# Tajima's Pi formulas

Let's suppose we have a minimum allele count of $b$, a pool of size $n$, and an observed coverage of $C$.
Our estimate of $\theta$ is $\pi_b$:

$$\pi_b(i_A, i_C, i_G, i_T) = \frac{C}{C-1}(1 - f_A^2 - f_C^2 - f_G^2 - f_T^2) = \tag{1}$$

$$= \frac{C}{C-1}(1 - \frac{i_A^2}{C^2} - \frac{i_C^2}{C^2} - \frac{i_G^2}{C^2} - \frac{i_T^2}{C^2}) =$$

$$= \frac{C}{C-1} - \frac{i_A^2}{C(C-1)} - \frac{i_C^2}{C(C-1)} - \frac{i_G^2}{C(C-1)} - \frac{i_T^2}{C(C-1)} =$$

$$= \frac{C^2}{C(C-1)} - \frac{i_A^2}{C(C-1)} - \frac{i_C^2}{C(C-1)} - \frac{i_G^2}{C(C-1)} - \frac{i_T^2}{C(C-1)} =$$

$$= \frac{C(C-1)}{C(C-1)} - \frac{i_A(i_A-1)}{C(C-1)} - \frac{i_C(i_C-1)}{C(C-1)} - \frac{i_G(i_G-1)}{C(C-1)} - \frac{i_T(i_T-1)}{C(C-1)} =$$

$$= 1 - \frac{i_A(i_A-1)}{C(C-1)} - \frac{i_C(i_C-1)}{C(C-1)} - \frac{i_G(i_G-1)}{C(C-1)} - \frac{i_T(i_T-1)}{C(C-1)}$$

(we have used that $C = i_A + i_C + i_G + i_T$)
Or in case of a simple SNP with allele count $i$:

$$\pi_b(i) = \frac{2i(C-i)}{C(C-1)}$$

Where $i_A$ is the count of allele A, $f_A$ is $i_A/C$ , and so on.
What is the expectation of $\pi_b$ conditioned by $C$ ? (we don't expect complex SNPs)

$$E(\pi_b|C) = P(SNP|n) \sum_{m=b}^{C-b} \frac{2m(C-m)}{C(C-1)} P(m|C,n))$$

$P(SNP|n)$ is the probability of observing a SNP in a pool of size $n$.

$$P(SNP|n) = \theta \sum_{k=1}^{n-1} 1/k$$

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

$$P(m|C,n)) = \sum_{k=1}^{n-1} P(m|C,n,k) \frac{1/k}{\sum_{j=1}^{n-1} 1/j}$$

$P(m|C,n,i)$ is the probability of having a first allele count of $m$ in $C$ reads from a pool of $n$ with first allele count of $i$ ($m$ is the allele count in the reads, $i$ is the allele count in the pool)

$$P(m|C,n,k) = \binom{C}{m} (\frac{k}{n})^m (\frac{n-k}{n})^{C-m} \tag{2}$$

We can re-write

$$E(\pi_b|C) = \theta \sum_{k=1}^{n-1} 1/k \sum_{m=b}^{C-b} \frac{2m(C-m)}{C(C-1)} P(m|C,n))$$

And from this we obtain our corrected estimate of $\theta$

$$\theta \approx \frac{E[\pi_b]}{\sum_{k=1}^{n-1} 1/k \sum_{m=b}^{C-b} \frac{2m(C-m)}{C(C-1)} P(m|C,n))}$$

So we define:

$$\pi_{b,pool}(i) = \frac{\pi_b(i)}{\sum_{k=1}^{n-1} 1/k \sum_{m=b}^{C-b} \frac{2m(C-m)}{C(C-1)} P(m|C,n))} =$$

$$= \frac{i(C-i)}{\sum_{m=b}^{C-b} m(C-m) \sum_{k=1}^{n-1} P(m|C,n,k) 1/k}$$

The only term you have to look for is in equation (2).
Note that the denominator of the last fraction could be calculated only once per each coverage level, and than be reused for different allele frequencies.

