PFcaller: A Frequency Caller for Pools and Polyploids.

May 22, 2023

The analysis in population genetics requires an unbiased estimate of the site frequency spectrum (SFS) to be accurate. The high-performance NGS methods produce a high number of reads for each base but also generates a similar or higher ratios of sequencing error than the population levels of variability that are wanted to be estimated. This effect becomes worst in case of having several numbers of lines per NGS lane. For diploid data, several methods to produce accurate results have been obtained (*e.g.*, MARUKI and LYNCH, 2015; NEVADO *et al.*, 2014; KORNELIUSSEN *et al.*, 2014; ROESTI *et al.*, 2012; HOHENLOHE *et al.*, 2010). Nevertheless, in case of using several or many lines within each NGS lane (pools or polyploids), the site frequency spectrum is difficult to calculate, specially for low frequency variants (FERRETTI *et al.*, 2013; RAINERI *et al.*, 2012). [Lynch, Kofler, Lynch et al. etc...]

In order to calculate accurately the SFS we think it is convenient to sacrifice the count of all individual lines included in the NGS lane (that may contain sequencing errors at low frequency) in order to have unbiased estimates of variability for any frequency considered at the new sample size (the number of lines accounted would be smaller than the included in the lane). Furthermore, not only the SFS would be unbiased but also the variance of the levels of variability would be similar to the

initial sample size included in the lane (Ferretti et al. pools) by averaging over all the studied region.

Algorithm

For a given site, we are interested in inferring the number of real lines that contributed to the observed reads. That is, the effective number of contributed lines having alternative sequence (ν_{CA}) and the effective number of contributed reference lines (ν_{CR}) that finally we observe in the form of reads (n_r) with some of them carrying the alternative (n_{rA}) or reference alleles (n_{rR}). We designate reference allele to the allele with a major frequency and alternative allele to the second allele (minor allele frequency in a biallelic context). Thus, for a given number of initial sampled lines (p) in a NGS lane, we are interested in calculating the following probability:

$$P(\nu_{CA}, \nu_{CR} | n_{rR}, n_{rA}, p). \tag{1}$$

In fact, the number of contributed lines and the frequency of alternative alleles would be affected by the level of the variability of the sampled population (per position, $\theta \ll 1$) and by the probability of having sequencing error (ξ) at this position an at each read variant. Then, reformulating the equation including new parameters we have:

$$P_{\nu_{CA}\nu_{CR}} = P(\nu_{CA}, \nu_{CR}, \pi_a, f_a, \nu_{\xi_A}, \nu_{\xi_R} | n_{rR}, n_{rA}, n_{r\xi}, \xi, \theta, p), \tag{2}$$

where π_a is the frequency of alternative allele samples in the pooled sample (of size p), f_a refers to the frequency of the alternative variant in the population and ν_{ξ_R} and ν_{ξ_A} refers to the number of errors in the reference and in the alternative

allele at that position, respectively. This probability can be decomposed in:

$$\begin{aligned} P_{\nu_{CA}\nu_{CR}} = P(\nu_{CA}, \nu_{CR}|\pi_a, f_a, \nu_{\xi_A}, \nu_{\xi_R}, n_{rR}, n_{rA}, n_{r\xi}, \xi, \theta, p)\cdot \\ P(\nu_{\xi_A}, \nu_{\xi_R}|\pi_a, f_a, n_{rR}, n_{rA}, n_{r\xi}, \xi, \theta, p)\cdot \\ P(\pi_a, f_a|n_{rR}, n_{rA}, n_{r\xi}, \xi, \theta, p). \end{aligned} \tag{3}$$

The first term of the equation are the parameters we are interested in. The other terms contain necessary parameters for estimating ν_{CA} and ν_{CR}.

Combinatorics

The first term probability of the equation 3 can be calculated as two independent probabilities, considering that the final number of reads for each variant (n_{rA}, n_{rR}) are consequence of sampling of the reads, in relation to their frequency (π_a, $p-\pi_a$), but also considering their error sequencing (ν_{ξ_A}, ν_{ξ_R}). Here we prefer to estimate ν_{CA} and $\nu_C = \nu_{CA} + \nu_{CR}$ instead of ν_{CR} directly:

$$\begin{aligned} P(\nu_{CA}, \nu_C|\pi_a, f_a, \nu_{\xi_A}, \nu_{\xi_R}, n_{rR}, n_{rA}, n_{r\xi}, \xi, \theta, p) = \\ P(\nu_{CA}|\nu_C, \pi_a, f_a, \nu_{\xi_A}, \nu_{\xi_R}, n_{rR}, n_{rA}, n_{r\xi}, \xi, \theta, p)\cdot \\ P(\nu_C|\pi_a, f_a, \nu_{\xi_A}, \nu_{\xi_R}, n_{rR}, n_{rA}, n_{r\xi}, \xi, \theta, p) = \\ P(\nu_{CA}|\nu_{CR}, \pi_a, \nu_{rA}) \cdot P(\nu_{CR}|\pi_r, \nu_{rR}), r \end{aligned} \tag{4}$$

where $\pi_r = p - \pi_a$, $\nu_{rR} = n_{rR} - \nu_{\xi_R} + \nu_{\xi_A}$ is the number of Reference reads without sequencing errors and $\nu_{rA} = n_{rA} - \nu_{\xi_A} + \nu_{\xi_R}$ is the number of Alternative reads without sequencing errors. The first term is obtained with a binomial and second term can be obtained from combinatorics (FERRETTI *et al.*, 2013, eq. 2)

using Stirling numbers. That is:

$$P(\nu_{CA}|\pi_a, \nu_{rA}) = \frac{\frac{\pi_a!}{(\pi_a-\nu_{CA})!}S(\nu_{rA}, \nu_{CA})}{\pi_a^{\nu_{rA}}} \text{ and}$$
$$P(\nu_{CR}|\pi_r, \nu_{rR}) = \frac{\frac{\pi_r!}{(\pi_r-\nu_{CR})!}S(\nu_{rR}, \nu_{CR})}{\pi_r^{\nu_{rR}}}, \tag{5}$$

where $S(\nu_r, \nu_C)$ are the Stirling numbers of the second kind (the number of ways to spread ν_r reads, $\nu_{rA} + \nu_{rR}$, into ν_C nonempty subsets from π independent samples).

In case of having a large ploidy and/or large n_r, (*i.e.*, total reads larger than 100) the calculation of combinatorics becomes unfeasible for calculating exactly the Stirling numbers of the second kind. For larger numbers, the average effective number of samples is estimated and fixed with probability 1. The estimated effective number of samples is (FERRETTI *et al.*, 2018):

$$E(\nu_{CA}) = \pi_a \left[1 - \left(1 - \frac{1}{\pi_a}\right)^{\nu_{rA}}\right] \tag{6}$$

Sequencing error

The sequencing error term (ξ) indicates the probability of a wrong classification of an observed read. Sequencing platforms provide an error term for each of the sequenced reads (Phred score error). This value is an approximate value and may be not accurate.

Estimating the read sequencing error from the whole dataset:

We propose two different approaches to estimate the number of sequencing errors per position. In case we have enough sequencing information it is possible to estimate the sequencing error rate (ξ) per position, instead of assuming the given phred scores values. LYNCH *et al.* (2014) developed a maximum likelihood (ML)

algorithm to estimate the sequencing error rate and the frequency of the alleles at each position, under the assumption of having population sequences with no more than biallelic variants and also assuming equal ratio of error for all four nucleotides. In this algorithm, the probability to observe a reference, alternative and error allele read (ϕ_R,ϕ_A and ϕ_ξ, respectively), assuming two real alleles at the given position, is (LYNCH *et al.*, 2014):

$$\begin{aligned}\phi_R &= \frac{\pi_r}{p}(1-\xi) + (1-\frac{\pi_r}{p})\frac{\xi}{3},\\ \phi_A &= \frac{\pi_a}{p}(1-\xi) + (1-\frac{\pi_a}{p})\frac{\xi}{3},\\ \phi_\xi &= \frac{2\xi}{3}.\end{aligned} \tag{7}$$

The likelihood function for real polymorphic positions is (LYNCH *et al.*, 2014):

$$L \propto \phi_R^{n_{rR}}\phi_A^{n_{rA}}\phi_\xi^{n_{r\xi}} \tag{8}$$

and an error estimate can be obtained from tri/tetra-allelic positions. The ML estimate of error rate is obtained by deriving L with respect ξ and equaling to zero: (LYNCH *et al.*, 2014):

$$\widehat{\xi} = \frac{3n_{r\xi}/n_r}{2}. \tag{9}$$

It is important to exclude highly variable regions, such as mutation hotspots or repetitive regions to estimate this error value. Note that multiple hits are also possible (with an approximated proportion of $(\theta a_n)^2$ per position in the sample). Thus, the estimated error rate per position should be compared with the expected proportion of multiple hits, although it is considered that error rate has a higher order of magnitude (on the order of θ).

