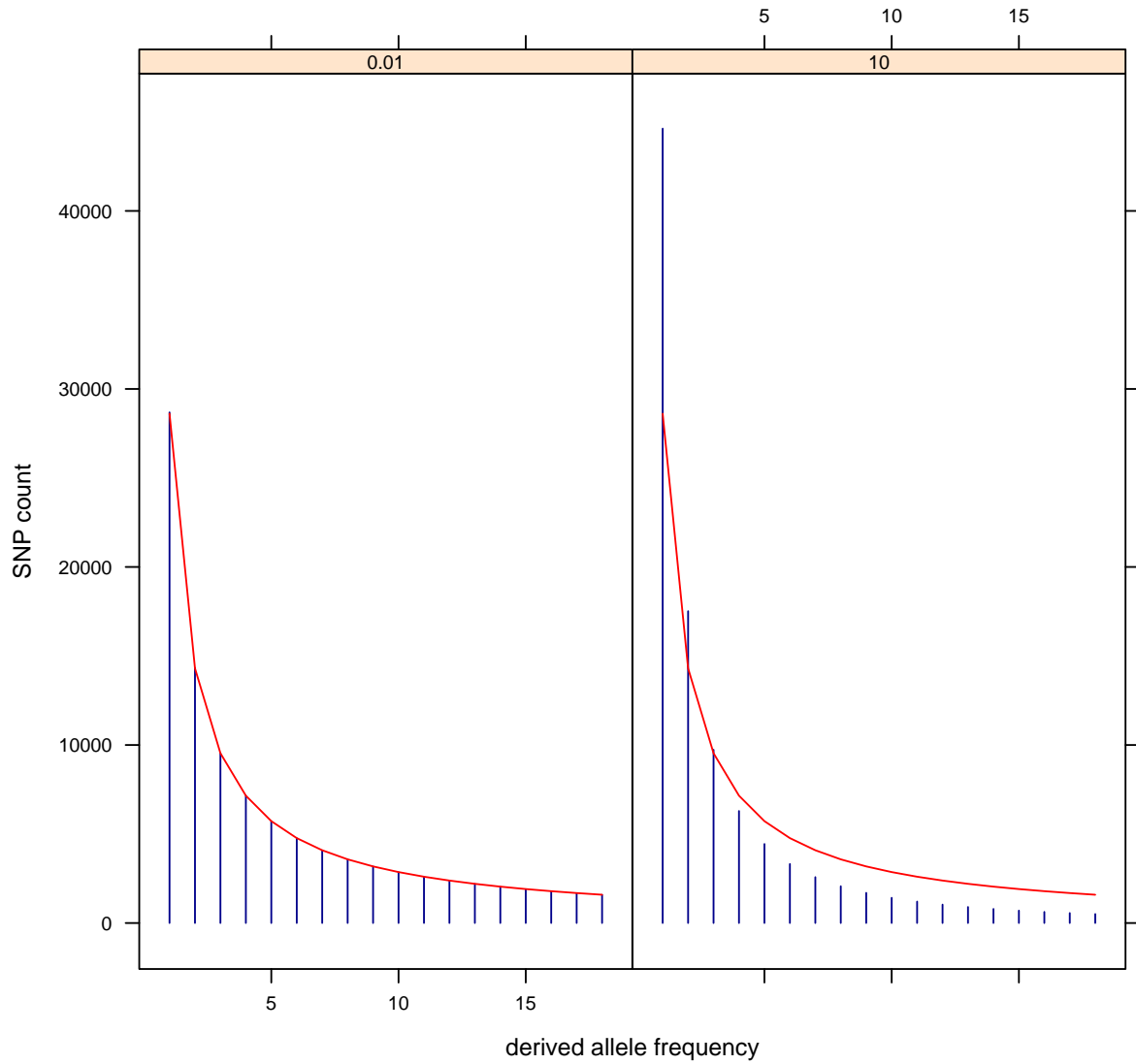


Figure 1: Histogram (blue) of the multiplicities of derived SNP alleles expected under correction for ascertainment as a SNP (14) compared to the expectation (red) under (7) under low (left, $\theta = 0.01$) and high (right, $\theta = 10$) neutral substitution rates.

