The relative Site Frequency Spectrum (*rSFS*) of subsamples versus a population

Sebastian E. Ramos-Onsins, Yuliaxis Ramayo-Caldas

July 10, 2025

1 Methods

1.1 Concept

The Site Frequency Spectrum (SFS) is the distribution of frequencies of the mutations that are contained in a sample or a population. This information is fundamental for studying the variability of the population and for inferring the evolutionary events occurred in the population. Under the Standard Neutral Model, the expected distribution of mutations for each frequency is $E(\xi_i) = \theta/i$ (Fu, 1995), where i is the frequency of the mutation (from 1 to n-1, being n the number of samples), ξ_i are the components of a vector that contains the unfolded frequency spectrum with the number of variants at frequency i/n and θ is the population mutation rate (that is, for a diploid species with N_e number of effective individuals and with a mutation rate μ is $\theta = 4N_e\mu$). Then, it is expected a large number of variants at low frequency and fewer variants at higher frequencies.

Each haplotype, individual or group of individuals contain information about the frequency of each of the derived mutations in relation to the total population. For example, for a single haplotype is possible to count the number of derived variants present at this haplotype that are at different frequencies in the entire population. Splitting the counts by the frequency at what this mutations are in the entire populations conforms a sort of Site Frequency Spectrum (we call relative SFS, rSFS) that is specific for this haplotype. This distribution can also be obtained for a diploid individual or a group of individuals, simply by counting the presence of mutations at this group and their frequency at the entire population. Therefore, we define $\xi_{i,j}$ as the components of a vector (from 1 to n-1) that contains the number of variants present at a given subsample of size j but at frequency i in the total sample size (see Figure 1 for an example). Differences between these subset samples can give information about specific evolutionary events occurring at this subset in comparison with other subsets, for example in populations that are experimenting a

gradual spread on large locations, with no clear structured subsets of populations.

1.2 Estimation of the levels of variability of rSFS for an unfolded frequency spectrum

The levels and the patterns of nucleotide diversity can be easily inferred from the rSFS, in relation to the total population. We are interested in estimating the levels of diversity of a subset of samples in relation to the total population and compare both estimates in order to detect differences in different locations (see Figure 1). Assuming neutrality, the estimation of the level of variability at the total population considering mutations at frequency i is $\hat{\theta}(i) = i\xi_i$.

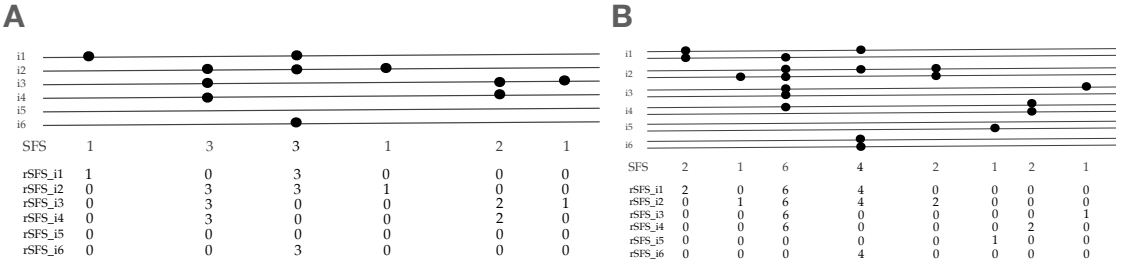

Figure 1: Example of the calculation of the rSFS with subsamples of size $j = 1$ (A) and for $j = 2$ (B). On the top section, the lines represent a sequence from an alignment of six individuals (haploid or diploid), the black dots indicate a different nucleotide from the line (a nucleotide polymorphism). On the bottom section, it is summarized the frequency of the polymorphisms at each position for the total sample (SFS) and for each subset.