## simplified $\pi$ (also good for individual sequencing)

The simplified version of the formula, which also works for individual sequencing with error, is obtained supposing (and substituting in the formula)

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

$$P(SNP|n) = \theta \sum_{k=1}^{C-1} 1/k$$

This corresponds to supposing that allele frequency distribution in the reads is about the same as in the real population.
After doing this substitution many terms cancel out and what remains is

$$\theta \approx \frac{C-1}{C-2b+1} E[\pi_b]$$

And so we define:

$$\pi_{b,simp}(i) = \frac{C-1}{C-2b+1}\pi_b(i) = \frac{2i(C-i)}{C(C-2b+1)}$$

In general for any SNP:

$$\pi_{b,simp}(i_{A,i_C,i_G,i_T}) = \frac{C-1}{C-2b+1}\pi_b(i_A, i_C, i_G, i_T) \tag{3}$$

These hypothesis hold for example when the coverage is small with respect to the pool size.

# Watterson theta formulas

We are in the same situation as before.
This time we define our estimate

$$\theta_{w_b}(i) = \frac{S_b(i)}{\sum_{k=1}^{C-1} 1/k}$$

Where $S_b(i)$ is 1 if $b \leq i \leq C-b$, otherwise is 0. Reasoning exactly in the same way as we did for $\pi$:

$$E(\theta_{w_b}|C) = \frac{P(SNP|n)\sum_{m=b}^{C-b} P(m|C,n))}{\sum_{k=1}^{C-1} 1/k}$$

Which brings to the corrected estimate:

$$\theta \approx \frac{E[\theta_{w_b}]\sum_{k=1}^{C-1} 1/k}{\sum_{k=1}^{n-1} 1/k \sum_{m=b}^{C-b} P(m|C,n))}$$

So we can define:

$$\theta_{w_{b,pool}}(i) = \frac{S_b(i)}{\sum_{k=1}^{n-1} 1/k \sum_{m=b}^{C-b} P(m|C,n))} =$$
$$= \frac{S_b(i)}{\sum_{m=b}^{C-b} \sum_{k=1}^{n-1} P(m|C,n,k)1/k} =$$

And the only term you need to substitute is in equation (2).
Again, the denominator of the last equation only depends on coverage, and can be calculated once per coverage level.

### simplified $\theta$ (also for individual sequencing)

Again, the same as for $\pi$: the same substitutions give

$$\theta \approx \frac{E[S_b]}{\sum_{k=b}^{C-b} \frac{1}{k}} = \frac{\sum_{k=1}^{C-1} \frac{1}{k}}{\sum_{k=b}^{C-b} \frac{1}{k}} E[\theta_{w_b}]$$

Which brings us to define:

$$\theta_{w_{b,simp}}(i) = \frac{S_b(i)}{\sum_{k=b}^{C-b} \frac{1}{k}}$$

## Tajima's D

So, we have our new $\pi$ and $\theta_w$-like parameters for pooled NGS data and we want to use them for a new Tajima's D-like test.
We than define:

$$d_{b,pool}(i) = \pi_{b,pool}(i) - \theta_{w_{b,pool}}(i)$$

From this we define the final Tajima's D-like parameter:

$$D_{b,pool}(i) = \frac{d_{b,pool}(i)}{\sqrt{Var(d_{b,pool})}}$$

So, the only problem here is actually the variance of $d_{b,pool}$.
But if we did our previous calculations correctly, $d_{b,pool}$ is unbiased, then has expected value 0.
in this case:

$$Var(d_{b,pool}) = E[d_{b,pool}^2]$$

So we now show how to calculate $E[d^2_{b,pool}]$.
We have then:

$$E[d^2_{b,pool}] = P(SNP|n)\sum_{m=b}^{C-b}(\pi_{b,pool}(m) - \theta_{w_{b,pool}}(m))^2 P(m|C,n)) =$$

$$= \theta \sum_{m=b}^{C-b}(\pi_{b,pool}(m) - \theta_{w_{b,pool}}(m))^2 \sum_{k=1}^{n-1} P(m|C,n,k)1/k$$

And you can find all the quantities used in the first section.
Note 1:
$\theta$ in the last equation needs to be estimated someway. I would suggest to estimate it with $\pi_{b,pool}$ on the same window on which $D_{b,pool}$ is estimated.
Note 2:
We don't need to calculate about $C$ times $\pi_{b,pool}$ and $\theta_{w_{b,pool}}$. In fact, while these two depend on $m$, their correction terms do not, so we can calculate the corrections once, and use them for each $\pi_{b,pool}(m)$ and $\theta_{w_{b,pool}}(m)$. This makes the calculations about $C$ times faster!!! (and comparable to the calculation times of the other quantities).