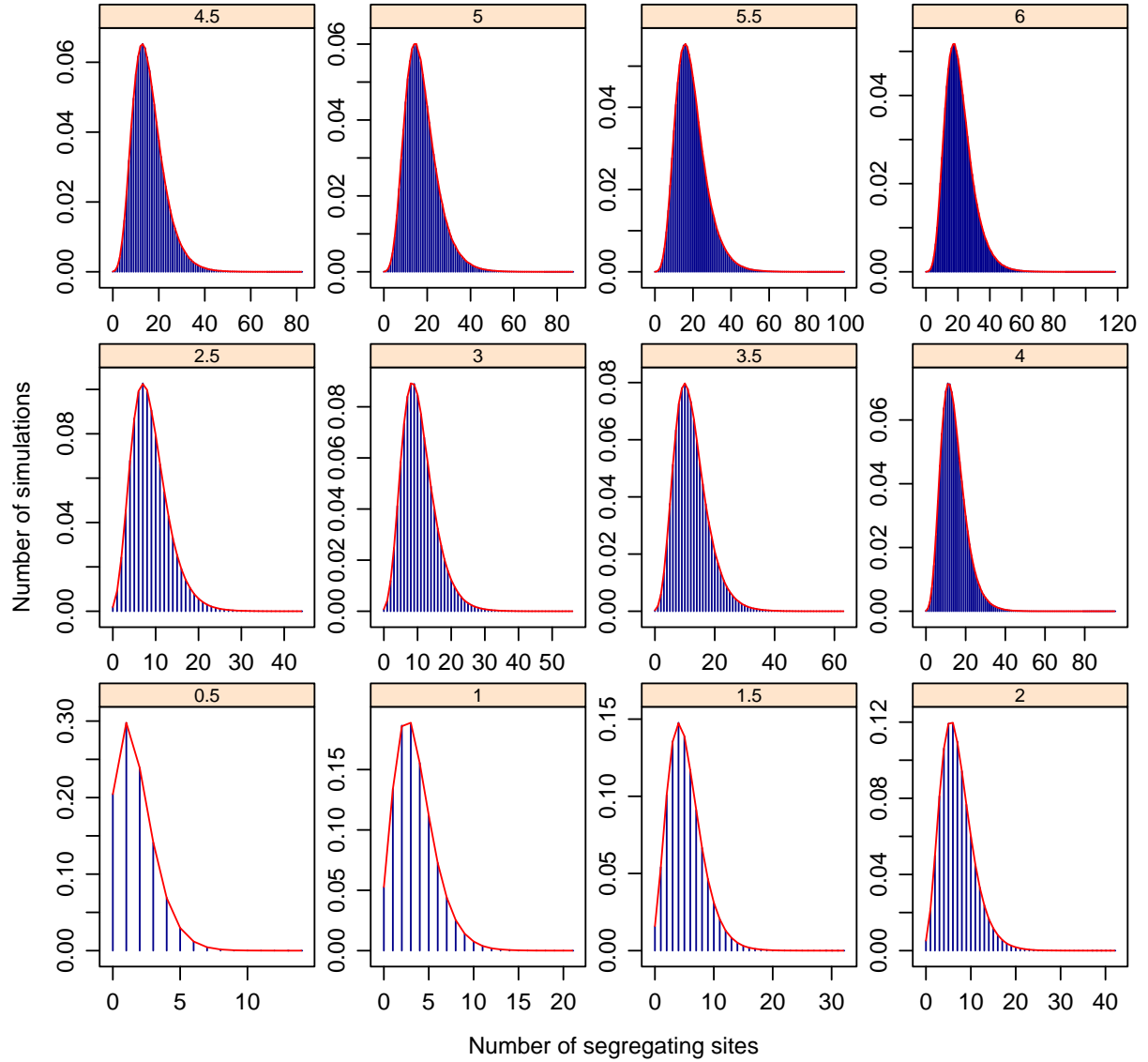


Figure 2: Histograms of observed number of trees that contain the number of segregating sites indicated on the abscissa compared to (5).