Assuming as error ratio prior the phred score error values:

Instead, if we have no enough information to estimate the sequencing error from the sequence dataset (that is, not enough 'third' alleles), we may assume that the Phred score supplied by any sequencing platform is enough accurate to estimate the number of sequencing errors. In biallelic variants, if the p-value from the Phred score is ϵ, we can estimate the error using the approach indicated in RAINERI *et al.* (2012):

$$
\begin{aligned}
P(\nu_A|n_A) &= 1-\epsilon_A \ , P(\nu_R|n_A) = \epsilon_A, \\
P(\nu_R|n_R) &= 1-\epsilon_R \ , P(\nu_A|n_R) = \epsilon_R,
\end{aligned}
$$

and

$$
\begin{aligned}
P(n_A|\nu_A) &= 1 - \frac{\epsilon_R(1-2\epsilon_A)}{1-\epsilon_A-\epsilon_R} = 1-\xi_A, \\
P(n_R|\nu_A) &= \frac{\epsilon_R(1-2\epsilon_A)}{1-\epsilon_A-\epsilon_R} = \xi_A, \\
P(n_R|\nu_R) &= 1 - \frac{\epsilon_A(1-2\epsilon_R)}{1-\epsilon_A-\epsilon_R} = 1-\xi_R, \\
P(n_A|\nu_R) &= \frac{\epsilon_A(1-2\epsilon_R)}{1-\epsilon_A-\epsilon_R} = \xi_R.
\end{aligned}
\tag{10}
$$

Estimating the number of sequencing errors per position:

The estimate of the number of sequencing errors is obtained from the second term of the equation 3. These probabilities can be calculated with two independent binomials and using the probabilities from equations 7.

$$
\begin{aligned}
P(\nu_{\xi_A}, \nu_{\xi_R}|\pi_a, f_r, n_{rR}, n_{rA}, n_{r\xi}, \xi, \theta, p) = \\
P(\nu_{\xi_A}|\pi_a, n_{rA}, \xi, p) \cdot P(\nu_{\xi_R}|\pi_a, n_{rR}, \xi, p).
\end{aligned}
\tag{11}
$$

Note that in case the allele a is not present whatever error to non-reference nucleotides can be considered as an alternative allele. That means that in case that the number of errors is equal to the number of alternative alleles, the probability to observe the first different allele from reference is not ϕ_A but $\phi_{\overline{R}}$ (not ref-

erence). Therefore, we have two different calculations for $\nu_{\xi_A} = n_{rA}$ and for $0 < \nu_{\xi_A} < n_{rA}$. For $\nu_{\xi_A} = 0$, this probability is calculated from the difference of the sum of the probabilities of all $\nu_\xi > 1$ to 1.

$$P(\nu_{\xi_A}|\pi_a, n_{rA}, \xi, p) = \begin{cases} \binom{n_{rA}}{\nu_{\xi_A}} {p_{\nu_{\xi_A}}}^{\nu_{\xi_A}} (1 - p_{\nu_{\xi_A}})^{(n_{rA} - \nu_{\xi_A})} & \text{if } 0 < \nu_{\xi_A} < n_{rA}, \\ p_{\nu_{\xi_{1A}}} \cdot \binom{n_{rA}-1}{\nu_{\xi_A}-1} {p_{\nu_{\xi_A}-1}}^{\nu_{\xi_A}-1} (1 - p_{\nu_{\xi_A}-1})^{(n_{rA} - \nu_{\xi_A})} & \text{if } \nu_{\xi_A} = n_{rA}, \\ 1 - \sum P(\nu_{\xi_A} > 0|\pi_a, n_{rA}, \xi, p) & \text{if } \nu_{\xi_A} = 0, \end{cases} \tag{12}$$

where $p_{\nu_{\xi_A}} = \frac{(1-\frac{\pi_a}{p})\frac{\xi}{3}}{\frac{\pi_a}{p}(1-\xi)+(1-\frac{\pi_a}{p})\frac{\xi}{3}}$ and $p_{\nu_{\xi_{1A}}} = \frac{(1-\frac{\pi_a}{p})\xi}{\frac{\pi_a}{p}(1-\xi)+(1-\frac{\pi_a}{p})\xi}$.

and $P(\nu_{\xi_R}|\pi_a, n_{rR}, \xi, p)$ is obtained using the same calculations than in 12. The probability to have as many errors as reference reads is very low, except if the number of reads is also very low.

