Measuring genetic differentiation from Pool-seq data

Valentin Hivert, Raphaël Leblois, Eric J. Petit, Mathieu Gautier and Renaud Vitalis

SUPPLEMENTAL FILE S1: DETAILED MATHEMATICAL DERIVATIONS

Analysis of variance for Pool-seq data

In the following, we first derive our model for a single locus. Consider a sample of n_d subpopulations, each of which is made of n_i genes ($i = 1, \ldots, n_\mathrm{d}$) sequenced in pools (hence n_i is the haploid sample size of the ith pool). We define c_{ij} as the number of reads sequenced from gene j ($j = 1, \ldots, n_\mathrm{i}$) in subpopulation i at the locus considered. Note that c_{ij} is a latent variable, that cannot be directly observed from the data. Let $X_{ijr:k}$ be an indicator variable for read r ($r = 1, \ldots, c_{ij}$) from gene j in subpopulation i, such that $X_{ijr:k} = 1$ if the rth read from the jth gene in the ith deme is of type k, and $X_{ijr:k} = 0$ otherwise. In the following, we use standard dot notations for sample averages, i.e.: $X_{ij\cdot:k} \equiv \sum_r X_{ijr:k}/c_{ij}$, $X_{i\cdots:k} \equiv \sum_j \sum_r X_{ijr:k} / \sum_j c_{ij}$ and $X_{\cdots:k} \equiv \sum_i \sum_j \sum_r X_{ijr:k} / \sum_i \sum_j c_{ij}$. The analysis of variance is based on the computation of sums of squares, as follows:

$$\begin{aligned}
\sum_i^{n_\mathrm{d}} \sum_j^{n_i} \sum_r^{c_{ij}} (X_{ijr:k} - X_{\cdots:k})^2 &= \sum_i^{n_\mathrm{d}} \sum_j^{n_i} \sum_r^{c_{ij}} (X_{ijr:k} - X_{ij\cdot:k})^2 \\
&+ \sum_i^{n_\mathrm{d}} \sum_j^{n_i} \sum_r^{c_{ij}} (X_{ij\cdot:k} - X_{i\cdots:k})^2 \\
&+ \sum_i^{n_\mathrm{d}} \sum_j^{n_i} \sum_r^{c_{ij}} (X_{i\cdots:k} - X_{\cdots:k})^2 \\
&\equiv SSR_{:k} + SSI_{:k} + SSP_{:k} \qquad \text{(A1)}
\end{aligned}$$

We express the sum of squares for reads within individuals as:

$$\begin{aligned} SSR_{:k} &= \sum_{i}^{n_{\mathrm{d}}} \sum_{j}^{n_i} \sum_{r}^{c_{ij}} \left(X_{ijr:k} - X_{ij\cdot:k}\right)^2 \\ &= 0 \end{aligned} \tag{A2}$$

since we assume that there is no sequencing error, i.e. all the reads sequenced from a single gene are identical (therefore $X_{ijr:k} = X_{ij\cdot:k}$, for all r). The sum of squares for genes within pools reads:

$$\begin{aligned} SSI_{:k} &= \sum_{i}^{n_{\mathrm{d}}} \sum_{j}^{n_i} \sum_{r}^{c_{ij}} \left(X_{ij\cdot:k} - X_{i\cdots:k}\right)^2 \\ &= \sum_{i}^{n_{\mathrm{d}}} \sum_{j}^{n_i} \sum_{r}^{c_{ij}} \left(X_{ij\cdot:k} - \pi_k\right)^2 - \sum_{i}^{n_{\mathrm{d}}} \sum_{j}^{n_i} \sum_{r}^{c_{ij}} \left(X_{i\cdots:k} - \pi_k\right)^2 \\ &= \sum_{i}^{n_{\mathrm{d}}} \sum_{j}^{n_i} c_{ij} \left(X_{ij\cdot:k} - \pi_k\right)^2 - \sum_{i}^{n_{\mathrm{d}}} C_{1i} \left(X_{i\cdots:k} - \pi_k\right)^2 \end{aligned} \tag{A3}$$

where π_k is the expectation of the frequency of allele k over independent replicates of the evolutionary process, and $C_{1i} \equiv \sum_j c_{ij}$ is the total number of observed reads in the ith pool. Likewise, the sum of squares for genes between pools reads:

$$\begin{aligned} SSP_{:k} &= \sum_{i}^{n_{\mathrm{d}}} \sum_{j}^{n_i} \sum_{r}^{c_{ij}} \left(X_{i\cdots:k} - X_{\cdots:k}\right)^2 \\ &= \sum_{i}^{n_{\mathrm{d}}} C_{1i} \left(X_{i\cdots:k} - \pi_k\right)^2 - C_1 \left(X_{\cdots:k} - \pi_k\right)^2 \end{aligned} \tag{A4}$$