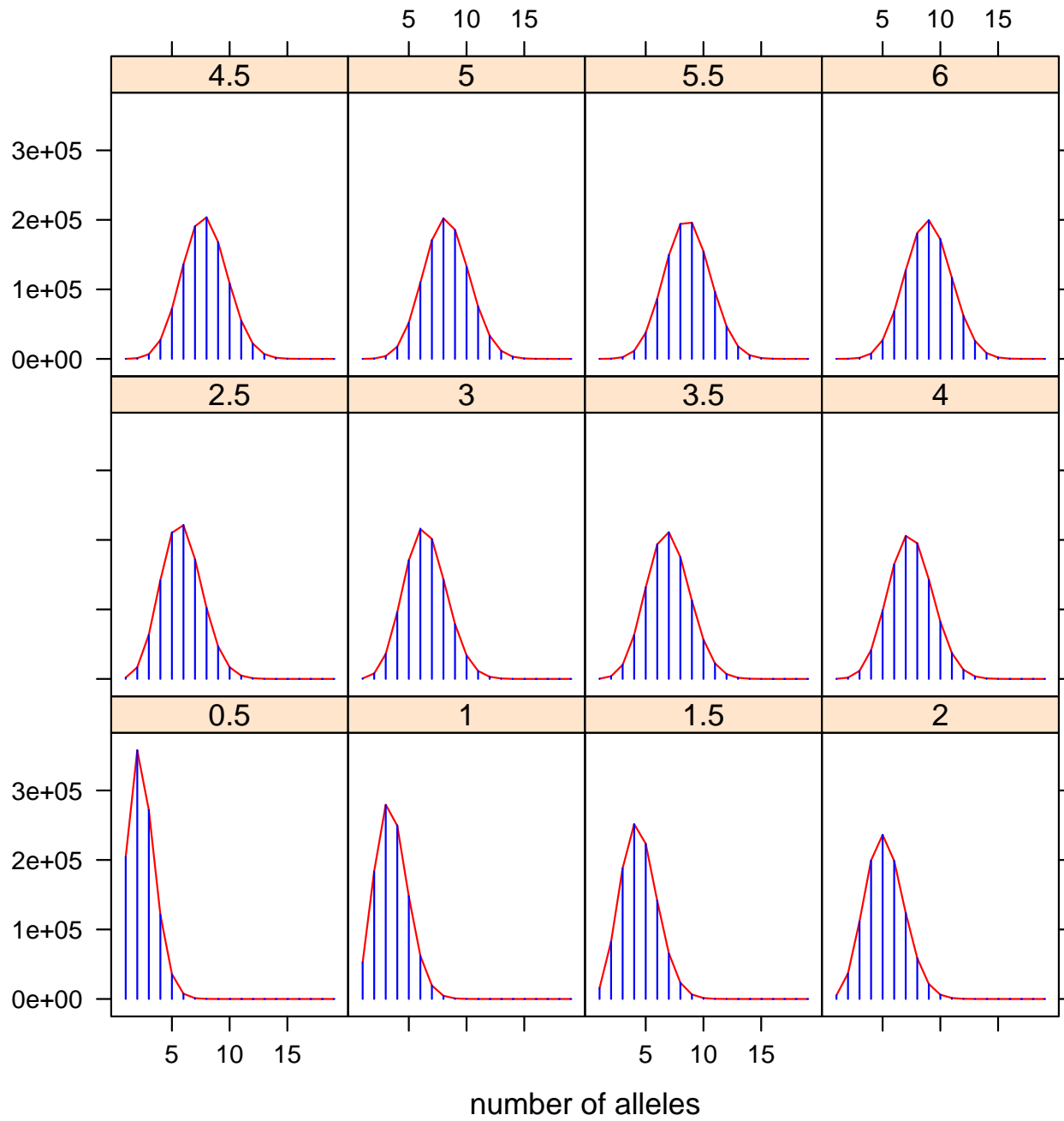


Figure 3: Histograms of observed number of trees that contain the number of alleles indicated on the abscissa compared to (9).

