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

Sebastian E. Ramos-Onsins, Yuliaxis Ramayo-Caldas

May 22, 2023

1 Methods

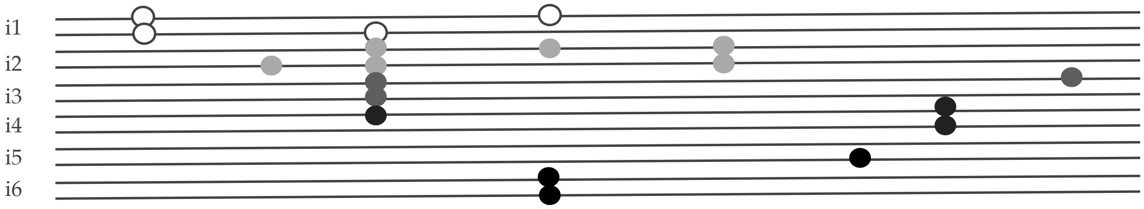
1.1 Concept

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 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. 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 $\hat{\theta}_{in} = i\xi_i$.



SFS	2	1	6	4	2	1	2	1
rSFS_i1	2	0	6	4	0	0	0	0
rSFS_i2	0	1	6	4	2	0	0	0
...								

Figure 1: Example of the calculation of the rSFS from the total SFS with subsamples of size $j = 2$.

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, ψ_{ij} indicates that is necessary to include an adjustment to achieve $\theta_{i,n} = \theta_{i,j}$ 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:

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

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

$$E(\xi_i)_j = E(\xi_i)_n(1 - \frac{\binom{n-i}{j}}{\binom{n}{j}}) = E(\xi_i)_n\psi_{i,j} = \frac{\theta}{i}\psi_{i,j}. \tag{3}$$

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 counfounded 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 $i \neq n-i$ and is 1 if $i = n-i$). 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:

$$E(\eta_i)_n = \frac{E(\xi_i)_n + E(\xi_{n-i})_n}{1+\delta_{i,n-i}} = \frac{\frac{\theta}{i} + \frac{\theta}{n-i}}{1+\delta_{i,n-i}} = \theta\frac{\frac{n}{i(n-i)}}{1+\delta_{i,n-i}} = \theta\phi_i. \tag{4}$$

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

$$E(\eta_i)_j \quad = E(\eta_i)_n\psi_{i,j} = \theta\phi_i\psi_{i,j}. \tag{5}$$

here $\psi_{i,j}$ is included to achieve $\theta_{i,n} = \theta_{i,j}$ when a subsampling of j haploypes is made.

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:

$$\begin{aligned}\hat{\theta}_n &= \frac{1}{\sum_{i=1}^{n-1}\omega_i}\sum_{i=1}^{n-1}\omega_i i\xi_i \text{ for total samples, and}\\ \hat{\theta}_j &= \frac{1}{\sum_{i=1}^{n-1}\omega_i}\sum_{i=1}^{n-1}\omega_i i\xi_{i,j}\psi_{i,j}^{-1} \text{ for a subset of } j \text{ samples.}\end{aligned} \tag{6}$$

For folded SFS:

$$\begin{aligned} \hat{\theta}^*_n &= \frac{1}{\sum_{i=1}^{n-1}\omega^*_i}\sum_{i=1}^{n-1}\omega^*_i\eta_i\phi_i^{-1} \text{ for total samples, and} \\ \hat{\theta}^*_j &= \frac{1}{\sum_{i=1}^{n-1}\omega^*_i}\sum_{i=1}^{n-1}\omega^*_i\eta_{i,j}\phi_i^{-1}\psi_{i,j}^{-1} \text{ for a subset of } j \text{ samples.} \end{aligned} \tag{7}$$

The weights for a number of different estimators for folded and unfolded SFS are described in the Table 1 in Achaz (2009).