Estimating the pool and the sample frequency

The third term probability of the equation 3 can be calculated considering the model of sequencing errors shown above. That is:

$$P(\pi_a, f_a|n_{rR}, n_{rA}, n_{r\xi}, \xi, \theta, p) = \frac{P(n_{rR}, n_{rA}, n_{r\xi}|\pi_a, f_a, \xi, \theta, p) \cdot P(\pi_a|f_a, p, \theta) \cdot P(f_a|p, \theta)}{\sum_{\pi_{Ai}=0}^{\pi_{Ai}=p} \sum_{f_{Ai}=0}^{f_{Ai}=1} P(n_{rR}, n_{rA}, n_{r\xi}|\pi_{Ai}, f_{Ai}, \xi, \theta, p) \cdot P(\pi_{Ai}|f_{Ai}, p, \theta) \cdot P(f_{Ai}|p, \theta)}. \tag{13}$$

When considering $P(n_{rR}, n_{rA}, n_{r\xi}|\pi_a, f_a, \xi, \theta, p) = P(n_{rR}, n_{rA}, n_{r\xi}|\pi_a, \xi, p)$, is possible to confound $n_{r\xi}$ with n_{rA} when the true n_{rA} is smaller than $n_{r\xi}$. In those cases, this probability is dependent whether comes from an error or from a real alternative variant. In order to consider this artefact, the probability should consider the number of different alleles observed. That is, the number of different alleles

and their frequency determine the value of n_{rR} and n_{rA} in the calculation, being n_{rR} the most frequent allele and n_{rA} the second most frequent. The probability of observations must consider alternative alleles all the combinations were error reads can substitute the alele A because they are observed at higher frequency. Error reads are separated in the two possible allele variants ($n_{r\xi} = n_{r\xi_1} + n_{r\xi_2}$, and $\phi_{\xi_1} = \phi_{\xi_2} = \xi/3$). In case we know the alleles that correspond to each of the reads, it can be formulated as a multinomial:

$$P'(n'_{rR},n'_{rA},n'_{r\xi_1},n'_{r\xi_2}|\pi_a,\xi,p) = \frac{n_r!}{n'_{rR}!n'_{rA}!n'_{r\xi_1}!n'_{r\xi_2}!} \cdot {\phi_r}^{n'_{rR}} \cdot {\phi_a}^{n'_{rA}} \cdot {\phi_{\xi_1}}^{n'_{r\xi_1}} \cdot {\phi_{\xi_2}}^{n'_{r\xi_2}} \tag{14}$$

Here, n'_{rR}, n'_{rA}, $n'_{r\xi_1}$ and $n'_{r\xi_2}$ are the real corresponding number of alleles from r and a and from each error sequencing, respectively.

The probability for each combination of observed reads must include all these combinations that have a major and a minor allele, whatever is the real allele. That is, the observed alleles are reordered from major to minor, and the frequency is folded. The observed $(n_{rR}, n_{rA}, n_{r\xi_1}, n_{r\xi_2}) = sort(n'_{rR}, n'_{rA}, n'_{r\xi_1}, n'_{r\xi_2})$. Therefore this probability must sum all cases where sorting the real alleles by size gives the same number of observed major and minor alleles:

$$\begin{aligned} &P(n_{rA},n_{rR},n_{r\xi_1},n_{r\xi_2}|\pi_a,\xi,p) = \\ &\sum \Big[P'(sort(n'_{rR},n'_{rA},n'_{r\xi_1},n'_{r\xi_2}) = (n_{rR},n_{rA},n_{r\xi_1},n_{r\xi_2})|\pi_a,\xi,p) + \\ &\quad P'(sort(n'_{rR},n'_{rA},n'_{r\xi_1},n'_{r\xi_2}) = (n_{rR},n_{rA},n_{r\xi_1},n_{r\xi_2})|\pi_r,\xi,p) \Big] / (2 - \delta_{\pi_a,\pi_r}) \end{aligned} \tag{15}$$

$P(\pi_a|f_a,p,\theta)$ can be estimated with binomials (considering a folded site frequency spectrum, that is, no outgroup):

$$P(\pi_a|f_a,p,\theta) = \left[\binom{p}{\pi_a} (\frac{f_a}{N})^{\pi_a} (1-\frac{f_a}{N})^{(p-\pi_a)} + \binom{p}{\pi_a} (\frac{f_a}{N})^{(p-\pi_a)} (1-\frac{f_a}{N})^{\pi_a} \right] / (1+\delta_{\pi_a,p-\pi_a}), \tag{16}$$

where N is the population size and $\delta_{\pi_a,p-\pi_a}$ = 1 if $\pi_a = p - \pi_a$, otherwise is 0. Finally, $P(f_a|\theta,p)$ is calculated assuming a SNM model as a prior (see next subsection).