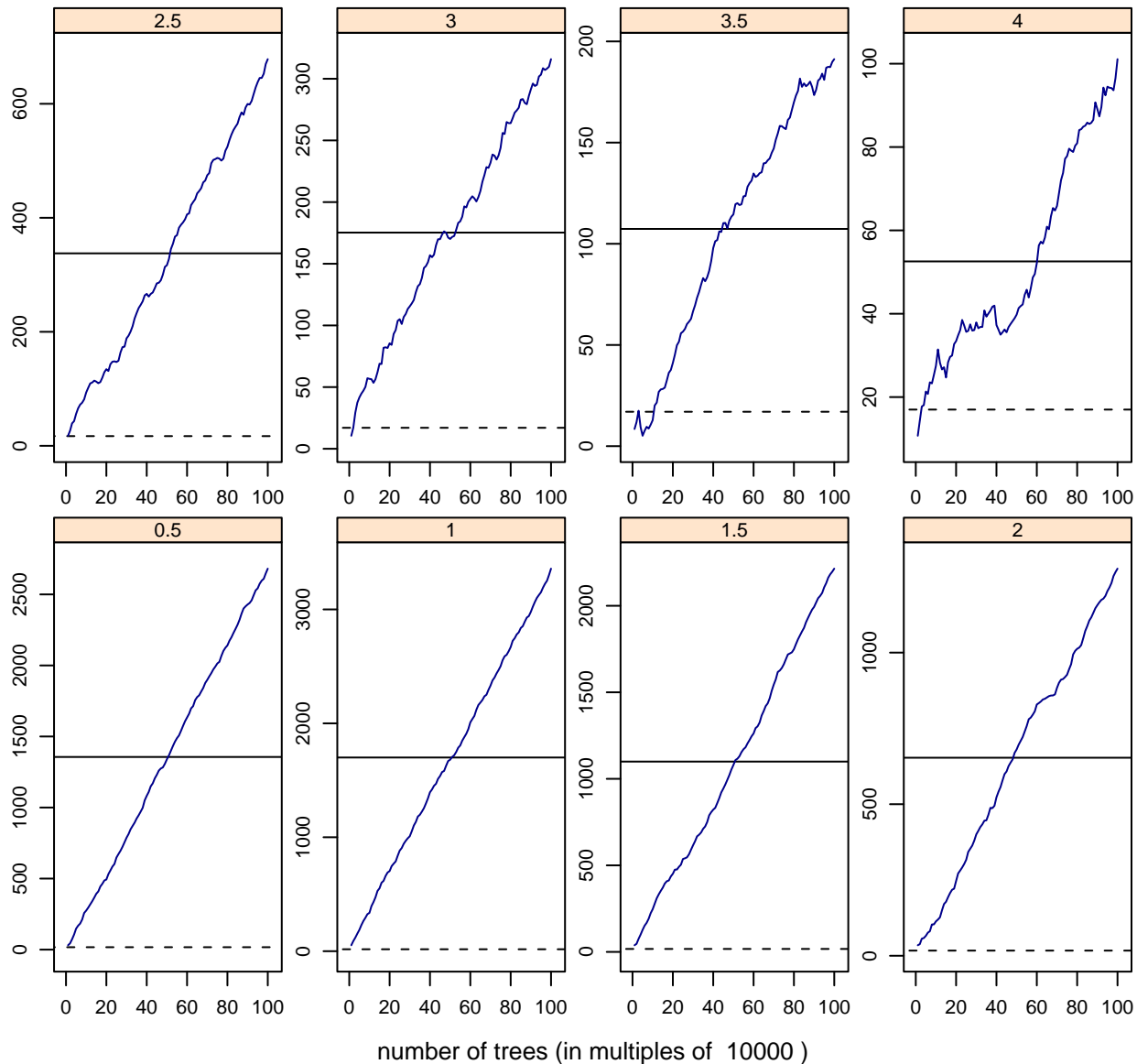


Figure 4: Sample X^2 values for the scaled multiplicity model (7) as a function of the number of loci (n) used to construct the SFS for SNP data (single segregating site) for samples of $m = 19$ genes. Shown on the abscissa are the numbers of sets of 10,000 trees in the group contributing to the spectrum, with 100 corresponding to all 10^6 simulated genealogies. The solid line indicates the average X^2 value and the dotted line corresponds to the expectation for fitting to a correct model (degrees of freedom=17, for multiplicities of the derived allele of 1 through 18).

