Neutrality Tests for Sequences with Missing Data

Luca Ferretti,^{*,1,2} Emanuele Raineri,^{†,2} and Sebastian Ramos-Onsins^{*}

^{*}Centre for Research in Agricultural Genomics, 08193 Bellaterra, Spain and [†]Centro Nacional de Análisis Genómico, 08028 Barcelona, Spain

ABSTRACT Missing data are common in DNA sequences obtained through high-throughput sequencing. Furthermore, samples of low quality or problems in the experimental protocol often cause a loss of data even with traditional sequencing technologies. Here we propose modified estimators of variability and neutrality tests that can be naturally applied to sequences with missing data, without the need to remove bases or individuals from the analysis. Modified statistics include the Watterson estimator θ_w , Tajima's D , Fay and Wu's H , and HKA. We develop a general framework to take missing data into account in frequency spectrum-based neutrality tests and we derive the exact expression for the variance of these statistics under the neutral model. The neutrality tests proposed here can also be used as summary statistics to describe the information contained in other classes of data like DNA microarrays.

NEUTRALITY tests are among the most widely used tools in population genetics. Many neutrality tests have been developed based on the levels and the patterns extracted from segregating sites, and in particular to be applied to biallelic SNP data. The simplest information that can be extracted from SNP data are the allele frequency spectrum; therefore, many tests focus on the difference between the observed and expected spectrum under the neutral Wright–Fisher model. Widespread tests of this kind include Tajima's D (Tajima 1989), Fu and Li's F and D (Fu and Li 1993), and Fay and Wu's H (Fay and Wu 2000). However, this class of tests is much larger, as recently shown by Achaz (2009) following an idea of Nawa and Tajima (2008), and includes among the others the tests by Fu (1997), Zeng *et al.* (2006), and Achaz (2009). A subclass of optimal neutrality tests against specific alternative scenarios was described by Ferretti *et al.* (2010). Some general results on the variances of these tests were provided by Fu (1995) and Pluzhnikov and Donnelly (1996).

All these statistics assume a complete knowledge of the alleles present in the n sequenced individuals for all the L positions genotyped. However, this is rarely the case: ex-

perimental problems in sample preparation or genotyping often result in missing data; *i.e.*, some individual alleles at some positions are actually unknown.

At present, most packages for population genetics analyses like DNAsp (Librado and Rozas 2009) deal with missing data simply by removing individuals and/or positions affected with incomplete data. This is a good strategy as long as missing data represent a very minor fraction of the alleles, since in this case they do not affect the power of the analysis. However, there could be situations in which a large amount of missing data are unavoidable. For example, in samples taken from natural populations the quality of the samples could be low or the amount of DNA available per individual could be insufficient; therefore, genotyping these samples could miss a significant fraction of the alleles.

There is another important reason to consider sequences with missing data. Many of the sequences that are being produced currently are not obtained through Sanger sequencing, but from next-generation sequencing (NGS) technologies. These technologies sequence a large amount of short reads that are then realigned to reconstruct the original sequence. The coverage of these reads is strongly inhomogeneous along the genome and there is often a large fraction of bases that is not covered by a sufficient number of reads, unless the coverage is very high. Missing data are therefore inherent to these technologies: hence, removing individuals or bases with missing alleles would imply a huge loss of information. Given the growing relevance of NGS technologies for population genetics studies, a different strategy is needed to deal with this circumstance. Several

Copyright © 2012 by the Genetics Society of America
doi: 10.1534/genetics.112.139949

Manuscript received February 22, 2012; accepted for publication May 24, 2012
Supporting information is available online at <http://www.genetics.org/content/suppl/2012/06/01/genetics.112.139949.DC1>.

¹Corresponding author: Centre de Recerca en Agrigenòmica (CRAG), Campus Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain.
E-mail: luca.ferretti@uab.cat

²These authors contributed equally to this work.

estimators of variability can be applied directly to sequenced reads (Lynch 2008; Hellmann *et al.* 2008; Jiang *et al.* 2009; Futschik and Schlötterer 2010; Kang and Marjoram 2011); however, no estimator is available for the sequences obtained after genotype call has been completed for each individual in each position. The difference between these two situations is that, once the genotype has been determined, all the information about the single read bases aligning on a given position and their qualities is (for our purposes) lost.

In this article we present a simple generalization of some estimators and tests that take missing data into account. In particular we consider the Watterson estimator of genetic variability (Watterson 1975), the Tajima estimator of nucleotide diversity (Tajima 1983), neutrality tests based on the frequency spectrum like Tajima's D and Fay and Wu's H , and the HKA test (Hudson *et al.* 1987) for neutral evolution based on the pattern of polymorphism and divergence. The most important result of this article is the general expression for the covariance between the frequency spectrum at two sites $\text{Cov}(\xi_i(x), \xi_j(y))$, which is the basis for the computation of the variances of the estimators and tests presented here.

