# Supplementary Text: Pool-Sequencing Equations

## GrENEnet pilot: A large-scale evole-and-resequence experiment...

Lucas Czech and Moisés Expósito-Alonso

This document describes our assessment and re-rendering of the equations originally presented in PoPoolation [3] and PoPoolation2 [4]. The aim of these equations is to correct for biases of pool sequencing when computing measures of diversity (such as Tajima's D) and differentiation (such as $\mathrm{F}_{\mathrm{ST}}$). This document is based on the equations as provided in the PoPoolation document `correction_equations.pdf` as found in their code repository; we here try to describe the equations in more detail, fix some of their mistakes, and improve some details of the computations. Apart from that, we do not add or change anything.

## 1 Definitions

We here assume basic familiarity with pool sequencing concepts. Please refer to Kofler *et al.* (2011a) [3] for details on pool sequencing data and its biases. TODO: maybe we should cite some more sources here?

### 1.1 Pool Sequencing Data

We first define the input that we assume to be given for all subsequent equations. In the implementation, these would either be based on the input data, or set by the user.

$C$ : Observed coverage. This is the number of reads that span the given position in the genome.

$b$ : Minimum allele count, provided by the user. We do not want to consider SNPs with fewer than $b$ alternative reads in the data, as they might be sequencing errors. Note that we assume $b$ to be a user-provided constant, and hence leave it out of (most) function arguments for simplicity.

$n$ : Pool size, provided by the user. This is the number of individuals that were pooled together for sequencing.

### 1.2 Notation

$\tau$ : Nucleotides, with $\tau \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$.

$i_\tau$ : Nucleotide counts, i. e., how many reads have a certain nucleotide $\tau$ at a given genomic position. Hence, $C = \sum_\tau i_\tau$.

$\boldsymbol{i}$ : Vector of nucleotide counts (for convenience), i. e., $\boldsymbol{i} = (\ i_\mathrm{A}, i_\mathrm{C}, i_\mathrm{G}, i_\mathrm{T}\ )$.

$f_\tau$ : Nucleotide frequencies, i. e., $f_\tau = i_\tau / C$.

$\boldsymbol{f}$ : Vector of nucleotide frequencies (for convenience), i. e., $\boldsymbol{f} = (\ f_\mathrm{A}, f_\mathrm{C}, f_\mathrm{G}, f_\mathrm{T}\ )$.

$u$, $v$ : For biallelic SNP positions, we simplify and instead of the four $i_\tau$ values, just use $u$ for the count of the reference allele, and $v$ for the count of the alternative allele.

$m$ : Index of summation over potential levels of coverage $C$.

$k$ : Index of summation over potential pool sizes $n$.

## 1.3 Harmonic Numbers

We define $a_1$ and $a_2$ based on (generalized) harmonic numbers, as the sum of (squared) reciprocals of the first $n-1$ positive integers:

$$a_1(n) = \sum_{k=1}^{n-1} \frac{1}{k} \tag{1}$$

$$a_2(n) = \sum_{k=1}^{n-1} \frac{1}{k^2} \tag{2}$$

These will be needed in several of the below equations.

# 2 Theta Pi

Here, we derive equations for $\theta_\pi$, also called Tajima's $\pi$, based on its original (classic) definition, but correcting for biases introduced by pool sequencing.

## 2.1 Pool-Sequencing Correction

### Definition for biallelic SNPs

In order to derive the pool-sequencing corrected equations, we first define $\theta_\pi$ as usual:

$$\theta_\pi(\boldsymbol{f}, C) = \frac{C}{C-1}\left(1 - \sum_\tau f_\tau^2\right) \tag{3}$$

using nucleotide frequencies $\boldsymbol{f} = (\ f_\mathrm{A}, f_\mathrm{C}, f_\mathrm{G}, f_\mathrm{T}\ )$, with $\tau \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$, and $C$ sequences in the pool in total.