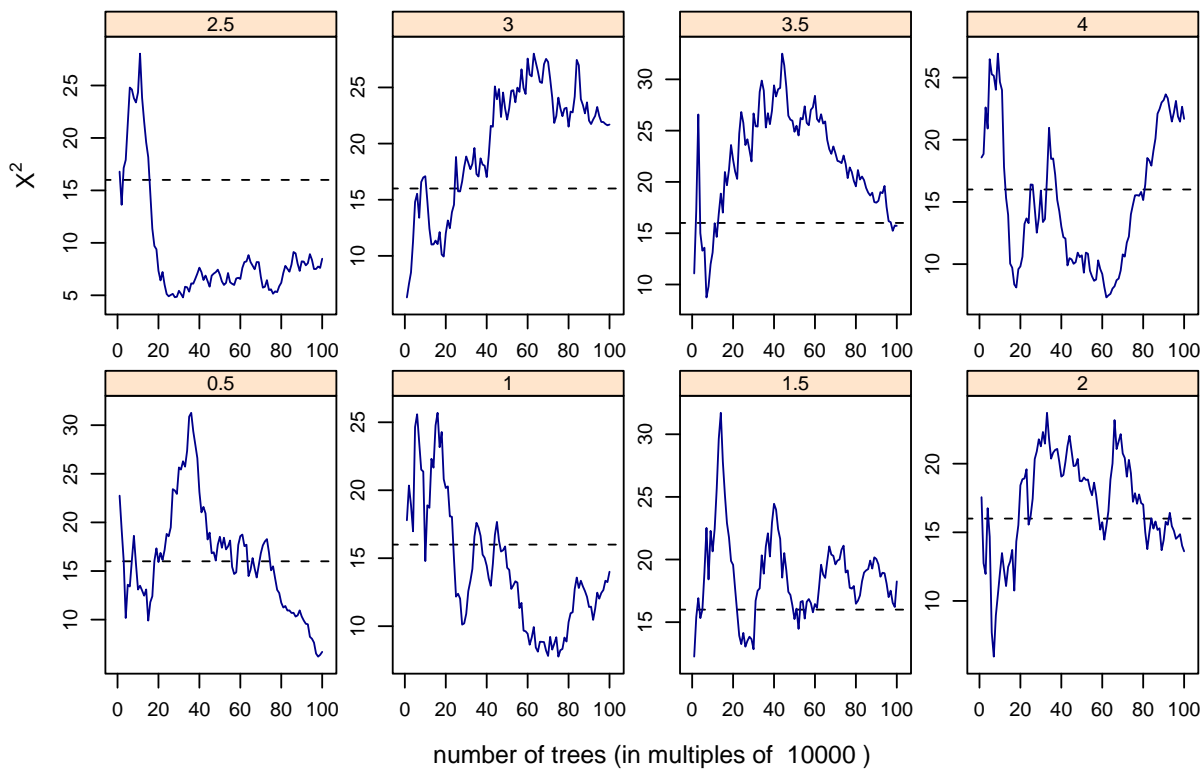


Figure 5: Sample X^2 values for the neutral SNP model (14) as a function of the number of loci (n) used to construct the SFS for SNP data (single segregating site) for samples of $m = 19$ genes. The horizontal line indicates the expectation (degrees of freedom=16, after estimation of θ), with other features as described for Fig. 4.

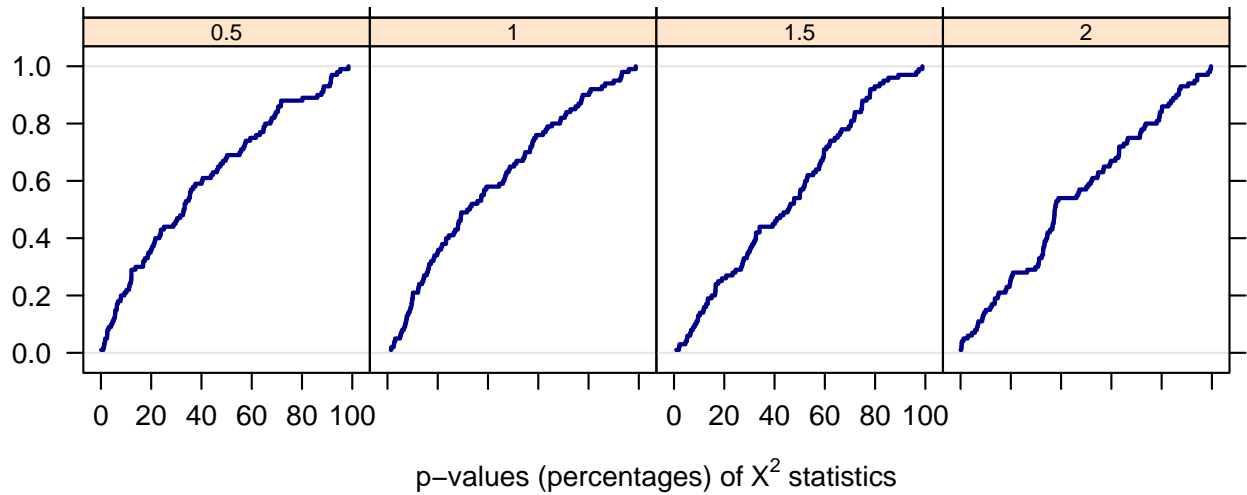


Figure 6: Comparison of the proportion of observed p -values among those obtained from 100 site frequency spectra that lie at or below the nominal significance level (abscissa) for the neutral SNP model (14) applied to SNP data (single segregating site) for the four indicated values of θ .