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:

$$\begin{aligned} \hat{\theta}_j &= \frac{1}{L}\sum_{x=1}^{L}\sum_{i=1}^{n_x-1} i\omega_{i,n_x}\xi_{i,j}(x)\psi_{i,j,n_x}, & \text{where } \frac{1}{L}\sum_{x=1}^{L}\sum_{n=1}^{n_x-1} i\omega_{i,n_x} = 1. \\ \hat{\theta}^*_j &= \frac{1}{L}\sum_{x=1}^{L}\sum_{i=1}^{n_x-1} i\omega^*_{i,n_x}\eta_{i,j}(x)\phi^{-1}_{i,n_x}\psi_{i,j,n_x}, & \text{where } \frac{1}{L}\sum_{x=1}^{L}\sum_{n=1}^{n_x-1} i\omega^*_{i,n_x} = 1. \end{aligned} \tag{8}$$

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 Simulation comparison of the levels of variability between subsamples and the whole sample using the unfolded and the folded site frequency spectrum

1.5.1 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

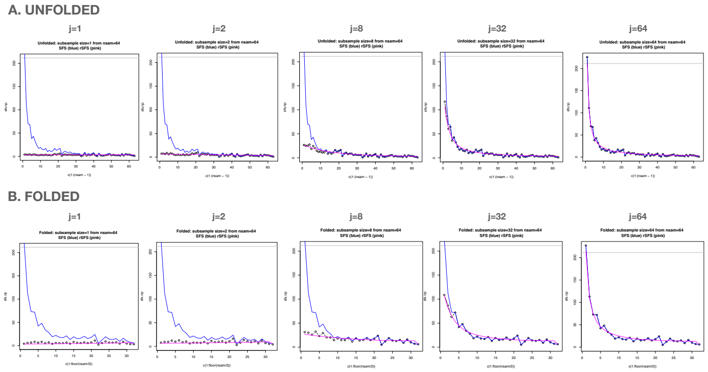


Figure 2: Validation of the expected rSFS for different subsample sizes given a total sample size of $n = 64$ under the SNM. The blue line indicates the SFS for the whole population. Black points indicate the mean of simulated rSFS observations for a subsample of size j, and pink line inidcates the expected 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 (Figure 3) and the folded (Figure 4) spectrum with 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.

1.5.2 Modeling a spatial gradient expanded population

We use SLiM v3.0 to simulate a population that expand into a 2D space matrix...

1.5.3 Modeling the effect of selection

We used *mlcoalsim* to simulate a selective sweep in a population. ...

Hard Selective Sweep.

Incomplete Selective Sweep.

Selective Sweep from standing variants.