We want to compare the variability estimated at the total population with any of the subset samples. In case we have a panmictic population, any subset of samples should estimate the same value of variability as the total. Then, for a sampling subset with j haplotypes the estimate of variability is:

$$\hat{\theta}(i)_j = i\xi_{i,j}\psi_{i,j}^{-1}. \tag{1}$$

Here, $\psi_{i,j}$ indicates that is necessary to include an adjustment to achieve $\theta(i) = \theta(i)_j$ under a stationary neutral model. The number of variants should be compensated given the smaller size of the subsample and the frequency of the variants in the subsample. The number of variants at a specific frequency i in the subsample with j haplotypes is given by a hypergeometrical distribution in relation to the total sample. The probability to have at least one haplotype with the derived variant in the subsample depends on the total sample size (n), the frequency of the variant at the total sample (i) and the size of the subsample (j). That is:

$$\psi_{i,j} = P(k > 0; j, i, n) = (1 - P(k = 0; j, i, n)) = 1 - \frac{\binom{n-i}{j}}{\binom{n}{j}}. \quad (2)$$

INCLUDE Fu 1995 equations for mean,var,cov.

The expected number of variants at each frequency for a subsample of size j, in relation to the total, is weighted considering this probability:

$$E(\xi_{i,j}) = E(\xi_i)\psi_{i,j} = \frac{\theta}{i}(1 - \frac{\binom{n-i}{j}}{\binom{n}{j}}), \quad (3)$$

For estimating the variance and the covariance of $\xi_{i,j}$, we have to consider that each subsample contains haplotypes that are different from other subsamples. That is, the variance of $\xi_{i,j}$ is also dependent on the combination of coalescent branches that finally give the variant frequency. For example, a haplotype that contains only singletons results from a combination where all (except the last) coalescent processes occurred in the rest of branches.

$$\begin{aligned} &TODO : Var(\xi_{i,j}) = Var(\psi_{i,j}\xi_i) = \psi_{i,j}^2 Var(\xi_i) = \psi_{i,j}^2(\frac{\theta}{i} + \sigma_{ii}\theta^2), \\ &TODO : Cov(\xi_{i,j}, \xi_{k,j}) = \end{aligned} \quad (4)$$

where σ_{ii} and σ_{ik} are defined in Fu (1995).

1.3 Estimation of the levels of variability of rSFS for a folded frequency spectrum

In case having no information about the allele that is derived or ancestral, we can analyze the folded spectrum. In that case, the frequencies i and $n - i$ are confounded and can not be treated separately. Define $\eta_i = \frac{(\xi_i + \xi_{n-i})}{(1+\delta_{i,n-1})}$, where $\delta_{i,n-i}$ is the kronecker delta (δ is equal to 0 if $i \neq n - i$, otherwise is 1). Kronecker delta is included to avoid count twice the mutations at frequency $i = n - i$. The expected number of mutations at frequency i for the total sample is:

INCLUDE Fu 1995 equations for mean,var,cov.

$$E(\eta_i) = \frac{E(\xi_i) + E(\xi_{n-i})}{1 + \delta_{i,n-i}} = \frac{\frac{\theta}{i} + \frac{\theta}{n-i}}{1 + \delta_{i,n-i}} = \theta \frac{\frac{n}{i(n-i)}}{1 + \delta_{i,n-i}} = \theta\phi_i. \quad (5)$$

Considering a subset of samples of size j from the total sample, the number variants at minor allele frequency i in the subsample is:

$$E(\eta_{i,j}) = E(\eta_i)\psi_{i,j} = \theta\phi_i\psi_{i,j}, \tag{6}$$

$$TODO : Var(\eta_{i,j}) = \quad TODO : Cov(\eta_{i,j}, \eta_{k,j}) = \tag{7}$$

where ϕ_i and ρ_{ii} and ρ_{ij} are defined in Fu (1995).