Note that in sequence data, missing data (usually represented by N 's located in the same position as the missing alleles) are not equivalent to gaps (represented by white spaces). Gaps correspond to insertions or deletions (indels) in some of the sequences. In this article we do not address indels, even if very short biallelic indels (a few bases long) are similar to SNPs as genetic variants and therefore could be analyzed by similar methods. In practice it is difficult to differentiate indels from missing data if the rate of missing data are high, and this is especially true for sequences obtained from NGS data. Here we consider sequences without indels.

Neutrality Tests Including Missing Data

In this article we consider estimators and tests based on the frequency spectrum. The basic population parameter involved in these tests is the nucleotide variability $\theta = 2N_e\mu$, where N_e is the haploid effective population size and μ is the mutation rate per base per generation. We assume a small variability $\theta \ll 1$ and a large window length $L \gg 1$, such that $\theta L \sim O(1)$.

All the tests and estimators belong to a general class of neutrality tests that can be parametrized in terms of weights ω_i , Ω_i (Achaz 2009),

$$\hat{\theta} = \frac{1}{L} \sum_{i=1}^{n-1} i\omega_i \xi_i, \quad \sum_{i=1}^{n-1} \omega_i = 1 \quad (1)$$

$$T = \frac{\hat{\theta} - \hat{\theta}'}{\text{Var}(\hat{\theta} - \hat{\theta}')} = \frac{\sum_{i=1}^{n-1} i\Omega_i \xi_i}{\text{Var}\left(\sum_{i=1}^{n-1} i\Omega_i \xi_i\right)}, \quad \sum_{i=1}^{n-1} \Omega_i = 0, \quad (2)$$

where ξ_i indicates the number of variants with frequency i for the derived allele. The weights ω_i , Ω_i multiply the normalized frequency spectrum $\theta_i = i\xi_i$ (Nawa and Tajima 2008). The estimators are unbiased estimators of θ , while the tests are normalized to have mean 0 and variance 1 under the standard neutral model without recombination.

Our definition of neutrality tests based on the frequency spectrum is the most general parametrization compatible with Equations 1–2 that takes explicitly into account the coverage for each site. We denote by n the total number of individuals sequenced and by n_x the number of individuals for which the allele at position x is known. The estimators and tests are defined as

$$\hat{\theta} = \frac{1}{L} \sum_{x=1}^L \sum_{i=1}^{n_x-1} i\omega_{i,n_x} \xi_i(x), \quad \frac{1}{L} \sum_{x=1}^L \sum_{i=1}^{n_x-1} \omega_{i,n_x} = 1 \quad (3)$$

$$T = \frac{\hat{\theta} - \hat{\theta}'}{\text{Var}(\hat{\theta} - \hat{\theta}')} = \frac{\sum_{x=1}^L \sum_{i=1}^{n_x-1} i\Omega_{i,n_x} \xi_i(x)}{\text{Var}\left(\sum_{x=1}^L \sum_{i=1}^{n_x-1} i\Omega_{i,n_x} \xi_i(x)\right)}, \quad \sum_{x=1}^L \sum_{i=1}^{n_x-1} \Omega_{i,n_x} = 0, \quad (4)$$

where $\xi_i(x)$ is an index variable that is 1 if there is a segregating site with i derived alleles in position x and 0 otherwise. The estimators are unbiased, *i.e.*, $E(\hat{\theta}) = \theta$, while the tests are normalized to $E(T) = 0$, $\text{Var}(T) = 1$ as in the usual framework.

The weights ω_{i,n_x} , Ω_{i,n_x} define the specific estimator or test (Achaz 2009). The most important estimator of variability is the Watterson estimator $\hat{\theta}_W = S/a_n L$, where we denote by S the total number of segregating sites and $a_n = \sum_{i=1}^{n-1} 1/i$. This estimator can be obtained as a maximum composite likelihood estimator (MCLE) (Hellmann *et al.* 2008). Its natural generalization is the MCLE with missing data,

$$\omega_{i,n_x}^W = \frac{1}{i \sum_{x=1}^L a_{n_x}/L} \Rightarrow \hat{\theta}_W = \frac{S}{\sum_{x=1}^L a_{n_x}}, \quad (5)$$

which depends only on S . For most of the other estimators like Tajima's Π , we choose the weights ω_{i,n_x} to be simply the same as the weights ω_i for the estimators (1), where n is substituted by n_x , that is, $\omega_{i,n_x}^\Pi = 2(n_x - i)/n_x(n_x - 1)$. This definition is equivalent to $\hat{\theta}_\Pi = \Pi/L$, where Π is the average pairwise diversity per base, which is naturally defined even with inhomogeneous coverage.