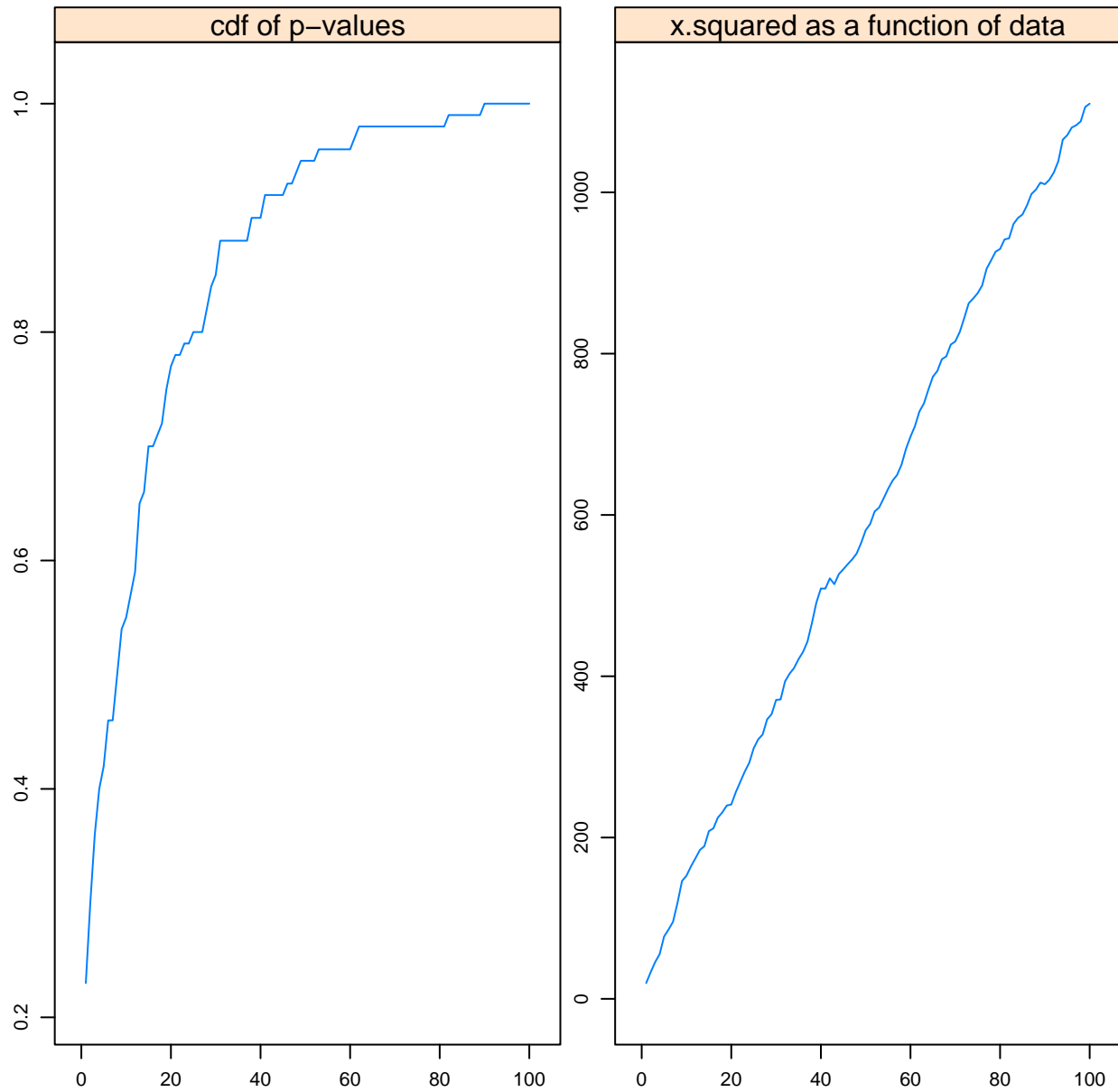


Figure 7: Results of tests of goodness of fit to the scaled multiplicity model (7) of site frequency spectra constructed from single random mutations in sample genealogies generated under $\theta = 0.5$. Left: Comparison between the empirical cumulative distribution of observed p -values (ordinate) to the nominal level of significance (abscissa). Right: X^2 values based on increasing numbers of trees, with the abscissa representing the number of groups of 10,000 trees used to construct the SFS.