The SNM as the prior frequency for the whole sample:

We consider the calculation of $P(f_a|\theta)$ assuming the SNM and a given variability value, although it is possible consider other *a priori* conditions. The probability of having a given sample frequency may be assumed as the theoretical SFS under the stationary Standard Neutral Model (SNM).

Assuming the derived variant is unknown (no ancestral variant or outgroup is available) and considering no more than two variants, the prior probability is calculated as:

$$P_n(f_a,|\theta) = \begin{cases} \theta \left(\frac{1}{f_a} + \frac{1-\delta_{f_a,N-f_a}}{N-f_a} \right) & \text{if } N/2 \geqslant f_a > 0 \\ 1 - \theta a_n & \text{if } f_a \text{ is } 0, \end{cases} \tag{17}$$

Here $a_n = \sum_{i=1}^{i=N-1} 1/i$ and $\theta \ll 1$. Assuming a large population size, we approximate $a_n \simeq 0.578 + log(N-1)$, f_a is the frequency of the minor allele, $\delta_{f_a,N-f_a}$ = 1 if $f_a = N - f_a$, otherwise is 0 . Note the subindex in the probability (P_n) to differentiate this probability with the probability calculated considering the ancestral variant.

Alternative priors: the Exponential model and the Uniform distribution:

The exponential model is a convenient model in case of using pools that have experimented a fast growth, such as a bacterial growth colony, infection process or cancer cell expansion. In such cases the prior used considering no more than two

variants is (OHTSUKI and INNAN, 2017) and considering the folded spectrum:

$$P_n(f_a|\theta) = \begin{cases} \theta\left(\frac{1}{(f_a(f_a+1))} + \frac{1-\delta_{f_a,N-f_a}}{(N-f_a)(N-(f_a+1))}\right) & \text{if } N/2 \geqslant f_a > 0 \\ 1-\theta a_e & \text{if } f_a \text{ is 0,} \end{cases} \quad (18)$$

Here $a_e = \sum_{i=1}^{i=N-1} 1/(i(i+1))$ and $\theta \ll 1$. Assuming a large population size, $a_e \simeq 1 + log(p-1) - log(N)$.

If the user does not want to consider informative priors and prefer to use a flat distribution with equal probabilities to each frequency, a uniform distribution is considered.

$$P_n(f_a|\theta) = \begin{cases} \theta(\frac{1}{N-1} + \frac{1-\delta_{f_a,N-f_a}}{N-1}) & \text{if } N/2 \geqslant f_a > 0 \\ 1-\theta & \text{if } f_a \text{ is 0,} \end{cases} \quad (19)$$

Estimation of the global levels of variability from the whole observed data:

In case the θ prior would not directly assigned by the researcher, an alternative is to estimate the level of variability from the global dataset. LYNCH *et al.* (2014) ML algorithm can be used to estimate the frequency of the major allele per position and then use this estimates to obtain the global level of variability. For each position x:

$$\widehat{f_{A_x}} = \frac{\frac{n_{rA}}{n_{rA}+n_{rR}}[1-\frac{2\widehat{\xi}}{3}] - \frac{\widehat{\xi}}{3}}{1-\frac{4\widehat{\xi}}{3}}. \quad (20)$$

The global estimation of θ is obtained with Tajima's heterozygosity estimation:

$$\widehat{\theta} = \frac{1}{L}\sum_{x=1}^{L} 2f_{A_x}(1-f_{A_x})\frac{n_{C_x}-1}{n_{C_x}}, \quad (21)$$

where n_{C_x} is the mean number of contributed lines at position x given by the number of pooled samples and the total number of reads at this position (following

equation 6).

Methodology to validate results

In order to estimate the accurateness of the algorithm, we used different estimators of the level of variability based on the frequency spectrum and considering missing data (FERRETTI *et al.*, 2012). The reason to use this estimators is because in real data we can not reconstruct a single frequency spectrum for the whole genome but the variability, because the sample size can be different at each position. The different estimators account for the different frequencies, so biases can be detected.