As for neutrality tests, Tajima's D (Tajima 1989) corresponds to $\hat{\theta}_\Pi - \hat{\theta}_W$, that is, to the weights

$$\Omega_{i,n_x}^D = \frac{2(n_x - i)}{n_x(n_x - 1)} - \frac{1}{i \sum_{y=1}^L a_{n_y}/L}, \quad (6)$$

and Fay and Wu's H (Fay and Wu 2000) corresponds to

$$\Omega_{i,n_x}^H = \frac{2(n_x - i)}{n_x(n_x - 1)} - \frac{2i}{n_x(n_x - 1)}. \quad (7)$$

The other tests can be generalized in a similar way. All the tests and estimators reduce to their usual expressions if no data are missing; *i.e.*, $n_x = n$ for all sites.

In this framework it is also possible to implement error corrections for error-prone data: for example, removing singletons (Achaz 2008) is equivalent to the choice of $\omega_{1,n_x} = \omega_{n_x-1,n_x} = 0$ and a rescaling of the other ω_{i,n_x} to match the normalization in Equation 3. For NGS data, information on base or SNP qualities is usually available; hence, a more refined error correction strategy consists in weighting each SNP in Equations 3–4 by the probability that it has been correctly identified. A detailed treatment can be found in Supporting Information, File S1. For NGS data it is also useful to filter out the bases with low coverage, *i.e.*, the ones for which information from most individuals is missing. If we assume that the minimum number of individuals covering reliable positions is $n_{\min}$, this filter can be easily implemented by removing all positions with $n_x < n_{\min}$ from the analysis.

To evaluate the tests (4), we need the variances in the denominators. Our basic result for these variances (leaving out subleading terms in θ and $1/L$) is

$$\text{Var}\left(\sum_{x=1}^L \sum_{i=1}^{n_x-1} i \Omega_{i,n_x} \xi_i(x)\right) = \sum_{x=1}^L \sum_{i=1}^{n_x-1} i \Omega_{i,n_x}^2 \theta + \sum_{\substack{x,y=1 \\ x \neq y}}^L \sum_{i=1}^{n_x-1} \sum_{j=1}^{n_y-1} ij \Omega_{i,n_x} \Omega_{j,n_y} \text{Cov}(\xi_i(x), \xi_j(y)) \quad (8)$$

since $\xi_i(x)$ are index variables with mean $E(\xi_i(x)) = \theta/i$ and $\xi_i(x), \xi_j(x)$ are mutually exclusive for $i \neq j$, so $E(\xi_i(x)\xi_j(x)) = 0$. The covariance $\text{Cov}(\xi_i(x), \xi_j(y))$ for the standard neutral model without recombination is presented in the next section.

The HKA (Hudson, Kreitman, Aguadé) test (Hudson *et al.* 1987) and the formulae for estimators and tests based on the folded spectrum are treated in File S1.

Covariance of the Frequency Spectrum at Different Sites

Since $\xi_i(x), \xi_j(y)$ are index variables, their covariance under the standard neutral model without recombination is

$$\text{Cov}(\xi_i(x), \xi_j(y)) = P_{ij}(n_x, n_y, n_{xy}) - \frac{\theta^2}{ij}, \quad (9)$$

where $P_{ij}(n_x, n_y, n_{xy})$ is the probability of observing SNPs of frequency i and j , n_x and n_y are the numbers of individuals with known alleles at the two sites, and n_{xy} is the number of individuals for which both alleles are known. This probability can be obtained as

$$P_{ij}(n_x, n_y, n_{xy}) = \sum_{k,l=1}^{n_x+n_y-n_{xy}-1} \left[C_{ij,kl}^S(n_x, n_y, n_{xy}) P_{kl}^S(n_x+n_y-n_{xy}) + C_{ij,kl}^E(n_x, n_y, n_{xy}) P_{kl}^E(n_x+n_y-n_{xy}) \right], \quad (10)$$

where $P_{kl(n)}^S$ and $P_{kl(n)}^E$ are the probabilities of shared (S) or exclusive (E) pairs of mutations of frequency k and l in n complete sequences. (We define a pair of mutations as shared if there are individuals with derived alleles in both loci and as exclusive if no individual sequence contains both the derived alleles.) The sum $P_{kl}^S + P_{kl}^E$ gives the probability for complete sequences $P_{kl} = \theta^2(1/k! + \sigma_{kl})$, where the matrix σ_{kl} is defined in Fu (1995, Equations 2–3).