where $C_1 \equiv \sum_i \sum_j c_{ij} = \sum_i C_{1i}$ is the total number of observed reads in the full sample. These sums can be expressed as functions of the average frequency of reads of type k for individual j: $\hat{\pi}_{ij:k} \equiv X_{ij\cdot:k}$, of the average fre-

quency of reads of type k within the ith pool: $\hat{\pi}_{i:k} \equiv X_{i\cdot:k}$, and of the average frequency of reads of type k in the full sample: $\hat{\pi}_k \equiv X_{\cdots:k}$. Note that from the definition of $X_{\cdots:k}$, $\hat{\pi}_k \equiv \sum_i \sum_j \sum_r X_{ijr:k} / \sum_i \sum_j c_{ij} = \sum_i C_{1i} \hat{\pi}_{i:k} / \sum_i C_{1i}$ is the weighted average of the sample frequencies with weights equal to the pool coverage. Our approach is therefore equivalent to the weighted analysis-of-variance in Cockerham (1973) (see also Weir and Cockerham 1984; Weir 1996; Weir and Hill 2002; Rousset 2007; Weir and Goudet 2017). Then, developing the square in the first term in the right-hand side of Equation A3, we get:

$$\begin{aligned}
(X_{ij\cdot:k} - \pi_k)^2 &= \left(\frac{\sum_r^{c_{ij}} (X_{ijr:k} - \pi_k)}{c_{ij}} \right)^2 \\
&= \frac{1}{c_{ij}^2} \left(\sum_r^{c_{ij}} X_{ijr:k} - c_{ij}\pi_k \right)^2 \\
&= \frac{1}{c_{ij}^2} \left(\sum_r^{c_{ij}} X_{ijr:k}^2 + \sum_{r \neq r'}^{c_{ij}} X_{ijr:k} X_{ijr':k} - 2c_{ij}^2 X_{ij\cdot:k}\pi_k + c_{ij}^2 \pi_k^2 \right) \\
&= \frac{1}{c_{ij}^2} \left(c_{ij} X_{ij\cdot:k} + c_{ij}(c_{ij} - 1) X_{ij\cdot:k} \right. \\
&- \left. 2c_{ij}^2 X_{ij\cdot:k}\pi_k + c_{ij}^2 \pi_k^2 \right) \\
&= \hat{\pi}_{ij:k} - 2\pi_k \hat{\pi}_{ij:k} + \pi_k^2 \qquad \text{(A5)}
\end{aligned}$$

The sums of squares also depend on the unobserved frequency of pairs of genes sampled in the ith pool that are both of type k, i.e. the probability of identity in state (IIS) for allele k, for two distinct genes in the ith pool: $\hat{Q}_{1i:k} \equiv \left(\sum_{j \neq j'} \sum_{r,r'} X_{ijr:k} X_{ij'r':k} \right) / \left(C_{1i}^2 - \sum_j c_{ij}^2 \right)$. Then, developing the

square in the second term in the right-hand side of Equation A3, we get:

$$
\begin{aligned}
(X_{i\cdot\cdot:k} - \pi_k)^2 &= \left(\frac{\sum_j^{n_i} \sum_r^{c_{ij}} (X_{ijr:k} - \pi_k)}{C_{1i}} \right)^2 \\
&= \frac{1}{{C_{1i}}^2} \left(\sum_j^{n_i} \sum_r^{c_{ij}} X_{ijr:k} - C_{1i}\pi_k \right)^2 \\
&= \frac{1}{{C_{1i}}^2} \left(\sum_j^{n_i} \sum_r^{c_{ij}} X_{ijr:k}^2 + \sum_j^{n_i} \sum_{r \neq r'}^{c_{ij}} X_{ijr:k} X_{ijr':k} \right. \\
&+ \left. \sum_{j \neq j'}^{n_i} \sum_{r,r'}^{c_{ij}} X_{ijr:k} X_{ij'r':k} - 2C_{1i}^2 X_{i\cdot\cdot:k}\pi_k + C_{1i}^2 \pi_k^2 \right) \\
&= \frac{1}{{C_{1i}}^2} \left(\sum_j^{n_i} c_{ij} X_{ij\cdot:k} + \sum_j^{n_i} c_{ij}(c_{ij} - 1) X_{ij\cdot:k} \right. \\
&+ \left. \left(C_{1i}^2 - \sum_j^{n_i} c_{ij}^2 \right) \hat{Q}_{1i:k} - 2C_{1i}^2 X_{i\cdot\cdot:k}\pi_k + C_{1i}^2 \pi_k^2 \right) \\
&= \frac{1}{{C_{1i}}^2} \left(\sum_j^{n_i} c_{ij}^2 \left(X_{ij\cdot:k} - X_{i\cdot\cdot:k} \right) + \left(C_{1i}^2 - \sum_j^{n_i} c_{ij}^2 \right) \left(\hat{Q}_{1i:k} - X_{i\cdot\cdot:k} \right) \right. \\
&+ \left. C_{1i}^2 X_{i\cdot\cdot:k} - 2C_{1i}^2 X_{i\cdot\cdot:k}\pi_k + C_{1i}^2 \pi_k^2 \right) \\
&= \hat{\pi}_{i:k} - 2\pi_k \hat{\pi}_{i:k} + \pi_k^2 + \sum_j^{n_i} \frac{c_{ij}^2}{C_{1i}^2} \left(\hat{\pi}_{ij:k} - \hat{\pi}_{i:k} \right) \\
&+ \left(1 - \sum_j^{n_i} \frac{c_{ij}^2}{C_{1i}^2} \right) \left(\hat{Q}_{1i:k} - \hat{\pi}_{i:k} \right) \qquad \text{(A6)}
\end{aligned}
$$