1.4 Estimates of variability considering different weights for calculations using subsamples

To obtain different estimates of variability from empirical data, we used the approach developed by Achaz (2009) to estimate θ_{ξ_1}, θ_S and θ_π based on singletons, the total mutations and the nucleotide diversity, respectively. The variability for the total and for the subsample using unfolded and folded spectrum are:

For unfolded SFS:

$$\begin{aligned} \hat{\theta} &= \frac{1}{\sum_{i=1}^{n-1}\omega_i}\sum_{i=1}^{n-1}\omega_i i\xi_i \text{ for total samples, and} \\ \hat{\theta}_j &= \frac{1}{\sum_{i=1}^{n-1}\omega_i}\sum_{i=1}^{n-1}\omega_i i\xi_{i,j}\psi_{i,j}^{-1} \text{ for a subset of } j \text{ samples.} \end{aligned} \tag{8}$$

For folded SFS:

$$\begin{aligned} \hat{\theta}^* &= \frac{1}{\sum_{i=1}^{n/2}\omega_i^*}\sum_{i=1}^{n/2}\omega_i^*\eta_i\phi_i^{-1} \text{ for total samples, and} \\ \hat{\theta}_j^* &= \frac{1}{\sum_{i=1}^{n/2}\omega_i^*}\sum_{i=1}^{n/2}\omega_i^*\eta_{i,j}\phi_i^{-1}\psi_{i,j}^{-1} \text{ for a subset of } j \text{ samples.} \end{aligned} \tag{9}$$

The weights for a number of different estimators for folded and unfolded SFS are described in the Table 1 in Achaz (2009). The weights for variability estimates using singletons (Fu and Li, 1993), Watterson (Watterson, 1975), π (Tajima, 1983) and H (Fay and Wu, 2000a) (this last only for unfolded SFS) estimates are (Achaz, 2009):

For unfolded:

$$\begin{aligned} \omega_{i,\xi_1} &= 1 \text{ if } i = 1, \text{ 0 otherwise,} && \text{for Fu and Li's estimate,} \\ \omega_{i,S} &= i^{-1} && \text{for Watterson's estimate,} \\ \omega_{i,\pi} &= n - i && \text{for Tajima's estimate,} \\ \omega_{i,H} &= i && \text{for Fay and Wu's estimate,} \end{aligned} \tag{10}$$

For folded:

$$\begin{aligned}
\omega^*_{i,\eta_1} &= n \text{ if } i = 1,\ 0 \text{ otherwise,} && \text{or Fu and Li's estimate,} \\
\omega^*_{i,S} &= \frac{n}{i(n-i)(1+\delta_{i,n-i})} && \text{for Watterson's estimate,} \\
\omega^*_{i,\pi} &= \frac{n}{1+\delta_{i,n-i}} && \text{for Tajima's estimate,}
\end{aligned} \tag{11}$$

1.4.1 Considering missing data

Estimates of variability can be obtained using the same framework when missing data is present (Ferretti *et al.*, 2012). It is only necessary to account the number of samples that are present at each nucleotide. For unfolded and folded spectrum in a subset of size j, considering possible missing data, the formulation is, respectively:

$$\begin{aligned}
\hat{\theta}_j &= \frac{1}{L}\sum_{x=1}^{L}\sum_{i=1}^{n_x-1} i\omega_{i,n_x}\xi_{i,j}(x)\psi^{-1}_{i,j,n_x}, && \text{where } \frac{1}{L}\sum_{x=1}^{L}\sum_{n=1}^{n_x-1}\omega_{i,n_x} = 1. \\
\hat{\theta}^*_j &= \frac{1}{L}\sum_{x=1}^{L}\sum_{i=1}^{n_x/2} \omega^*_{i,n_x}\eta_{i,j}(x)\phi^{-1}_{i,n_x}\psi^{-1}_{i,j,n_x}, && \text{where } \frac{1}{L}\sum_{x=1}^{L}\sum_{n=1}^{n_x/2}\omega^*_{i,n_x} = 1.
\end{aligned} \tag{12}$$

and where ω_{i,n_x}, ω^*_{i,n_x}, ϕ_{i,n_x} and ψ_{i,j,n_x} consider the calculation in relation to the sample size n_x (with no missing samples) at the site x.

1.5 Neutrality test to compare variability estimates

Statistics based on estimates of variability given neutrality have been developed to contrast the neutral theory in empirical sequence data (*e.g.*, Tajima's D test, Fu and Li's D and F test, Fay and Wu's H test; Tajima, 1989; Fu and Li, 1993; Fay and Wu, 2000b). We are interested in constructing neutrality tests adapted to the relative Site Frequency Spectrum.