See for example Equation (3.1) of Hahn (2018) [2] for the original definition for individuals; we here use this as the starting point for pools of individuals, by replacing the number of individuals with the number of sequences $C$ in the pool.

Using nucleotide counts $i_\tau$ instead, that is, $f_\tau = i_\tau / C$, we can then reformulate this as:

$$\begin{aligned}\theta_\pi(\boldsymbol{i}, C) &= \frac{C}{C-1}\left(1 - \sum_\tau \frac{i_\tau^2}{C^2}\right) \qquad (4)\\ &= \frac{C}{C-1} - \sum_\tau \frac{i_\tau^2}{C(C-1)}\\ &= \frac{C^2}{C(C-1)} - \sum_\tau \frac{i_\tau^2}{C(C-1)}\end{aligned}$$

Now, we use $C = \sum_\tau i_\tau$ to extend the numerators:

$$\begin{aligned}&= \frac{C^2 - C}{C(C-1)} - \sum_\tau \frac{i_\tau^2 - i_\tau}{C(C-1)}\\ &= \frac{C(C-1)}{C(C-1)} - \sum_\tau \frac{i_\tau(i_\tau - 1)}{C(C-1)}\\ &= 1 - \sum_\tau \frac{i_\tau(i_\tau - 1)}{C(C-1)} \qquad (5)\end{aligned}$$

For a simple biallelic SNP, we only have two counts with $C = u + v$ instead of four $i_\tau$ counts. Substituting this in Eq. (4), we get:

$$\begin{aligned}
\theta_\pi(u, v, C) &= \frac{C}{C-1}\left(1 - \frac{u^2}{C^2} - \frac{v^2}{C^2}\right) \\
&= \frac{C^2}{C(C-1)}\left(1 - \frac{u^2}{C^2} - \frac{v^2}{C^2}\right) \\
&= \frac{C^2}{C(C-1)} - \frac{u^2}{C(C-1)} - \frac{v^2}{C(C-1)}
\end{aligned}$$

This contains redundant information; let's further simplify using $v = C - u$:

$$\begin{aligned}
\theta_\pi(u, C) &= \frac{C^2}{C(C-1)} - \frac{u^2}{C(C-1)} - \frac{(C-u)^2}{C(C-1)} \\
&= \frac{C^2 - u^2 - (C-u)^2}{C(C-1)} \\
&= \frac{(C+u)(C-u) - (C-u)^2}{C(C-1)} \\
&= \frac{[(C+u) - (C-u)]\,(C-u)}{C(C-1)} \\
&= \frac{2u(C-u)}{C(C-1)} \qquad (6)
\end{aligned}$$

This is the basic equation for $\theta_\pi$ for a biallelic SNP that we will use for pool sequencing, with a coverage of $C$, composed of two SNP counts with $u$ as the first and $v$ as the second allele count, and $u + v = C$.

### Expected value

We now use this to compute the expected value of $\theta_\pi$ for biallelic SNPs, conditioned on the total coverage $C$:

$$\mathbb{E}(\theta_\pi|C, n) = P(\text{SNP}|n) \cdot \sum_{m=b}^{C-b} \theta_\pi(m, C) \cdot P(m|C, n) \qquad (7)$$

In words, the expected value is computed by summing all possible SNP counts (that exceed the minimum count $b$) that can occur in a pool with coverage $C$ (using the first allele count $m$ here, with second allele count $C - m$ implicit), weighted by the probability to have each of those counts, and scaled by the probability to have a SNP in the first place.

Here, we are using the minimum allele count $b$ that we want to consider (as provided by the user), meaning that we only consider SNPs that have at least $b$ reads for either the first or second allele. As we are only using the first allele count $m$ in the equation above, and do not know which of the two counts is the larger one, we "sandwhich" our potential values for the coverage between $b$ and $C - b$.

The two probabilities used above are computed as follows.