Last, the sums of squares depend on the unobserved frequency of pairs of genes sampled in the same pool that are both of type k, i.e. the IIS probability for allele k for two distinct genes in the same pool: $\hat{Q}_{1:k} \equiv \left(\sum_i \sum_{j \neq j'} \sum_{r,r'} X_{ijr:k} X_{ij'r':k} \right) / \left(C_2 - \sum_i \sum_j c_{ij}^2 \right)$, and of the unobserved frequency of pairs of genes sampled in different pools that are both of type k: $\hat{Q}_{2:k} \equiv \left(\sum_{i \neq i'} \sum_{j,j'} \sum_{r,r'} X_{ijr:k} X_{i'j'r':k} \right) / \left(C_1^2 - C_2 \right)$, where $C_2 \equiv \sum_i C_{1i}^2$.

Developing the second term in the right-hand side of Equation A4, we get:

$$
\begin{aligned}
(X_{\cdots:k} - \pi_k)^2 &= \left(\frac{\sum_i^{n_\mathrm{d}} \sum_j^{n_i} \sum_r^{c_{ij}} (X_{ijr:k} - \pi_k)}{C_1} \right)^2 \\
&= \frac{1}{C_1^2} \left(\sum_i^{n_\mathrm{d}} \sum_j^{n_i} \sum_r^{c_{ij}} X_{ijr:k} - C_1 \pi_k \right)^2 \\
&= \frac{1}{C_1^2} \left(\sum_i^{n_\mathrm{d}} \sum_j^{n_i} \sum_r^{c_{ij}} X_{ijr:k}^2 + \sum_i^{n_\mathrm{d}} \sum_j^{n_i} \sum_{r \neq r'}^{c_{ij}} X_{ijr:k} X_{ijr':k} \right. \\
&+ \sum_i^{n_\mathrm{d}} \sum_{j \neq j'}^{n_i} \sum_{r,r'}^{c_{ij}} X_{ijr:k} X_{ij'r':k} + \sum_{i \neq i'}^{n_\mathrm{d}} \sum_{j,j'}^{n_i} \sum_{r,r'}^{c_{ij}} X_{ijr:k} X_{i'j'r':k} \\
&- \left. 2C_1^2 X_{\cdots:k} \pi_k + C_1^2 \pi_k^2 \right) \\
&= \frac{1}{C_1^2} \left(\sum_i^{n_\mathrm{d}} \sum_j^{n_i} c_{ij} X_{ij\cdot:k} + \sum_i^{n_\mathrm{d}} \sum_j^{n_i} c_{ij}(c_{ij} - 1) X_{ij\cdot:k} \right. \\
&+ \left. \left(C_2 - \sum_i^{n_\mathrm{d}} \sum_j^{n_i} c_{ij}^2 \right) \hat{Q}_{1:k} + \left(C_1^2 - C_2 \right) \hat{Q}_{2:k} - 2C_1^2 X_{\cdots:k} \pi_k + C_1^2 \pi_k^2 \right) \\
&= \frac{1}{C_1^2} \left(\sum_i^{n_\mathrm{d}} \sum_j^{n_i} c_{ij}^2 \left(X_{ij\cdot:k} - X_{\cdots:k} \right) + \left(C_2 - \sum_i^{n_\mathrm{d}} \sum_j^{n_i} c_{ij}^2 \right) \left(\hat{Q}_{1:k} - X_{\cdots:k} \right) \right. \\
&+ \left. \left(C_1^2 - C_2 \right) \left(\hat{Q}_{2:k} - X_{\cdots:k} \right) + C_1^2 X_{\cdots:k} - 2C_1^2 X_{\cdots:k} \pi_k + C_1^2 \pi_k^2 \right) \\
&= \hat{\pi}_k - 2\pi_k \hat{\pi}_k + \pi_k^2 + \sum_i^{n_\mathrm{d}} \sum_j^{n_i} \frac{c_{ij}^2}{C_1^2} \left(\hat{\pi}_{ij:k} - \hat{\pi}_k \right) \\
&+ \left(\frac{C_2}{C_1^2} - \sum_i^{n_\mathrm{d}} \sum_j^{n_i} \frac{c_{ij}^2}{C_1^2} \right) \left(\hat{Q}_{1:k} - \hat{\pi}_k \right) + \left(1 - \frac{C_2}{C_1^2} \right) \left(\hat{Q}_{2:k} - \hat{\pi}_k \right) \qquad \text{(A7)}
\end{aligned}
$$