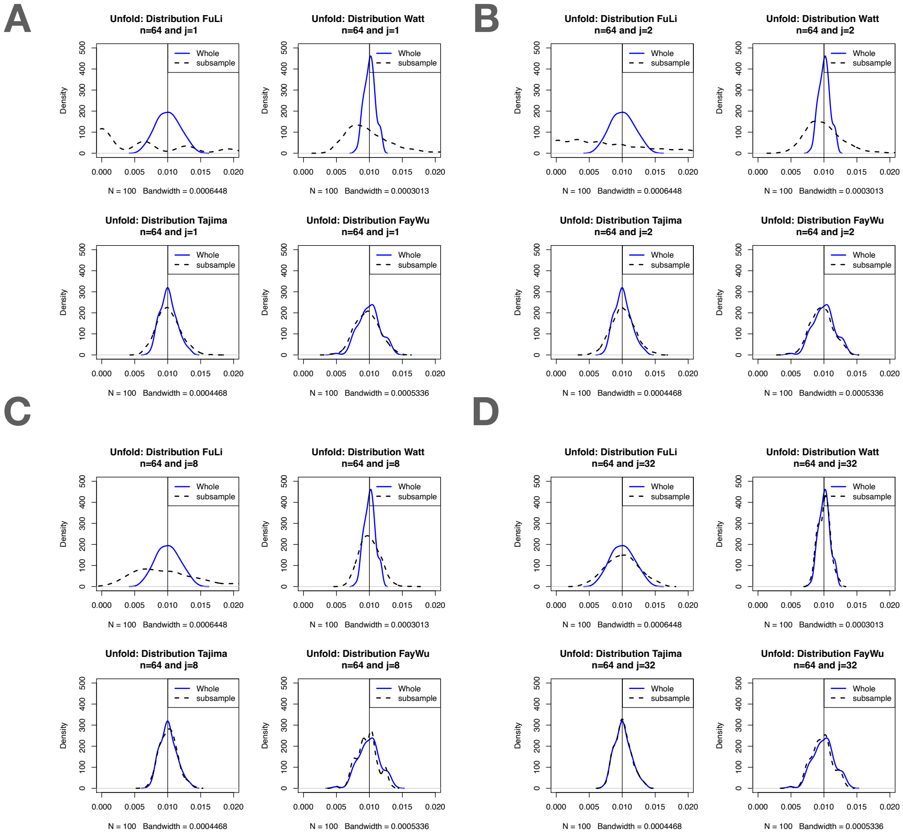


Figure 3: Distribution of θ unfolded estimators for whole sample and for subsamples under the SNM. A. Size subsample $j = 1$. B. Size subsample $j = 2$. C. Size subsample $j = 8$. D. Size subsample $j = 32$.

1.5.4 Modeling the effect of selection on specific samples under a spatial gradient expanded population

1.6 Neutrality test to compare variability estimates of total versus subsamples

1.6.1 Neutrality Test for a given subsample versus whole population

Achaz (2009) developed a general framework to develop neutrality tests by contrasting two estimators having different weights at each frequency. Here we propose 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

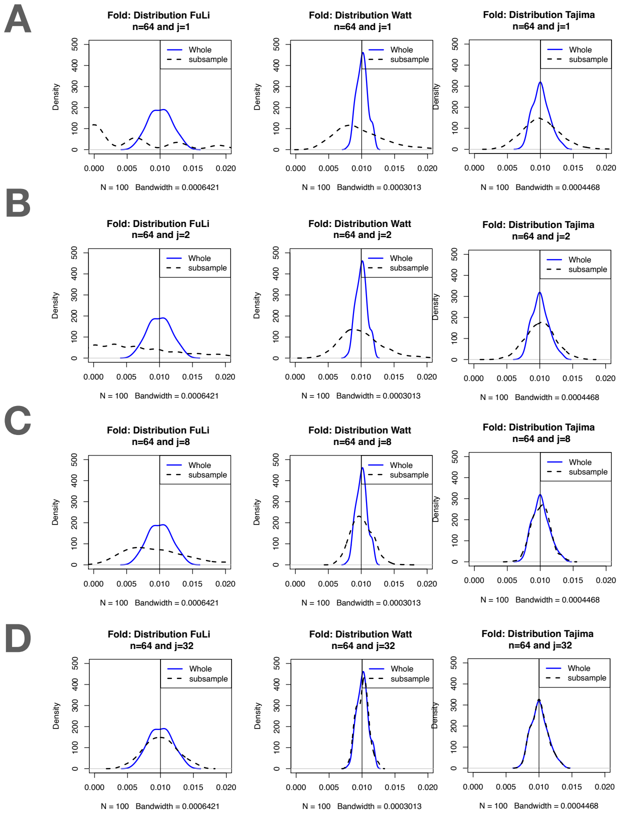


Figure 4: Distribution of θ folded estimators for whole sample and for subsamples under the SNM. A. Size subsample $j = 1$. B. Size subsample $j = 2$. C. Size subsample $j = 8$. D. Size subsample $j = 32$.

diploid individual, a group of individuals) or a subgroup of size j.

$$rT_{n,j} = \frac{\hat{\theta}_n - \hat{\theta}_j}{\sqrt{Var(\hat{\theta}_n - \hat{\theta}_j)}} \tag{9}$$

1.6.2 Neutrality Test considering different estimators within a given subsample

Following the same reasoning, it also possible to contrast the different subsamples of size j and $j^{'}$:

$$rT_{j,j'} = \frac{\hat{\theta}_j - \hat{\theta}'_{j'}}{\sqrt{Var(\hat{\theta}_n - \hat{\theta}'_{j'})}} \tag{10}$$

1.7 Calculation of F_{st} for each subsample versus total using different θ estimators

1.7.1 Using Mantel test for detecting differential rSFS patterns in relation to distance

1.8 Analyzing a real dataset (Pig/Almond?)

2 Discussion

2.1 The expectations of the patterns of the samples versus the whole population

2.2 An approach to study populations that are gradually changing without clear substructures. The variability at single individuals

2.3 rSFS and SDFS as complementary information for analysis on autopolyploid species

References

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

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.

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