1.5.1 Neutrality Test considering different estimators of θ for each subsample

Achaz (2009) developed a general framework to develop neutrality tests by contrasting two estimators having different weights at each frequency. Here we propose to contrast the different subsamples of size j and $j^{'}$ using the same framework.

$$rT_{j,j'} = \frac{\hat{\theta}_j - \hat{\theta}'_{j'}}{\sqrt{Var(\hat{\theta}_j - \hat{\theta}'_{j'})}} \tag{13}$$

To do it, we need to calculate the variance of the denominator. The variance of θ under SNM is known and is (Achaz, 2009; Fu, 1995; Tajima, 1989):

$$Var(\hat{\theta}) \quad = Var(\frac{1}{\sum_i \omega_i} \sum_i \omega_i i \xi_i) = (\frac{1}{\sum_i \omega_i})^2 (\sum_i \omega_i^2 i^2 Var(\xi_i) + 2 \sum_i \sum_k ik\omega_i\omega_k Cov(\xi_i, \xi_k)) \tag{14}$$

where $var(\xi_i) = \frac{\theta}{i} + \sigma_{ii}\theta^2$, $cov(\xi_i, \xi_k) = \sigma_{ik}\theta^2$ and σ_{ii} and σ_{ik} are given in Fu (1995). The variance of θ estimates based on relative SFS are:

$$Var(\hat{\theta_j}) = Var(\frac{1}{\sum_i \omega_i} \sum_i \omega_i i \xi_{i,j} \psi_{i,j}^{-1}) = TODO \tag{15}$$

The variance of the difference of relative θ estimators ($Var(\hat{\theta_j} - \hat{\theta}'_{j'})$) is, TO INCLUDE.

1.5.2 Neutrality Test for a given subsample versus whole population

It is also possible to use the same framework to contrast the variability obtained from the whole population in relation to the variability estimated from the relative SFS at a given subsample (*e.g.*, an haplotype, a diploid individual, a group of individuals) or a subgroup of size j.

$$rT_{n,j} = \frac{\hat{\theta}_n - \hat{\theta}_j}{\sqrt{Var(\hat{\theta}_n - \hat{\theta}_j)}} \tag{16}$$

1.6 Simulation study

1.6.1 Modeling Standard Neutral Model

We have empirically validated the expectations of the site frequency spectrum using R, creating subsamples of j = 1, 2, 8, 32 and 64 from a total sample size of $n = 64$ with a total S = 1000 variants ($\theta = 211$). The next plots (Figure 2) show the fit of expectations versus observations of the mean rSFS for each subsample size, plus the variability and the SFS of the total sample

More, in order to validate the code and the pipeline to estimate the variability, we did coalescent simulations with *mlcoalsim* (Ramos-Onsins and Mitchell-Olds, 2007) using the stationary neutral model (SNM) for creating 100 alignments of $n = 64$ samples using