# Tajima's D: how do I deal with windows?

Let's calculate the mean Watterson's $\theta$ for a window $W$ (sum of $\theta$s divided by $|W|$) and let's call it $\theta_{b,pool_W}$.
For each SNP with MAF $i$ let's calculate $D_{b,pool}(i)$ using $\theta = \theta_{b,pool_W}$ in the variance.
This will be our SNP-specific Tajima's D.
Now in order to calculate Tajima's D for the window $W$ we have to account for the fact that the standard deviation of the mean is $\frac{1}{\sqrt{|W|}}$ times the standard deviation of the single SNP (under independent uniformly distributed assumption, this is a good assumption when the pool size is much larger than the coverage):

$$D_{b,pool_W} = \frac{\sqrt{|W|}\sum_{base=1}^{|W|} D_{b,pool}(MAF(base))}{|W|} = \frac{\sum_{base=1}^{|W|} D_{b,pool}(MAF(base))}{\sqrt{|W|}}$$

Clearly the $D_{b,pool}(MAF(base))$ will be zero if $base$ is not a SNP.

# Tajima's D: now I want something comparable to the classic D

When calculating the new Tajima's D we have to consider that on large data the classical Tajima's D is not a measure of significance (in number of standard deviations away from null hypothesis) but is a measure of the magnitude of the divergence from neutrality. This happens because all loci are considered completely linked even if they are not.
The new Tajima's D instead considers all loci as completely unlinked, and thus represents the number of standard deviations away from neutrality, giving a completely different result from the classical Tajima's D (much higher absolute value).
Now we present a method in order to obtain values comparable with the classic D when non-small windows are analyzed, that is, a measure of the magnitude of divergence from neutrality.

The results are obtained modifying the $Y^*$ estimator of Achaz (2008, *Testing for neutrality with sequencing errors*), and as so, it is assumed to work only excluding singletons. But I would also give a try after excluding doubletons.
In his paper Achaz defines:

$$f^* = \frac{n-3}{a_n(n-1)-n}$$

And it is used to define

$$\alpha^* = f^{*2}(a_n - \frac{n}{n-1}) + f^*(a_n\frac{4(n+1)}{(n-1)^2} - 2\frac{n+3}{n-1}) - a_n\frac{8(n+1)}{n(n-1)^2} + \frac{n^2+n+60}{3n(n-1)},$$

and

$$\beta^* = f^{*2}(b_n - \frac{2n-1}{(n-1)^2}) + f^*(b_n\frac{8}{n-1} - a_n\frac{4}{n(n-1)} - \frac{n^3+12n^2-35n+18}{n(n-1)^2}) -$$
$$-b_n\frac{16}{n(n-1)} + a_n\frac{8}{n^2(n-1) + \frac{2(n^4+110n^2-255n+126)}{9n^2(n-1)^2}},$$

(you can counter-check the correctness of the formulas on Achaz's paper, appendix B). Also note we used the notation

$$a_n = \sum_{i=1}^{n-1}\frac{1}{i}$$

$$b_n = \sum_{i=1}^{n-1} \frac{1}{i^2}$$

The only parameter until now is $n$, the number of individuals sequenced in individual sequencing. What does this parameter mean in our data? We could reasonably substitute it with the expected number of distinct individuals sequenced:

$$\tilde{n}_{base} = \sum_{k=1}^{max(C,n_p)} k \frac{\binom{n_p}{k} \sum_{j=1}^{k} (-1)^{k-j} \binom{k}{j} j^C}{n_p^C}$$

And $\tilde{n}$ is obtained averaging $\tilde{n}_b ases$ over the window $W$. Again, $C$ is coverage, $n_p$ is pool size (do not confound with the previous $n$). If $n_p$ is much larger than $C$ than one can assume $\tilde{n} \approx C$.

Now comes the new formula: following Achaz we define

$$\tilde{D}_{b,pool_W} = \frac{\pi_{b,pool_W} - \theta_{b,pool_W}}{\sqrt{\frac{\alpha^*_{\tilde{n}}}{|W|}\theta_{b,pool_W} + \beta^*_{\tilde{n}}\theta^2_{b,pool_W}}}$$

This is mostly suggested for not too small windows and with $b = 2$.
In case $b = 1$ the same procedure can be applied to classic Tajima's D instead of Achaz's $Y^*$.