$P(\text{SNP}|n)$ is the probability of observing a SNP in a pool of $n$ individuals:

$$P(\text{SNP}|n) = \theta \sum_{k=1}^{n-1} \frac{1}{k} = \theta a_1(n) \qquad (8)$$

Assuming that all variation is neutral, and that the population is of constant size and in mutation-drift equilibrium, by definition, $\theta = \mathbb{E}(S/a_1(n))$ with $S$ segregating sites, meaning that Eq. (8) yields the proportion of variable sites.

$P(m|C, n)$ is the probability of observing $m$ as first allele count in a SNP with $C$ reads from a pool of dimension $n$:

$$P(m|C, n) = \frac{1}{a_1(n)} \sum_{k=1}^{n-1} \frac{1}{k} P(m|C, n, k) \qquad (9)$$

$P(m|C,n,k)$ is the probability of having a first allele count of $m$ observed in $C$ reads that were taken from a pool of $n$ individuals with first allele count of $k$. That is, $m$ is the allele count in the reads, and $k$ is the allele count in the pool:

$$P(m|C,n,k) = \binom{C}{m} \left(\frac{k}{n}\right)^{m} \left(\frac{n-k}{n}\right)^{C-m} \tag{10}$$

In words, $P(m|C,n,k)$ follows a binomial distribution, with $m$ successes out of $C$ trials with a success probability of $k/n$ for each trial. That is, we compute how likely it is to observe $m$ as the first allele count in $C$ reads, given the frequency $k/n$ of the first allele in the pool. Again, the count of the second allele is implicitly used here as $C - m$.

Starting from Eq. (7), we can now put this together:

$$\begin{aligned} \mathbb{E}(\theta_\pi|C,n) &= P(\text{SNP}|n) \cdot \sum_{m=b}^{C-b} \theta_\pi(m,C) \cdot P(m|C,n) \\ &= \theta a_1(n) \cdot \sum_{m=b}^{C-b} \frac{2u(C-m)}{C(C-1)} \cdot \frac{1}{a_1(n)} \sum_{k=1}^{n-1} \frac{1}{k} \binom{C}{m} \left(\frac{k}{n}\right)^{m} \left(\frac{n-k}{n}\right)^{C-m} \end{aligned} \tag{11}$$

#### Final estimate of Theta Pi

We can now solve this for $\theta$ to define our final corrected estimate $\theta_{\pi,\text{pool}}$. Note that the $a_1$ terms cancel out.

$$\theta \approx \frac{\mathbb{E}(\theta_\pi|C,n)}{a_1(n) \cdot \sum_{m=b}^{C-b} \theta_\pi(m,C) \cdot P(m|C,n)} \tag{12}$$

This only leaves the $\mathbb{E}(\theta_\pi|C,n)$ term unresolved, which we however can estimate from our data using the classic estimator as shown in Eq. (6); note however that this is only evaluated on SNPs that have at least a count of $b$. TODO: This step feels weird and I have not fully understood why doing this is reasonable... In total, this yields:

$$\theta_{\pi,\text{pool}}(u,C,n) := \frac{\frac{2u(C-u)}{C(C-1)}}{\sum_{m=b}^{C-b} \frac{2m(C-m)}{C(C-1)} \cdot \sum_{k=1}^{n-1} \frac{1}{k} \binom{C}{m} \left(\frac{k}{n}\right)^{m} \left(\frac{n-k}{n}\right)^{C-m}} \tag{13}$$

Note that the denominator only depends on the total coverage $C$ and the pool size $n$, and hence only needs to be computed once per coverage level, yielding a significant computational speedup.

### 2.2 Simplified Theta Pi

Not sure if needed.