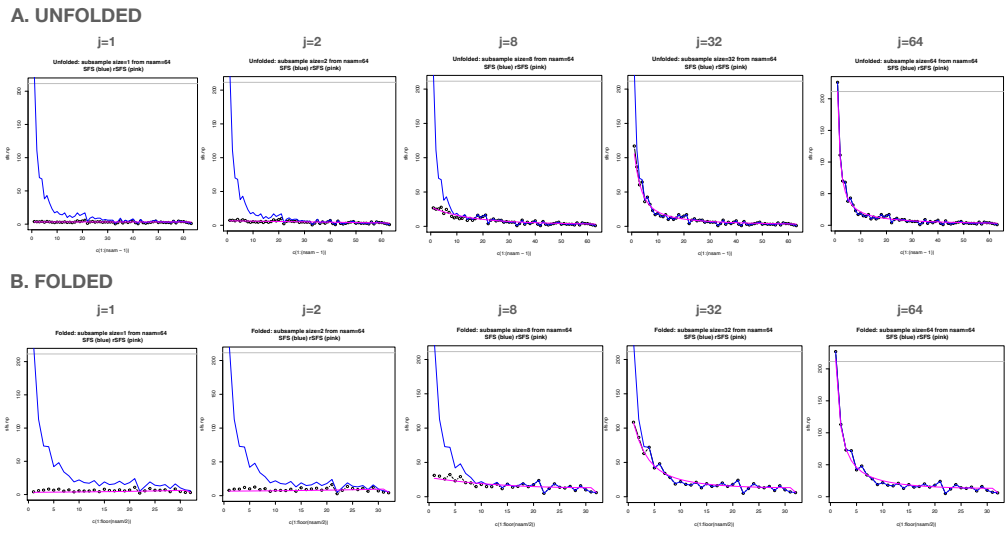

Figure 2: Validation of the expected rSFS for different subsample sizes given a total sample size of $n = 64$ under the SNM. The blue line indicates the SFS for the whole population. Black points indicate the mean of simulated rSFS observations for a subsample of size j, and pink line indicates the expected rSFS for a subsample of size j. A. Unfolded rSFS. B. Folded rSFS

$\theta = 0.01$ per nucleotide and $L = 10000$ base pairs. Each matrix was divided in subsamples of $j = 1, 2, 8, 32$ and 64 haplotypes and the estimates of variability were calculated considering the unfolded and the folded spectrum (Figure 3) using R. The mean estimates obtained with different subsamples fit perfectly to estimates obtained with the whole sample. We observe that the larger variance is for the Fu & Li variability estimator at small subsamples, both in unfolded and in the folded spectrum.

1.6.2 Modelling a colonization process in a 2D space

We use SLiM v4 to simulate a model of colonization into a 2D space matrix and compared with a simple stationary and panmictic neutral model where the geographical location was randomly assigned. First we simulated a grid of 5 x 5 spaces where a unique diploid population with Ne=4000 expands from the left bottom corner space to the adjacent spaces every 800 generations (0.8 Ne generations) in four consecutive steps until fill the whole space, where gene flow between adjacent grids is allowed with a migration rate of 0.01 per generation and the population size is modified in each step according to their capacity to survive in more adverse environmental conditions. We applied the equation $Np[i, j] = ratio_{col}^{(min(i,5-j))}$. That is, the population size is reduced at each ring. Once