The coefficients $C_{ij,kl(n_x, n_y, n_{xy})}^{S,E}$ represent the probabilities that, given a pair of shared or exclusive mutations with frequencies k and l in $n_x + n_y - n_{xy}$ complete sequences, i and j derived alleles are found among the n_x, n_y alleles in x and y , respectively, assuming that the n_x, n_y individuals (with n_{xy} in common between the two sets) are randomly extracted from the complete set of $n_x + n_y - n_{xy}$ individuals. The combinatorial formulae for these probabilities are

$$C_{ij,kl(n_x, n_y, n_{xy})}^S = \frac{\binom{n_x - n_{xy}}{l-j} \binom{n_y - n_{xy}}{k-i}}{\binom{n_x + n_y - n_{xy}}{l, k-l, n_x + n_y - n_{xy} - k}} \sum_{k_x = \max(i - n_{xy}, l-j)}^{\min(i, n_x - n_{xy}, k-j)} \binom{n_{xy}}{i - k_x} \binom{n_x - n_{xy} + j - l}{k_x + j - l} \binom{k - k_x}{j} \quad (11)$$

if $k \geq l$; otherwise use the identity $C_{ij,kl(n_x, n_y, n_{xy})}^S = C_{ji,lk(n_y, n_x, n_{xy})}^S$,

$$C_{ij,kl(n_x, n_y, n_{xy})}^E = \frac{\binom{n_x - n_{xy}}{l-j} \binom{n_y - n_{xy}}{k-i}}{\binom{n_x + n_y - n_{xy}}{k, l, n_x + n_y - n_{xy} - k - l}} \sum_{k_x = \max(0, i - n_{xy}, k - n_y + j)}^{\min(i, n_x - n_{xy} + j - l)} \binom{n_{xy}}{i - k_x} \binom{n_x - n_{xy} + j - l}{k_x} \binom{n_y - k + k_x}{j}, \quad (12)$$

where $\binom{a}{b, c, a-b-c}$ is the multinomial coefficient $a!/b!c!(a-b-c)!$. We define $C_{ij,kl(n_x, n_y, n_{xy})}^S$ or $C_{ij,kl(n_x, n_y, n_{xy})}^E$ to be zero if there are negative arguments in the binomial or multinomial coefficients in the above Equations 11 or 12.

The formulae for the probabilities $P_{kl(n)}^{S,E}$ can be obtained by breaking the derivation of $E(\xi_k \xi_l)$ by Fu (1995) into the contributions from shared mutations (Fu 1995, Equations 24 and 28) and exclusive mutations (Equations 25, 29, and 30):

$$P_{kl(n)}^S = \theta^2 \delta_{kl} \beta_n(k) + \theta^2 (1 - \delta_{kl}) \left[\frac{\beta_n(\min(k, l)) - \beta_n(\min(k, l) + 1)}{2} \right] \quad (13)$$

$$P_{kl(n)}^E = \begin{cases} \theta^2 \left[\frac{1}{kl} - \frac{\beta_n(k) - \beta_n(k+1) + \beta_n(l) - \beta_n(l+1)}{2} \right] & \text{for } k+l < n \\ \theta^2 \left[\frac{a_n - a_k}{n-k} + \frac{a_n - a_l}{n-l} + \frac{\beta_n(k) + \beta_n(l)}{2} \right] & \text{for } k+l = n \\ 0 & \text{for } k+l > n \end{cases} \quad (14)$$

Figure 1 Variance of the Watterson estimator θ_W on a window of $L = 100$ bases for $\theta = 0.1$. Computed by drawing out randomly $N_s = 100$ triples (n_x, n_y, n_{xy}) in two different ways. First, we fix the sample size $n = 20$ and remove alleles randomly according to the probability p_m of missing an allele (solid blue line). In this case the number of individuals is fixed but the actual depth may vary along the sequence. Second, the number of individuals sequenced in each position is adjusted to keep the average depth constant at $n(1 - p_m) \simeq 20$ (dashed green line) and then we remove alleles with probability p_m . In this case the sample size is $n \simeq 20/(1 - p_m)$.

with

$$a_n = \sum_{i=1}^{n-1} \frac{1}{i}, \quad \beta_n(i) = \frac{2n}{(n-i+1)(n-i)} (a_{n+1} - a_i) - \frac{2}{n-i}. \quad (15)$$

Some special cases of these formulae are treated in [File S1](#), [Figure S1](#), and [Figure S2](#).

Finally, note that the computation of the variances requires an estimate of $\hat{\theta}$ and $\hat{\theta}^2$. These estimates are usually obtained by the method of moments (MM). In our approach, $\hat{\theta}$ is given by the Watterson estimator (5), while the MM estimate of $\hat{\theta}^2$ is given by

$$\hat{\theta}^2 = \frac{S^2 - S}{\left(\sum_{x=1}^L a_{n_x}\right)^2 + \sum_{x,y=1}^L \sum_{i=1}^{n_x-1} \sum_{j=1}^{n_y-1} \text{Cov}(\xi_i(x), \xi_j(y)) / \theta^2}. \quad (16)$$

Discussion

In this article we presented a general framework for estimators of variability and neutrality tests based on the frequency spectrum that take into account missing data in a natural way. This is particularly interesting in the light of sequences obtained from NGS data, since for these technologies a relevant fraction of bases is often not sequenced or sequenced at very low read depth.