PoPoolation notes: also good for individual sequencing

## 3 Theta Watterson

### 3.1 Pool-Sequencing Correction

For Watteron's estimator $\theta_w$, we follow the same approach as above. In order to derive the pool-sequencing corrected equations, we first define $\theta_w$ as usual:

$$\theta_w(u,C) = \frac{S(u)}{\sum_{k=1}^{C-1} 1/k} \tag{14}$$

where classically, $S$ is the number of segregating sites, see for example Equation (3.5) of Hahn (2018) [2]. We are here working with a biallelic SNP at a single site, which as before we only want to consider if its count is within the limits of the minimum allele count $b$, and so we define:

$$S(u) = \begin{cases} 1 & \text{if } b \le u \le C-b \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

Reasoning the same as above, we get the expected value of $\theta_w$ as:

$$\mathbb{E}(\theta_w|C,n) = P(\mathrm{SNP}|n) \cdot \frac{\sum_{m=b}^{C-b} P(m|C,n)}{\sum_{k=1}^{C-1} 1/k}$$

with the two probability terms again as in Eq. (8) and Eq. (9). For conciseness, we here only resolve $P(\mathrm{SNP}|n)$:

$$= \theta a_1(n) \cdot \frac{\sum_{m=b}^{C-b} P(m|C,n)}{\sum_{k=1}^{C-1} 1/k} \tag{16}$$

We can again solve this for $\theta$, to get our corrected estimate:

$$\theta \approx \mathbb{E}(\theta_w|C,n) \cdot \frac{\sum_{k=1}^{C-1} 1/k}{a_1(n) \cdot \sum_{m=b}^{C-b} P(m|C,n)} \tag{17}$$

Again using the classic value of Eq. (14) for the expected value, we can now define our estimate:

$$\theta_{w,\mathrm{pool}}(u,C,n) := \frac{S(u)}{a_1(n) \cdot \sum_{m=b}^{C-b} P(m|C,n)} \tag{18}$$

The summation over $1/k$ cancel out here. As before, the denominator only depends on the coverage $C$, and hence only needs to be computed once per coverage level that is present in the data.

### 3.2 Simplified Theta Watteron

Not sure if needed.

PoPoolation notes: also good for individual sequencing

## 4 Tajima's D

Above, we have defined pool-sequencing corrected estimators $\theta_\pi$ and $\theta_w$. Now, we want to use them to define a test akin to Tajima's D for pool sequencing.

### 4.1 Pool-Sequencing Correction

First, we define:

$$d_{\mathrm{pool}}(u,C) \;:=\; \theta_{\pi,\mathrm{pool}}(u,C) \;-\; \theta_{w,\mathrm{pool}}(u,C) \tag{19}$$

and use this to define our statistic:

$$D_{\mathrm{pool}}(u,C) \;:=\; \frac{d_{\mathrm{pool}}(u,C)}{\sqrt{\mathrm{Var}(d_{\mathrm{pool}}(u,C))}} \tag{20}$$

In order to compute the variance of $d_{\mathrm{pool}}$ (leaving out function arguments for simplicty), we start with the standard expansion of the variance:

$$\mathrm{Var}(d_{\mathrm{pool}}) = \mathbb{E}(d_{\mathrm{pool}}^2) - \mathbb{E}(d_{\mathrm{pool}})^2$$

At this point, we use that $d_{\mathrm{pool}}$ is unbiased, and hence has an expected value of 0, that is, $\mathbb{E}(d_{\mathrm{pool}})^2 = 0$. PoPoolation notes that this is only true if they did their previous calculations correctly, but we trust they did.

Then, we can compute the variance as:

$$\begin{aligned}\mathrm{Var}(d_{\mathrm{pool}}) &= \mathbb{E}(d_{\mathrm{pool}}^2) \\ &= P(\mathrm{SNP}|n) \cdot \sum_{m=b}^{C-b} d_{\mathrm{pool}}^2(m,C) \cdot P(m|C,n)\end{aligned}$$

which can be resolved using equations Eq. (8) and Eq. (9) from previous sections:

$$= \theta \cdot \sum_{m=b}^{C-b} \left(\theta_{\pi,\text{pool}}(m,C) - \theta_{w,\text{pool}}(m,C)\right)^2 \cdot \sum_{k=1}^{n-1} \frac{1}{k} \binom{C}{m} \left(\frac{k}{n}\right)^m \left(\frac{n-k}{n}\right)^{C-m} \tag{21}$$

This leaves $\theta$ to be estimated. PoPoolation suggests to estimate it as $\theta_{\pi,\text{pool}}$ on the same window on which we are computing $D_{\text{pool}}$. This assumes that all individuals contribute the same number of reads to the pool.