The estimators are based in this expression:

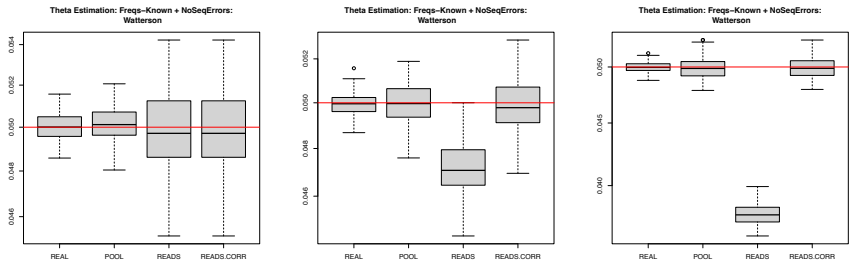
$$\hat{\theta} = \frac{1}{L}\sum_{x=1}^{L}\sum_{i=1}^{n_x-1} i\omega_{i,n_x}\xi_i(x) \quad , \quad \frac{1}{L}\sum_{x=1}^{L}\sum_{i=1}^{n_x-1}\omega_{i,n_x} = 1 \tag{22}$$

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 weights ω_{i,n_x}, Ω_{i,n_x} define the specific estimator (ACHAZ, 2009), where n_x is the sample size at position x.

The used estimators are Watterson (WATTERSON, 1975), Tajima (TAJIMA, 1983), Fu & Li (FU and LI, 1993), Fay & Wu (FAY and WU, 2000), Zeng (ZENG *et al.*, 2006), Ferretti (FERRETTI *et al.*, 2017) and Watterson and Tajima estimators without singletons (ACHAZ, 2009). The weights are obtained from (ACHAZ, 2009) and HELLMANN *et al.* (2008) as is indicated in (FERRETTI *et al.*, 2012). Under a stationary Neutral Model, all these estimates of the variability should give statistically equivalent values. These estimators are focused on different sections of the SFS (they have different weights) and thus can give us an accurate idea about the estimation of the SFS using this method.

Validation of each of the methodological steps using R scripts

Validation of the three main steps (Combinatorics, Sample frequency and Sequencing error estimation) is initially tested using R scripts. First we tested the estimation of the contributed samples using the combinatorics expression, where the frequency of the sample are known and there are no sequencing errors. We tested for a pool sample size of $n = 50$ having a variation of $theta = 0.05$ under the SNM and having 2 (no repeated samples), 20 (several are repeated) and 200 (all samples are many times repeated) reads per position (Figure 1). We see in that the estimation of variability is accurate in all combinations and estimates of θ.

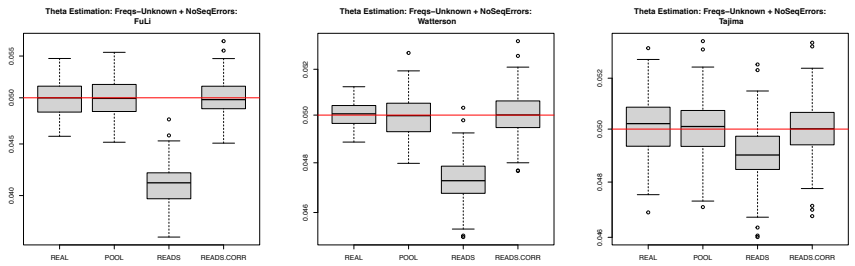


(a) Watterson θ Estimate for number of reads = 2 (b) Watterson θ Estimate for number of reads = 20 (c) Watterson Estimate for number of reads = 200

Figure 1: Testing the inference of contributed lines with the Combinatorics expressions. For each subfigure, the first column shows the variability in the population for 100 bins of 10000 positions, the second column is the variability for a pooled sample of 50 samples, the third is the variability for the sample reads, and the fourth is the variability for the contributed sample reads (*i.e.*, after correction with this algorithm).

Next, we tested the inference of the sample frequency (together with combinatorics) using a pool sample size of $n = 50$, $theta = 0.05$ under the SNM and having 20 reads per position. We tested the estimation of theta using the Fu & Li, Wattterson and Tajima θ estimators, which are estimating the variability using only singletons, the number of total variants and the average number of pairwise differences, respectively (see Figure 2). We observe that the different estimators are estimating the correct value of variability, which suggest no deviations of the

inferred frequency in relation to the expected frequency of the variants.