Figure 2 Variance of Tajima's D (bottom line) and Fay and Wu's H (top line) on a window of $L = 100$ bases for $\theta = 0.1$. Computed as in Figure 1 for fixed average depth $n(1 - p_m) \simeq 20$. (Note that Fay and Wu's H variance is divided by 4 to appear in scale with Tajima's D variance.)

The approach discussed here is based on results that are conditional on the distribution of the missing data, as summarized by the distribution of all the triples (n_x, n_y, n_{xy}) . An effective way of implementing numerically the above variances is to sample N_s random values of (n_x, n_y, n_{xy}) from the empirical distribution and compute the covariances using only these values and then rescale the second term in Equation 8 by a factor L^2/N_s .

The modifications presented in this article can be applied to all estimators and tests included in the framework of Achaz (2009) and represent, therefore, a complete tool with which to deal with missing data. However, it would be interesting to know the impact of the missing data on the performance of the estimators and tests. If we fix the sample size, an increase in the amount of missing data leads to an increase in the variance of the estimators (Figure 1), as is to be expected given that this is equivalent to loss of information. On the other hand, if the loss of information associated with missing data is compensated by sequencing more individuals, the performances of the estimators actually *increase* (i.e., their variances decrease) with respect to complete sequences with the same coverage (Figure 1). A similar effect can be observed for neutrality tests (Figure 2). The explanation for this counterintuitive behavior lies in the fact that in the case of complete sequences, all individuals share the same genealogical tree at all positions, i.e., the same evolutionary history, while in this case different positions are covered by different sets of individuals with partly independent histories in the same population; therefore, the number of available histories is actually larger and the variance is reduced similarly to what happens with recombination. Our results imply that with the same amount of information per base, missing data could improve the power of neutrality tests.

Acknowledgments

Work funded by grant CGL2009-09346 to S.R.O., grant AG2010-14822 to Miguel Pérez-Enciso, and Consolider grant CSD2007-00036 “Centre for Research in Agrigenomics” (Ministerio de Ciencia e Innovación, Spain). S.R.O. is recipient of a Ramón y Cajal position (Ministerio de Ciencia e Innovación, Spain). L.F. acknowledges support from Consejo Superior de Investigaciones Científicas (Spain) under the JAE-doc program.

Literature Cited

- Achaz, G., 2008 Testing for neutrality in samples with sequencing errors. *Genetics* 179: 1409.
- Achaz, G., 2009 Frequency spectrum neutrality tests: one for all and all for one. *Genetics* 183: 249.
- Fay, J., and C.-I. Wu, 2000 Hitchhiking under positive Darwinian selection. *Genetics* 155: 1405.
- Ferretti, L., M. Perez-Enciso, and S. Ramos-Onsins, 2010 Optimal neutrality tests based on the frequency spectrum. *Genetics* 186: 353.
- Fu, Y., and W.-H. Li, 1993 Statistical tests of neutrality of mutations. *Genetics* 133: 693.
- Fu, Y.-X., 1995 Statistical properties of segregating sites. *Theor. Popul. Biol.* 48: 172–197.
- Fu, Y.-X., 1997 Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* 147: 915.
- Futschik, A., and C. Schlötterer, 2010 The next generation of molecular markers from massively parallel sequencing of pooled dna samples. *Genetics* 186: 207.
- Hellmann, I., Y. Mang, Z. Gu, P. Li, M. Francisco *et al.*, 2008 Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals. *Genome Res.* 18: 1020–1029.
- Hudson, R., M. Kreitman, and M. Aguadé, 1987 A test of neutral molecular evolution based on nucleotide data. *Genetics* 116: 153.
- Jiang, R., S. Tavaré, and P. Marjoram, 2009 Population genetic inference from resequencing data. *Genetics* 181: 187.
- Kang, C., and P. Marjoram, 2011 Inference of population mutation rate and detection of segregating sites from next-generation sequence data. *Genetics* 189: 595–605.
- Librado, P., and J. Rozas, 2009 DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25: 1451.
- Lynch, M., 2008 Estimation of nucleotide diversity, disequilibrium coefficients, and mutation rates from high-coverage genome-sequencing projects. *Mol. Biol. Evol.* 25: 2409–2419.
- Nawa, N., and F. Tajima, 2008 Simple method for analyzing the pattern of dna polymorphism and its application to snp data of human. *Genes Genet. Syst.* 83: 353–360.
- Pluzhnikov, A., and P. Donnelly, 1996 Optimal sequencing strategies for surveying molecular genetic diversity. *Genetics* 144: 1247.
- Tajima, F., 1983 Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105: 437.
- Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123: 585.
- Watterson, G., 1975 On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* 7: 256.
- Zeng, K., Y.-X. Fu, S. Shi, and C.-I. Wu, 2006 Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics* 174: 1431–1439.