The first summation in Eq. (21) involves computing $\theta_{\pi,\text{pool}}$ and $\theta_{w,\text{pool}}$ repeatedly $C - 2b$ many times, with each of these computations involving to compute their respective denominators, as shown in Eq. (13) and Eq. (18). However, as $C$ remains constant throughout this computation, these denominators (the correction terms) are identical, so that we only need to compute them once, to gain a $\approx C$-fold speedup.

TODO: I am not sure that this whole section is implemented at all. Maybe it is in PoPoolation, but not in grenedalf.

## 4.2 Computation in Windows

TODO: not sure if needed

## 4.3 Integration with Classic Tajima's D

On large windows, the classic Tajima's D is not a measure of significance (in number of standard deviations away from the null hypothesis), but instead is a measure of the magnitude of the divergence from neutrality. This is because all loci are considered completely linked, even if they are not in reality.

However, our pool-sequencing Tajima's D instead consideres all loci as completely unlinked, and thus represents the number of standard deviations away from neutrality. Therefore, it gives a different numerical result that has a much higher absolute value compared to classic Tajima's D.

Now, we want to obtain a correction term for the pool-sequence Tajima's D to obtain values that are comparable to classic Tajima's D in non-small windows, that is, we want a measure of the magnitude of the divergence from neutrality.

### Approach by Achaz

To this end, we use a modified version of the $Y^*$ test of Achaz (2008) [1], which was originally developed as a test for neutrality despite the presence of sequencing errors. This test only works when excluding singletons, that is, we set $b := 2$ from here on.

Following PoPoolation and Achaz (2008) [1], we first define:

$$f^*(n) = \frac{n-3}{a_1(n) \cdot (n-1) - n} \tag{22}$$

which is then used to define:

$$\alpha^*(n) = f^{*2} \cdot \left(a_n - \frac{n}{n-1}\right) + f^* \cdot \left(a_n \cdot \frac{4(n+1)}{(n-1)^2} - 2 \cdot \frac{n+3}{n-1}\right) - a_n \cdot \frac{8(n+1)}{n(n-1)^2} + \frac{n^2+n+60}{3n(n-1)} \tag{23}$$

and:

$$\begin{aligned}\beta^*(n) = {} & f^{*2} \cdot \left(b_n - \frac{2n-1}{(n-1)^2}\right) + f^* \cdot \left(b_n \cdot \frac{8}{n-1} - a_n \cdot \frac{4}{n(n-1)} - \frac{n^3+12n^2-35n+18}{n(n-1)^2}\right) \\ & - b_n \cdot \frac{16}{n(n-1)} + a_n \cdot \frac{8}{n^2(n-1)} + \frac{2(n^4+110n^2-255n+126)}{9n^2(n-1)^2}\end{aligned} \tag{24}$$

Note that these equations were originally developed for data from individuals, and hence here, $n$ denotes the number of individuals *as if* we were doing individual sequencing.

NB: The PoPoolation document recommends to counter-check the correctness of their equation with the original of Achaz (2008) [1], which we did. In fact, PoPoolation introduced a mistake in the last term of $\beta^*$, which we have fixed here. Above is the (hopefully) correct one, following Achaz (2008) [1]. Note that the mistake only concerns the PoPoolation equations document, but not their implementation.

### The number of individuals sequenced

The only unresolved parameter is $n$, which corresponds to the number of individuals sequenced – if we were to do individual sequencing. In our case of pool sequencing, according to PoPoolation, we can reasonably substitute this with the expected number of distinct individuals sequenced.