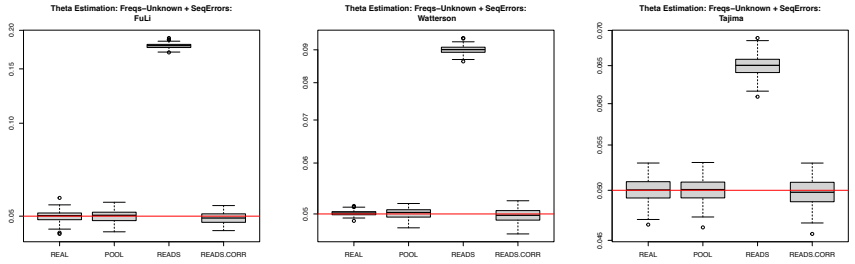
(a) Fu & Li θ Estimate (b) Watterson θ Estimate (c) Tajima θ Estimate

Figure 2: Testing the inference of frequency variants lines with the Frequency Sample expressions. See explanation of the columns in legend of Figure 1

Finally, we tested the inference of sequencing errors (together with combinatorics and frequency inference) using a sequencing error rate $\xi = 0.01$ and two different levels of variability ($\theta = 0.05$ and $\theta = 0.01$), considering 20 reads per position. Different estimators of the levels of variability were also tested in order to detect deviations (Figure 3). We see that the patterns of variability are correctly estimated, even in case of a¡having the same variability ratio than the sequencing error ratio.

Implementation

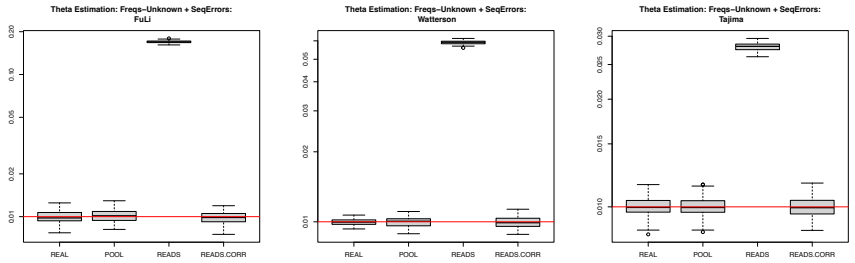
First, a filtering step for homologous positions of NGS data was performed as indicated above. Afterwards, the calculation of the posterior distribution was performed with our algorithm using the Metropolis algorithm method. In our implementation, the Metropolis algorithm starts using the mean observed frequency and p values for the given sampled position to obtain the initial parameters for the first iteration (rounded to the lowest integer for each frequency) and to calculate the log-probability for this parameters ($f(x) = log(P_{\nu_{CA}\nu_{CR}})$). The second step implies to do a number of iterations. They are performed in the following way: For each iteration generate a new set of parameters (let's say x') according to the parameters of the previous data, using the conditional probability for calculating the new pro-



(a) Fu & Li θ Estimate for $\theta = 0.05$ and $\xi = 0.01$

(b) Watterson θ Estimate for $\theta = 0.05$ and $\xi = 0.01$

(c) Tajima θ Estimate for $\theta = 0.05$ and $\xi = 0.01$



(d) Fu & Li θ Estimate for $\theta = 0.01$ and $\xi = 0.01$

(e) Watterson θ Estimate for $\theta = 0.01$ and $\xi = 0.01$

(f) Tajima θ Estimate for $\theta = 0.01$ and $\xi = 0.01$

Figure 3: Testing the inference of sequencing errors with the Sequencing errors expressions. See explanation of the columns in legend of Figure 1.

posal, $P(\nu_{CA'}, \nu_{CR'}, \pi_{R'}, \nu_{\xi'_A}, \nu_{\xi'_R} | \nu_{CA}, \nu_{CR}, \pi_R, \nu_{\xi_A}, \nu_{\xi_R})$. Next, it is calculated the acceptance ratio $\alpha = \frac{f(x')}{f(x)}$If f $\alpha \geqslant ran(1)$ (where $ran(1)$ is a random number between 0 and 1) then the new parameters set is accepted and becomes the new step in the Markov Chain. Otherwise, reject the new set of parameters and the current set will be again the new step in the Markov Chain.