Communicating editor: N. A. Rosenberg

GENETICS

Supporting Information

<http://www.genetics.org/content/suppl/2012/06/01/genetics.112.139949.DC1>

Neutrality Tests for Sequences with Missing Data

Luca Ferretti, Emanuele Raineri, and Sebastian Ramos-Onsins

FILE S1
SUPPORTING INFORMATION

General framework for tests with missing data:

The general framework proposed by ACHAZ (2009) for estimators $\hat{\theta}$ and neutrality tests T based on the frequency spectrum ξ_i is based on these assumptions:

1. the estimator/test statistics is a linear function of the frequency spectrum ξ_i and a general function of the variability θ ;
2. the expected value of the statistics under the standard neutral model (SNM) is $E(\hat{\theta}|\theta) = \theta$ for the estimators and $E(T|\theta) = 0$ for the tests;
3. the tests are normalized such that their variance under the SNM without recombination is $\text{Var}(T|\theta) = 1$.

Finally, the actual values of θ , θ^2 in the statistics are estimated from the Watterson estimator and the MM estimator for θ^2 . It is easy to check that these conditions imply the equations (1), (2) for general estimators and tests.

In the framework of sequences with missing data, the estimators and tests should be actually based on the site frequency spectrum $\xi_i(x)$. We propose a set of assumptions which is a slight generalization of the one above:

1. the estimator/test statistics is a linear function of the frequency spectrum $\xi_i(x)$ and a general function of the variability θ ;
2. the expected value of the statistics under the standard neutral model (SNM) is $E(\hat{\theta}|\theta) = \theta$ for the estimators and $E(T|\theta) = 0$ for the tests;
3. the tests are normalized such that their variance under the SNM without recombination is $\text{Var}(T|\theta) = 1$;
4. the relative weight of the site frequency spectrum $\xi_i(x)$ for a given position x depends on local information only.

Assumption 4 is not compulsory (in fact, more general tests can be obtained), but it helps to reduce considerably the complexity of the class of tests without a sensible reduction of their power. The assumptions 1-4 imply immediately the general form of equations (3), (4) for estimators and tests.

Accounting for base/SNP calling errors in sequences from NGS:

Sequences called from NGS data could contain a relatively high number of incorrectly called bases. As a result of these errors, false SNPs could appear and affect the statistics. (In principle, these base errors could change SNP frequencies or avoid detection of true SNPs; however, the fraction of SNPs in a sequence is generally low enough that these effects are rare and not relevant.)

Depending on the way the sequences have been obtained, two kind of quality data could be available: base qualities (often available when all sequences have been called separately) and SNP qualities (available as an output of SNP callers). These qualities are actually given in terms of error probabilities; for example, if the qualities are Phred scaled, the error probability is $10^{-\text{quality}/10}$, so quality 10 means error probability 0.1, quality 20 means error probability 0.01, quality 30 means error probability 0.001, etc.

We assume that all the sites are biallelic (this can be done by SNP calling, or by taking only the two most abundant alleles, or the two alleles with lowest product of base error probabilities). For each position x where multiples alleles are present in the data, we want to obtain the probability of true SNP $p_{SNP}(x)$. The way to do it depends on the available data:

- SNP qualities: $p_{SNP}(x)$ is simply 1 minus the SNP error probability;
- base qualities: for each allele in position x compute the product of the base error probabilities, then take $p_{SNP}(x)$ to be 1 minus the higher of the two products.

Once $p_{SNP}(x)$ has been obtained, the estimators and tests can be corrected for sequencing errors as follows:

$$\hat{\theta} = \frac{1}{L} \sum_{x=1}^L \sum_{i=1}^{n_x-1} i \omega_{i,n_x} p_{SNP}(x) \xi_i(x) \quad (\text{SI-1})$$

$$T = \frac{\sum_{x=1}^L \sum_{i=1}^{n_x-1} i \Omega_{i,n_x} p_{SNP}(x) \xi_i(x)}{\text{Var} \left(\sum_{x=1}^L \sum_{i=1}^{n_x-1} i \Omega_{i,n_x} \xi_i(x) \right)} \quad (\text{SI-2})$$

where in the denominator of T we neglect terms of order $p_{SNP}(1 - p_{SNP})$ since we assume $p_{SNP} \simeq 1$.