To this end, we use the coverage $C$, as well as the pool size $n_p$, which is not to be confused with the previous $n$. Then, we define:

$$\tilde{n}_{\text{base}} = \sum_{k=1}^{T} \sum_{j=1}^{k} (-1)^{k-j} \cdot k \binom{n_p}{k} \binom{k}{j} \left( \frac{j}{n_p} \right)^C \tag{25}$$

where $T = \max(C, n_p)$; if $n_p$ is much larger than $C$, we can assume $n_p \approx C$. Our substitute $\tilde{n}$ is then obtained by averaging $\tilde{n}_{\text{base}}$ over the window $W$.

Computing the expected number of distinct individuals sequenced corresponds to the following statistical question: Given a set of integers $A = \{1, \dots, n_p\}$ (corresponding to individuals), pick a set $B$ of $C$ elements from set $A$ with replacement (corresponding to reads); what is the expected number of distinct values (individuals) that have been picked in $B$ (that we have reads from)?

PoPoolation computes this value by brute force using Eq. (25), that is, by trying all possible ways to pick numbers from the set. However, there exists a closed form solution to this question, which yields massive speedups for larger coverages.

One way to arrive at the closed form expression is as follows: Define an indicator random variable $I_i$ for $1 \leq i \leq n_p$ as 1 if individual $i$ is present in the set $B$ (that is, if individual $i$ has been sequenced), and as 0 if not. Then, the size of set $B$ is simply $\sum_{i=1}^{n_p} I_i$.

The probability that $I_i$ equals 1 (that is, that individual $i$ has been sequenced) for any $i$ is given by:

$$P(I_i = 1) = 1 - \left( \frac{n_p - 1}{n_p} \right)^C \tag{26}$$

In words, this is the complement of *not* picking $i$ in all of the $C$ picks from set $A$.

The expected size of the set $B$ can then be computed by linearity of expectation, yielding our closed form expression:

$$\tilde{n}_{\text{base}} = n_p \left( 1 - \left( \frac{n_p - 1}{n_p} \right)^C \right) \tag{27}$$

### Final estimator for D

Now that we have a way of computing a reasonable value for the number of individuals sequenced, we can finally define the estimator:

$$\tilde{D}_{\text{pool}} = \frac{\theta_\pi - \theta_w}{\sqrt{|W|^{-1} \cdot \alpha^*(\tilde{n}) \cdot \theta \ + \ \beta^*(\tilde{n}) \cdot \theta^2}} \tag{28}$$

TODO: theta b pool W?! following PoPoolation and Achaz (2008) [1]. This requires $b = 2$; furthermore, PoPoolation suggests to use “not too small” windows.

## 4.4 Simplified Tajima’s D

TODO: not sure what this is needed for at the moment...

### 4.5 PoPoolation Bugs

In the implementation of the above $\tilde{D}_{\text{pool}}$, there are two bugs in PoPoolation $\leq$ v1.2.2, which alter the numerical results of the computation of Tajima's D. Firstly, they compute $\tilde{n}$ not by using the pool size $n_p$ and the coverage $C$, but by using the pool size $n_p$ for both arguments. Secondly, instead of computing $\alpha^*$ and $\beta^*$, they only compute $\beta^*$, and use this as the value for $\alpha^*$ as well. We have examined the effect of these bugs, and present results of the numerical changes induced by them in TODO: Supplementary document X.

## 5 Fixation Index $\mathsf{F_{ST}}$

### 5.1 Large Regions

### 5.2 Simplified Regional $\mathsf{F_{ST}}$

### 5.3 Single SNPs

### 5.4 Weights

## References

[1] Achaz, G. (2008). Testing for neutrality in samples with sequencing errors. *Genetics*, **179**(3), 1409–1424.

[2] Hahn, M. W. (2018). *Molecular Population Genetics*.

[3] Kofler, R. *et al.* (2011a). PoPoolation: A Toolbox for Population Genetic Analysis of Next Generation Sequencing Data from Pooled Individuals. *PLoS ONE*, **6**(1), e15925.

[4] Kofler, R. *et al.* (2011b). PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics*, **27**(24), 3435–3436.