A period of burnin (that is, an initial process in which accepted chains are not used for the posterior distribution) becomes unnecessary when using the whole prior distribution for sampling new parameters. This algorithm is some less efficient than MHMCMC but simpler. Only 1/10 iterations were kept for the analysis of the posterior distribution in order to avoid correlation among contiguous sets in the chain. The number of effective iterations performed to calculate the posterior distribution of the parameters for each position was set to 5000.

The inference of the best n_C and ν_{CA} (or all four bases, that is ν_{CAs}, n_{CTs}, n_{CCs} and n_{CGs}) were obtained from sampling the joint posterior distribution. We translated the obtained values to a multi-fasta file by rounding the number of effective contributed lines and sampling the frequencies from the values of the posterior distribution.
The implementation will works well for $min(n_r, ploidy) < 5000$.

Validation with NGS simulated data

A number of simulations using different number of initial samples per NGS lane and different read depth were performed in order to study the behaviour of the method. The ranges of initial sample sizes per lane vary from 2, 8, 64, and 128, which includes a wide number of different experiments (from diploid and polyploid individuals to moderated size pools). The ranges of read depth per lane included were 2, 4, 8, 16 and 64.

The simulations were performed using the Stationary Standard Neutral Model. That means that all the different used estimations of the nucleotide variability should give the same result under this conditions. Finally, in order to test the prior distributions included (which assume the SNM), we also tested other evolutionary models: (i) expansion model, which results in an excess of singletons in relation to the SNM; and (ii) an inbreeding diploid population, which results in a defect of singletons in relation to the SNM.

Empirical data analysis

Sequencing data from Drosophila pools is used and the results are contrasted and validated with results obtained from individual sequencing and with results using other methodologies.

References

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

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

FERRETTI, L., A. KLASSMANN, E. RAINERI, S. E. RAMOS-ONSINS, T. WIEHE, *et al.*, 2018 The neutral frequency spectrum of linked sites. Theor Popul Biol **123**: 70–79.

FERRETTI, L., A. LEDDA, T. WIEHE, G. ACHAZ, and S. E. RAMOS-ONSINS, 2017 Decomposing the site frequency spectrum: The impact of tree topology on neutrality tests. Genetics **207**: 229–240.

FERRETTI, L., E. RAINERI, and S. RAMOS-ONSINS, 2012 Neutrality tests for sequences with missing data. Genetics **191**: 1397–401.

FERRETTI, L., S. E. RAMOS-ONSINS, and M. PÉREZ-ENCISO, 2013 Population genomics from pool sequencing. Mol Ecol **22**: 5561–76.

FU, Y., and W.-H. LI, 1993 Statistical tests of neutrality of mutations. Genetics **133**: 693.

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 research **18**: 1020–1029.

HOHENLOHE, P. A., S. BASSHAM, P. D. ETTER, N. STIFFLER, E. A. JOHNSON, *et al.*, 2010 Population genomics of parallel adaptation in threespine stickleback using sequenced rad tags. PLoS Genet **6**: e1000862.

KORNELIUSSEN, T. S., A. ALBRECHTSEN, and R. NIELSEN, 2014 Angsd: Analysis of next generation sequencing data. BMC Bioinformatics **15**: 356.

LYNCH, M., D. BOST, S. WILSON, T. MARUKI, and S. HARRISON, 2014 Population-genetic inference from pooled-sequencing data. Genome Biol Evol **6**: 1210–8.

MARUKI, T., and M. LYNCH, 2015 Genotype-frequency estimation from high-throughput sequencing data. Genetics **201**: 473–86.

NEVADO, B., S. E. RAMOS-ONSINS, and M. PEREZ-ENCISO, 2014 Resequencing studies of nonmodel organisms using closely related reference genomes: optimal experimental designs and bioinformatics approaches for population genomics. Mol Ecol **23**: 1764–79.

OHTSUKI, H., and H. INNAN, 2017 Forward and backward evolutionary processes and allele frequency spectrum in a cancer cell population. Theor Popul Biol **117**: 43–50.

RAINERI, E., L. FERRETTI, A. ESTEVE-CODINA, B. NEVADO, S. HEATH, *et al.*, 2012 Snp calling by sequencing pooled samples. BMC Bioinformatics **13**: 239.

ROESTI, M., A. P. HENDRY, W. SALZBURGER, and D. BERNER, 2012 Genome divergence during evolutionary diversification as revealed in replicate lake-stream stickleback population pairs. Mol Ecol **21**: 2852–62.

TAJIMA, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105**: 437.

WATTERSON, G., 1975 On the number of segregating sites in genetical models without recombination. Theoretical population biology **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.