HKA test with missing data:

We propose also a modified version of the HKA test (HUDSON *et al.* 1987) that deals with missing data. The HKA test is a widely used multi-locus test for neutral sequence evolution, based on the statistics

$$X^2 = \sum_l \frac{(S_l - E(S_l))^2}{\text{Var}(S_l)} + \sum_l \frac{(S'_l - E(S'_l))^2}{\text{Var}(S'_l)} + \sum_l \frac{(D_l - E(D_l))^2}{\text{Var}(D_l)} \quad (\text{SI-3})$$

which has an approximate χ^2 distribution in the neutral case. In the above equation, D_l is the divergence between the two species for the l th locus (i.e. the number of fixed differences) and S_l , S'_l denote the numbers of segregating sites of the two species. This statistics can be applied to incomplete sequences by substituting the correct values for $E(S)$, $E(D)$, $\text{Var}(S)$ and $\text{Var}(D)$. For sequences with missing data, $E(S) = \theta \sum_{x=1}^L a_{n_x}$ while $\text{Var}(S) = \text{Var}(\hat{\theta}_W) \left(\sum_{x=1}^L a_{n_x} \right)^2$. The expected value and variance of the divergence are given by the standard formulae, taking into account that sites with no coverage in one or both populations must be discarded and do not count in $E(D)$ or $\text{Var}(D)$.

The variance of the Watterson estimator $\text{Var}(\hat{\theta}_W)$, as well as the variance of all the estimators (3), can be obtained from equation (8) by substituting Ω_{i,n_x} with $\omega_{i,n_x}/L$. In particular, $\text{Var}(\hat{\theta}_W)$ is given by

$$\text{Var}(\hat{\theta}_W) = \frac{\theta}{\sum_{x=1}^L a_{n_x}} + \frac{1}{\left(\sum_{x=1}^L a_{n_x}\right)^2} \sum_{x,y=1}^L \sum_{i=1}^{n_x-1} \sum_{j=1}^{n_y-1} \text{Cov}(\xi_i(x), \xi_j(y)) \quad (\text{SI-4})$$

Covariance formulae - special cases:

There are two special cases of the formulae for the covariances (10-14) given in the Main Text. The first case occurs when the allele in y is known for all individuals with known allele in x , i.e. $n_y = n_{xy}$. In this case $P_{ij(n_x, n_y, n_{xy})}$ reduces to a simpler expression in terms of an hypergeometric distribution:

$$P_{ij(n_x, n_{xy}, n_{xy})} = \sum_{l=j}^{n_x - n_{xy} + j} P_{il(n_x)} \frac{\binom{n_x - n_{xy}}{l-j} \binom{n_{xy}}{j}}{\binom{n_x}{l}} \quad (\text{SI-5})$$

where $P_{il(n_x)} = \theta^2(1/il + \sigma_{il})$ is the probability obtained by FU (1995) for n_x complete sequences. The second special case corresponds to $n_{xy} = 0$, i.e. there are no individuals for which both alleles at x and y are known. In this case both C^S and C^E reduce to generalized hypergeometric distributions:

$$C_{ij,kl(n_x, n_y, 0)}^S = \frac{\binom{n_x}{l-j, i+j-l, n_x-i} \binom{n_y}{j, k-i-j, n_y-k+i}}{\binom{n_x+n_y}{l, k-l, n_x+n_y-k}} \quad , \quad k \geq l \quad (\text{SI-6})$$

$$C_{ij,kl(n_x, n_y, 0)}^E = \frac{\binom{n_x}{i, l-j, n_x-l+j-i} \binom{n_y}{k-i, j, n_y-k+i-j}}{\binom{n_x+n_y}{k, l, n_x+n_y-k-l}} \quad (\text{SI-7})$$

Formulae for folded spectrum:

The results in Main Text have been obtained for the case where the ancestral allele is known, for example from an outgroup sequence, and therefore the frequency spectrum is unfolded. In many situations an outgroup is not available and it is not possible to discriminate between derived and ancestral alleles; in this case, the estimators and tests should be based on the folded frequency spectrum $\eta_i(x) = (\xi_i(x) + \xi_{n_x-i}(x))/(1 + \delta_{n_x, 2i})$. As discussed by ACHAZ (2009), estimators and tests for folded data should satisfy additional conditions, which in our framework read $i\omega_{i,n_x} = (n_x - i)\omega_{n_x-i, n_x}$ and $i\Omega_{i,n_x} = (n_x - i)\Omega_{n_x-i, n_x}$. We explain here how to obtain the estimators and tests based on the folded spectrum.