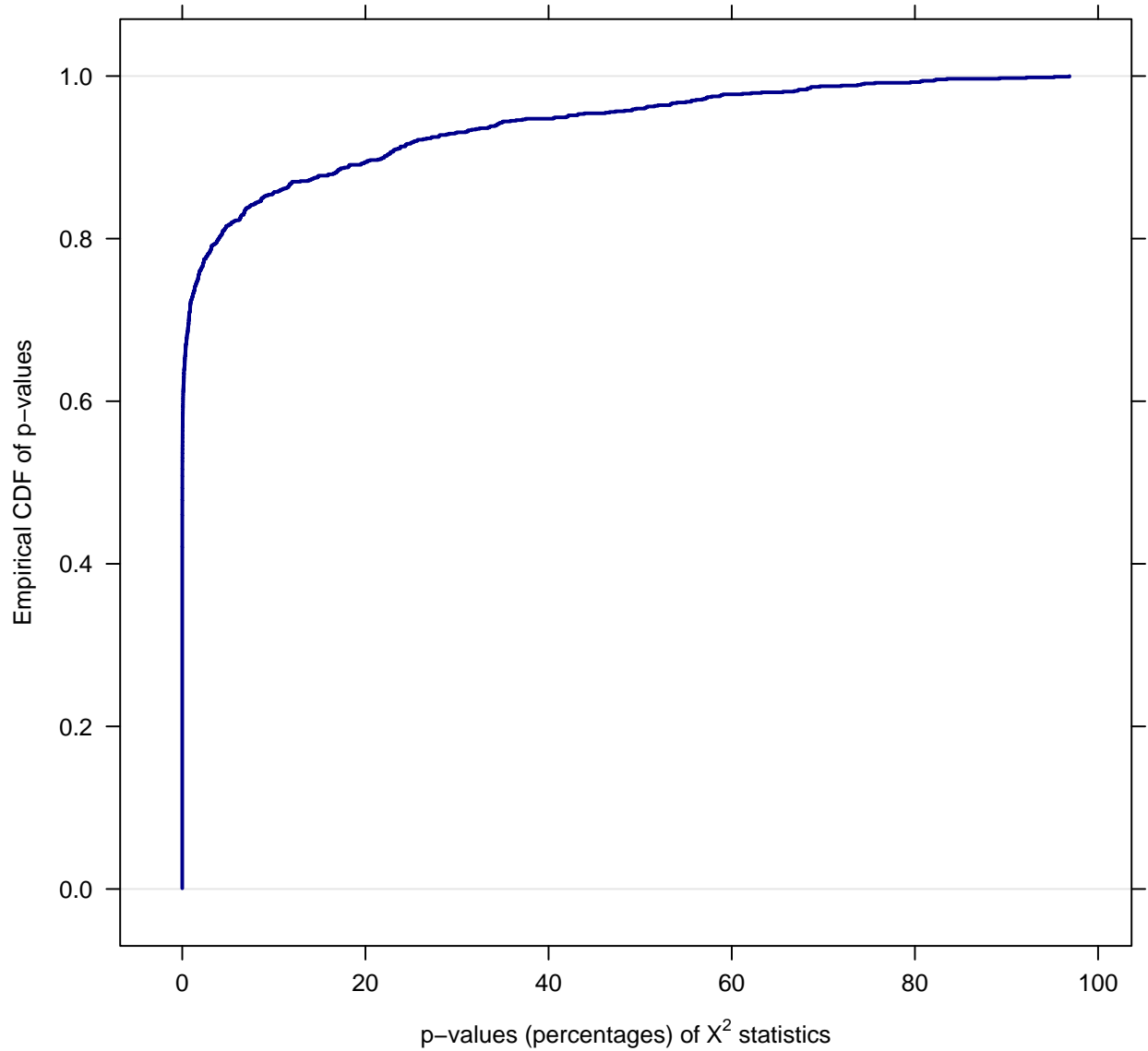


Figure 8: Comparison between the nominal level of significance and the cumulative frequency of p -values obtained from analyses of the goodness of fit of spectra of the multiplicities of all observed mutations to the scaled multiplicity model (7) across all θ values.