# simplified Tajima's D (also for individual sequencing)

The same assumptions with the other simplified formulas lead us to define:

$$d_{b,simp}(i) = \pi_{b,simp}(i) - \theta_{w_{b,simp}}(i)$$

And

$$D_{b,simp}(i) = \frac{\pi_{b,simp}(i) - \theta_{w_{b,simp}}(i)}{\sqrt{Var(d_{b,simp})}}$$

And the same reasoning as before leads to:

$$Var(d_{b,simp}) = E[d^2_{b,simp}] = \theta \sum_{m=b}^{C-b} (\pi_{b,simp}(m) - \theta_{w_{b,simp}}(m))^2 \frac{1}{m}$$

same warnings as for the other simplified formulas.

# correcting $F_{ST}$

We want a parameter similar to $F_{ST}$.
Let's suppose we have $J$ subpopulations and $K$ sites. we define $n_1..n_J$ the subpopulation sizes so that $n = \sum_{j=1}^{J} n_j$, and for each site we define $C_1..C_J$ the subpopulation coverages so that $C = \sum_{j=1}^{J} C_j$.
We also call $m_1, ..., m_J$ and $m = m_1 + ... + m_J$ the number of reads for the minor allele (minor with respect to the total population).

## Strategy for large regions

We fix $b_1..b_J$ and $b$ with the usual meaning, this time $b$ and $b_j$ are not necessarily correlated.
We define our parameter over a certain window as:

$$F_{ST_{pool}} = 1 - \frac{\pi_{b_1,pool} + .... + \pi_{b_J,pool}}{J * \pi_{b,split,pool}}$$

Where $\pi_{b_1,pool}$ is the average of $\pi_{b_1,pool}(.)$ over the $K$ loci, and so on.
Yet we stll have to define $\pi_{b,split,pool}$. We first define a statistic for data from many pools, and then correct for its bias as usual.
Given a certain system of weights $w_1, ..., w_J$ we could define our statistic :

$$\pi_{b,split} = f(w_1, C_1, n_1, m_1 ... w_J, C_J, n_J, m_J)$$

for some function $f$ similar to Tajima's $\pi$, for example one option could be:

$$\pi_{b,split} = 2p(1-p)$$

with

$$p = \frac{m_1/C_1 + ... + m_J/C_J}{J}$$

Or more generally using the weights:

$$p = \frac{\frac{w_1 * m_1}{C_1} + ... + \frac{w_J * m_J}{C_J}}{w}$$

In any case $\pi_{b,split} = 0$ when $m < b$.
We will retain the notation with $f$ in the following for simplicity.

We have in the case J=2:

$$E(\pi_{b,split}) = P(SNP|n)\sum_{k=1}^{n-1}\frac{1/k}{\sum_{j=1}^{j=n-1}1/j}\sum_{k_1=k-min(n_2,k)}^{min(k,n_1)}\frac{\binom{k_1}{k}\binom{n_1-k_1}{n-k}}{\binom{n_1}{n}}*$$

$$*\sum_{m_1=0}^{C_1}P(m_1|C_1,n_1,k_1)\sum_{m_2=max(b-m_1,0)}^{C_2-max(0,b-C_1+m_1)}P(m_2|C2,n_2,k_2)\pi_{b,split}(n_1,m_1,C_1,n_2,m_2,C_2)$$

So we can finally define:

$$\pi_{b,split,pool}(m_1,m_2) = \pi_{b,split}(m_1,m_2)/\left[\sum_{k=1}^{n-1}\frac{1}{k}\sum_{k_1=k-min(n_2,k)}^{min(k,n_1)}\frac{\binom{k_1}{k}\binom{n_1-k_1}{n-k}}{\binom{n_1}{n}}*\right.$$

$$\left.*\sum_{m_1=0}^{C_1}P(m_1|C_1,n_1,k_1)\sum_{m_2=max(b-m_1,0)}^{C_2-max(0,b-C_1+m_1)}P(m_2|C2,n_2,k_2)\pi_{b,split}(n_1,m_1,C_1,n_2,m_2,C_2)\right] \quad (4)$$

## Simplified regional Fst (individual sequencing)

We can use the simplifying assumptions used for equation (3) to get a simplified version of equation (4). Same warnings as usual.

$$\pi_{b,split,simp}(m_1,m_2) = \pi_{b,split}(m_1,m_2)/\left[\sum_{m=b}^{C-b}\frac{1}{m}\sum_{m_1=m-min(C_2,m)}^{min(m,C_1)}\frac{\binom{m_1}{m}\binom{C_1-m_1}{C-m}}{\binom{C_1}{C}}\pi_{b,split}(m_1,m_2)\right] \quad (5)$$