In our framework, all estimators and tests depend on the folded site frequency spectrum $\eta_i(x) = (\xi_i(x) + \xi_{n_x-i}(x))/(1 + \delta_{\lfloor n_x/2 \rfloor, i})$. The general form for estimators and tests is

$$\hat{\theta} = \frac{1}{L} \sum_{x=1}^L \sum_{i=1}^{\lfloor n_x/2 \rfloor} \frac{i(n_x - i)}{n_x(1 + \delta_{\lfloor n_x/2 \rfloor, i})} \omega_{i,n_x} \eta_i(x) \quad , \quad \frac{1}{L} \sum_{x=1}^L \sum_{i=1}^{\lfloor n_x/2 \rfloor} \omega_{i,n_x} = 1 \quad (\text{SI-8})$$

$$T = \frac{\hat{\theta} - \hat{\theta}'}{\text{Var}(\hat{\theta} - \hat{\theta}')} = \frac{\sum_{x=1}^L \sum_{i=1}^{\lfloor n_x/2 \rfloor} \frac{i(n_x-i)(1+\delta_{\lfloor n_x/2 \rfloor, i})}{n_x} \Omega_{i, n_x} \eta_i(x)}{\text{Var}\left(\sum_{x=1}^L \sum_{i=1}^{\lfloor n_x/2 \rfloor} \frac{i(n_x-i)(1+\delta_{\lfloor n_x/2 \rfloor, i})}{n_x} \Omega_{i, n_x} \eta_i(x)\right)}, \quad \sum_{x=1}^L \sum_{i=1}^{\lfloor n_x/2 \rfloor} \Omega_{i, n_x} = 0 \quad (\text{SI-9})$$

The variances are

$$\begin{aligned} \text{Var}\left(\sum_{x=1}^L \sum_{i=1}^{\lfloor n_x/2 \rfloor} \frac{i(n_x-i)(1+\delta_{\lfloor n_x/2 \rfloor, i})}{n_x} \Omega_{i, n_x} \xi_i(x)\right) &= \sum_{x=1}^L \sum_{i=1}^{\lfloor n_x/2 \rfloor} \frac{i(n_x-i)(1+\delta_{\lfloor n_x/2 \rfloor, i})}{n_x} \Omega_{i, n_x}^2 \theta + \\ &+ \sum_{\substack{x, y=1 \\ x \neq y}}^L \sum_{i=1}^{\lfloor n_x/2 \rfloor} \sum_{j=1}^{\lfloor n_y/2 \rfloor - 1} \frac{i(n_x-i)(1+\delta_{\lfloor n_x/2 \rfloor, i})}{n_x} \frac{j(n_y-j)(1+\delta_{\lfloor n_y/2 \rfloor, j})}{n_y} \Omega_{i, n_x} \Omega_{j, n_y} \text{Cov}(\eta_i(x), \eta_j(y)) \end{aligned} \quad (\text{SI-10})$$

in terms of the covariance $\text{Cov}(\eta_i(x), \eta_j(y))$ between different sites, which can be obtained as

$$\text{Cov}(\eta_i(x), \eta_j(y)) = \frac{\text{Cov}(\xi_i(x), \xi_j(y)) + \text{Cov}(\xi_{n_x-i}(x), \xi_j(y)) + \text{Cov}(\xi_i(x), \xi_{n_y-j}(y)) + \text{Cov}(\xi_{n_x-i}(x), \xi_{n_y-j}(y))}{(1+\delta_{\lfloor n_x/2 \rfloor, i})(1+\delta_{\lfloor n_y/2 \rfloor, j})} \quad (\text{SI-11})$$

Numerical results and discussion:

In the main text we provide evidence of an increase in performance when the read depth is fixed and more individuals are sequenced, both for Watterson estimator (Figure 1 in the main text) and for neutrality tests (Figure S1). This decrease in variance (i.e., the increase in performance) is apparent again if we compare these variances with fixed sample size and $p_m > 0$ with the variances at the same average depth but without missing data, as in Figure S2. The effect is stronger at lower depth. Interestingly, missing data could therefore result in a loss of power for haplotype tests, but they increase the performance of tests and estimators based on the frequency spectrum as long as they are compensated by an higher number of sequences.

LITERATURE CITED

- ACHAZ, G., 2009 Frequency Spectrum Neutrality Tests: One for All and All for One. *Genetics* **183**: 249.
- FU, Y.-X., 1995 Statistical properties of segregating sites. *Theoretical Population Biology* **48**: 172–197.
- HUDSON, R., M. KREITMAN, and M. AGUADÉ, 1987 A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**: 153.

Figure S1: Variance of Tajima's D (lower lines) and Fay and Wu's H (upper lines) on a window of $L = 100$ bases for $\theta = 0.1$. Computed as in Figure 1 of the main text for fixed sample size $n = 20$ (solid lines) and fixed average depth $n(1 - p_m) \simeq 20$ (dashed lines). The decrease in variance for fixed sample size is due to the reduced effective sample size $20(1 - p_m)$. (Note that Fay and Wu's H variance is divided by 4 to appear in scale with Tajima's D variance.)

Figure S2: Ratio of the variances of Tajima's D (blue line) and Fay and Wu's H (green line) between two cases with the same average depth $20(1 - p_m)$: first, with missing data ($p_m > 0$) and fixed sample size $n = 20$, second, with sample size $\simeq 20(1 - p_m)$ but without missing data ($p_m = 0$). Computed as in Figure S1 on a window of $L = 100$ bases for $\theta = 0.1$.