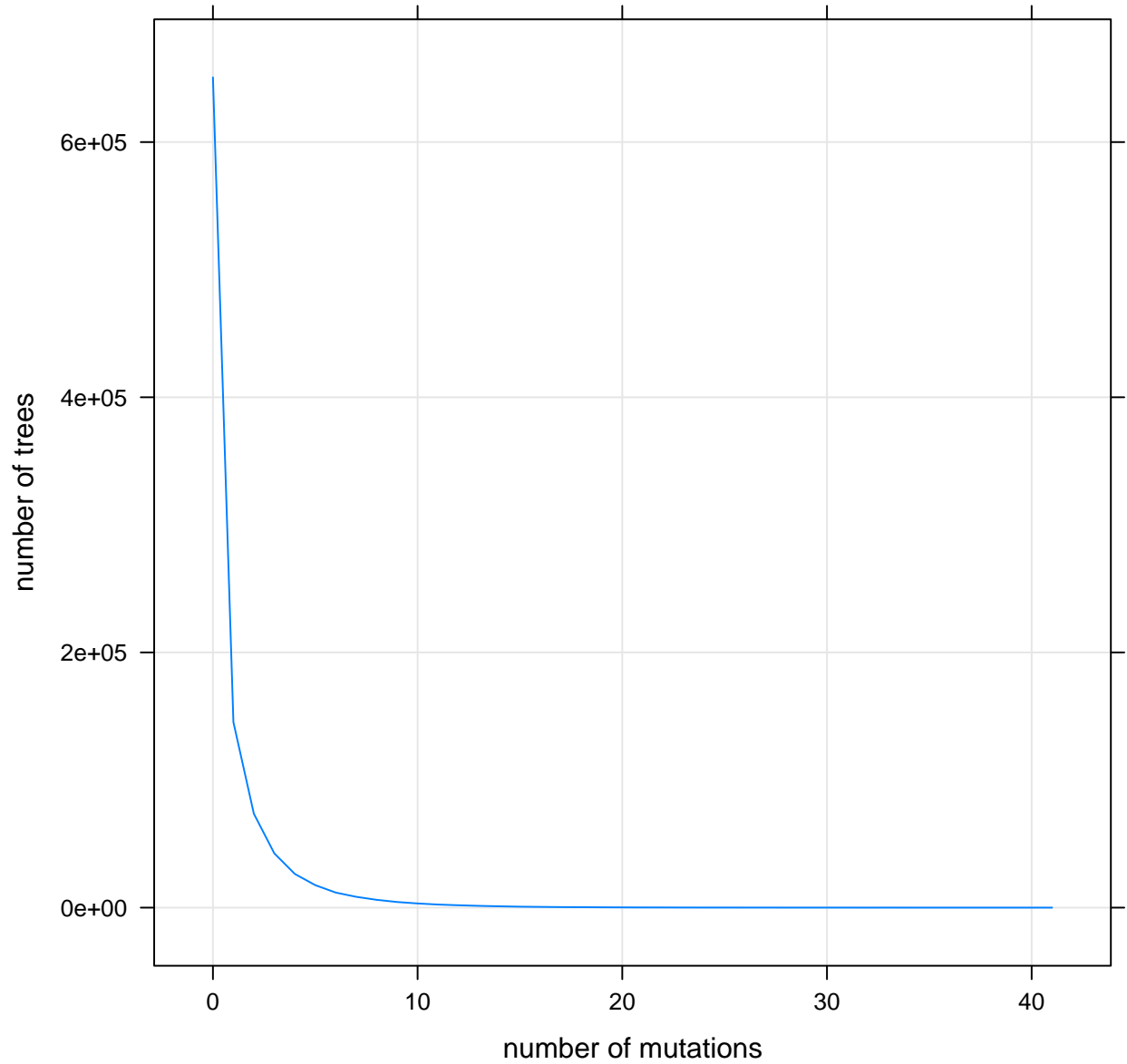


Figure 9: Histogram of the number of trees, among 10^6 samples of size 19 simulated under $\theta = 6.0$, that contain the number of mutations on the abscissa in multiplicity 5.

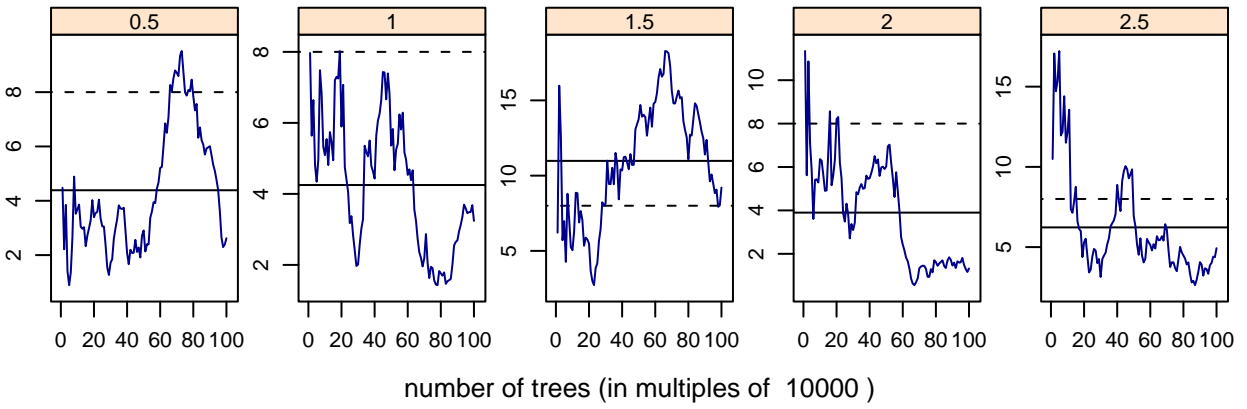


Figure 10: X^2 values for goodness of fit to the scaled multiplicity model (7) of biallelic frequency spectra generated under the θ value indicated above each panel as a function of an increasing number of trees per spectrum. The solid horizontal line represents the average X^2 value and the dashed line the expectation (degrees of freedom=8).