## Fst over single SNPs

When working with a single SNP we don't have to account for the missing singletons (and so on) any more, and we don't want to exclude data from any population. A solution to this problem would be using previous section formulae but setting $b_1 = ... = b_J = b = 1$, but this also doesn't work perfectly.
An alternative (more complex) proceeding, which accounts for the small amount of data, is calculating the expected value of Tajima's $\pi$ given the data of a SNP:

$$\pi_{site}(m) := E(\pi|C,n,m) = \sum_{k=1}^{n-1}P(k|C,n,m)\frac{2k(n-k)}{n(n-1)}$$

where, using Bayes theorem, we can substitute

$$P(k|C,n,m) = \frac{P(m|k,n,C)\frac{1/k}{\sum_{j=1}^{n-1} 1/j}}{\sum_{z=1}^{n-1} P(m|z,n,C)\frac{1/z}{\sum_{j=1}^{n-1} 1/j}} .$$

We plan to use this strategy for the $F_{ST}$ test, so we define the same parameter for the total population:

$$\pi_{site,split}(m_1,m_2) := E(\pi|C_1,C_2,n_1,n_2,m_1,m_2) = \sum_{k=1}^{n-1} \sum_{k_1=k-min(n_2,k)}^{min(n_1,k)} P(k_1,k_2|...)\pi_{split}(k_1,k_2,w_1,w_2)$$

and using Bayes theorem again:

$$P(k_1,k_2|...) = \frac{P(m_1,m_2|k_1,k_2,...)\frac{1/k}{\sum_{j=1}^{n-1} 1/j}\frac{\binom{k_1}{k}\binom{n_1-k_1}{n-k}}{\binom{n_1}{n}}}{\sum_{z=1}^{n-1}\sum_{z_1=z-min(n_2,z)}^{min(n_1,z)} P(m_1,m_2|z_1,z_2,...)\frac{1/z}{\sum_{j=1}^{n-1} 1/j}\frac{\binom{z_1}{z}\binom{n_1-z_1}{n-z}}{\binom{n_1}{n}}}$$

where $z = z_1 + z_2$ and $k = k_1 + k_2$ as usual, then also:

$$P(m_1,m_2|k_1,k_2,n_1,n_2,C_1,C_2) = P(m_1|C_1,n_1,k_1)P(m_2|C_2,n_2,k_2) .$$

We then define:

$$F_{ST_{site}} = 1 - \frac{\pi_{site}(m_1) + \pi_{site}(m_2)}{2 * \pi_{site,split}(m_1,m_2)} .$$

## Weights

Now it would be nice to introduce some weights for the populations, so that populations with less sequenced individual have less weight in the $F_{ST}$ value.
We have already introduced $w_1, ...w_J$, if we also define $a_1...a_J$ and $a = a_1+...+a_J$ we can modify $F_{ST}$ in this way:

$$F_{ST_{pool}} = 1 - \frac{a_1\pi_{b_1,pool} + .... + a_J\pi_{b_J,pool}}{a\pi_{b,pool}}$$

How to to define the $a_j$s?
One simple way would be $a_j = min(n_j, C_j)$ or $a_j = min(n_j, C_j) * [min(n_j, C_j) -$

$1]/2$.
Similarly we can define $w_j = min(n_j, C_j)$.
But maybe we can do better.
Let's call $I_j$ the number of different individuals actually sequenced in subpopulation $j$. Then it would be nice to define $w_j = a_j = E[I_j]$, or also $a_j = E[I_j] * (E[I_j] - 1)/2$ or even better $a_j = E[I_j * (I_j - 1)/2]$.
One way to achieve this is using this recursive equation:

$$P(I_j = i|n_j, C_j) = \sum_{m_i=1}^{C_j-i+1} \frac{n_j - i + 1}{n_j}(i/n_j)^{m_i-1} P(I_j = i - 1|n_j, C_j - m_i)$$

with starting values:

$$P(I_j = 1|n_j, C_j) = (1/n_j)^{C_j-1}$$

This recursive approach would fit well a dynamic programming algorithm.
Is there a simpler way?
sincerely I don't know, it is worth consulting a combinatorics book since the problem is quite general.
Last problem: the coverage is usually not constant over a window. One solution would be to calculate the average coverage for the window for each population, and use this in the equations.
The second more precise solution would be to calculate $a_j$ for each position on the window, and then to take the average.

contact me for questions!
Nicola De Maio