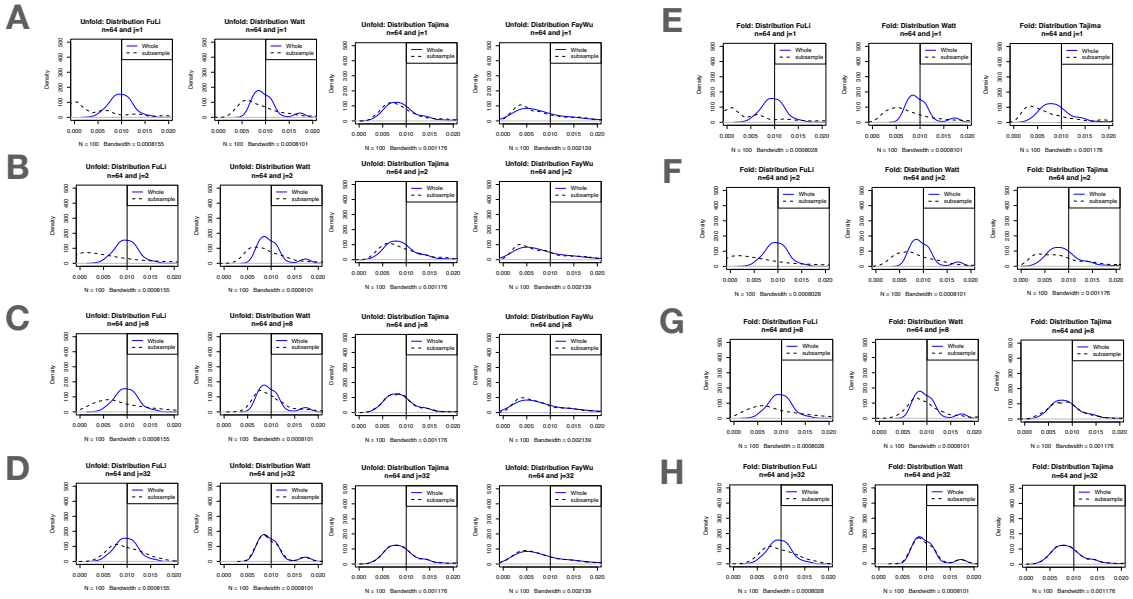

Figure 3: Distribution of θ for unfolded and folded estimators for whole sample and for subsamples under the SNM. Vertical line indicates the expected value. A. Unfolded θ estimators for size subsample $j = 1$. B. Unfolded θ estimators for size subsample $j = 2$. C. Unfolded θ estimators for size subsample $j = 8$. D. Unfolded θ estimators for size subsample $j = 32$. E. Folded θ estimators for size subsample $j = 1$. F. Folded θ estimators for size subsample $j = 2$. G. Folded θ estimators for size subsample $j = 8$. H. Folded θ estimators for size subsample $j = 32$.

simulation is finished, one, two or 20 chromosomes are sampled per grid, which will be used for the estimation of variability.

We collected one sample per each of the 5x5 spaces to see if the relative variability for each one has a differential value versus others. We did three different sampling strategies: one haploid individual per space, one diploid individual per space and a pool of 20 individuals per space. To confirm the robustness of the pattern we performed 100 iterations for each condition. pattern The colonization model shows clear differential patterns of variability across the space, even having a single haploid individual per space (Figure 4).

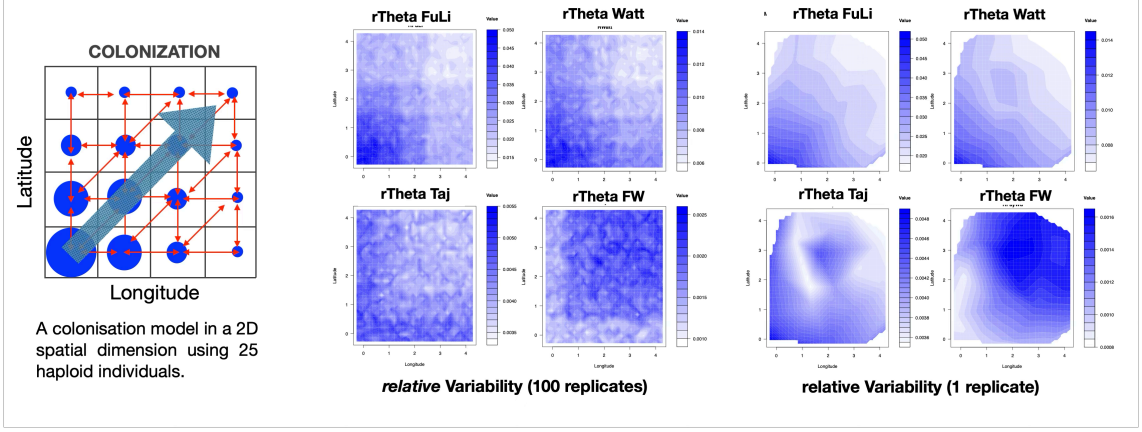

Figure 4: Write Legend

1.6.3 Application on GWAs considering effects affecting the frequency of the populations

1.7 Analyzing real dataset ?

2 Discussion

2.1 New estimators and SNM test to interpret the variability within species

2.2 An approach to complement the study of species in spatial gradients

2.3 An approach to study the relationship of variability among individuals of interacting species

References

Achaz, G., 2009 Frequency Spectrum Neutrality Tests: One for All and All for One. Genetics **183**: 249.

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

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

Ferretti, L., E. Raineri, and S. Ramos-Onsins, 2012 Neutrality tests for sequences with missing data. Genetics **191**: 1397–401.

Fu, Y.-X., 1995 Statistical properties of segregating sites. Theoretical Population Biology **48**: 172–197.

Fu, Y.-X., and W.-H. Li, 1993 Statistical tests of neutrality of mutations. Genetics **133**: 693.

Ramos-Onsins, S. E., and T. Mitchell-Olds, 2007 Mlcoalsim: multilocus coalescent simulations. Evol Bioinform Online **3**: 41–44.

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. Theoretical population biology **7**: 256.