Hence, developing the first term in the right-hand side of Equation A3 using Equation A5, we have:

$$
\sum_i^{n_\mathrm{d}} \sum_j^{n_i} c_{ij} \left(X_{ij\cdot:k} - \pi_k \right)^2 = C_1 \left(\hat{\pi}_k - 2\pi_k \hat{\pi}_k + \pi_k^2 \right) \qquad \text{(A8)}
$$

Likewise, developing the second term in the right-hand side of Equation A3 using Equation A6, we get:

$$\begin{aligned}\sum_{i}^{n_{\mathrm{d}}} C_{1i}\left(X_{i\cdot:k}-\pi_k\right)^2 &= C_1\left(\hat{\pi}_k-2\pi_k\hat{\pi}_k+\pi_k^2\right)+\sum_{i}^{n_{\mathrm{d}}}\sum_{j}^{n_i}\frac{c_{ij}^2}{C_{1i}}\left(\hat{\pi}_{ij:k}-\hat{\pi}_{i:k}\right)\\ &+ \sum_{i}^{n_{\mathrm{d}}}\left(C_{1i}-\sum_{j}^{n_i}\frac{c_{ij}^2}{C_{1i}}\right)\left(\hat{Q}_{1i:k}-\hat{\pi}_{i:k}\right) \qquad \text{(A9)}\end{aligned}$$

Last, developing the second term in the right-hand side of Equation A4 using Equation A7, we get:

$$\begin{aligned}C_1\left(X_{\cdots:k}-\pi_k\right)^2 &= C_1\left(\hat{\pi}_k-2\pi_k\hat{\pi}_k+\pi_k^2\right)+\sum_{i}^{n_{\mathrm{d}}}\sum_{j}^{n_i}\frac{c_{ij}^2}{C_1}\left(\hat{\pi}_{ij:k}-\hat{\pi}_k\right)\\ &+ \left(\frac{C_2}{C_1}-\sum_{i}^{n_{\mathrm{d}}}\sum_{j}^{n_i}\frac{c_{ij}^2}{C_1}\right)\left(\hat{Q}_{1:k}-\hat{\pi}_k\right)\\ &+ \left(C_1-\frac{C_2}{C_1}\right)\left(\hat{Q}_{2:k}-\hat{\pi}_k\right) \qquad \text{(A10)}\end{aligned}$$

Then, from Equations A3, A8 and A9:

$$\begin{aligned}SSI_{:k} &= \sum_{i}^{n_{\mathrm{d}}}\sum_{j}^{n_i}\frac{c_{ij}^2}{C_{1i}}\left(\hat{\pi}_{i:k}-\hat{\pi}_{ij:k}\right)\\ &+ \sum_{i}^{n_{\mathrm{d}}}\left(C_{1i}-\sum_{j}^{n_i}\frac{c_{ij}^2}{C_{1i}}\right)\left(\hat{\pi}_{i:k}-\hat{Q}_{1i:k}\right) \qquad \text{(A11)}\end{aligned}$$

and from Equations A4, A9 and A10:

$$\begin{aligned}SSP_{:k} &= \sum_{i}^{n_{\mathrm{d}}}\sum_{j}^{n_i}\frac{c_{ij}^2}{C_{1i}}\left(\hat{\pi}_{ij:k}-\hat{\pi}_{i:k}\right)-\sum_{i}^{n_{\mathrm{d}}}\left(C_{1i}-\sum_{j}^{n_i}\frac{c_{ij}^2}{C_{1i}}\right)\left(\hat{\pi}_{i:k}-\hat{Q}_{1i:k}\right)\\ &+ \sum_{i}^{n_{\mathrm{d}}}\sum_{j}^{n_i}\frac{c_{ij}^2}{C_1}\left(\hat{\pi}_k-\hat{\pi}_{ij:k}\right)+\left(\frac{C_2}{C_1}-\sum_{i}^{n_{\mathrm{d}}}\sum_{j}^{n_i}\frac{c_{ij}^2}{C_1}\right)\left(\hat{\pi}_k-\hat{Q}_{1:k}\right)\\ &+ \left(C_1-\frac{C_2}{C_1}\right)\left(\hat{\pi}_k-\hat{Q}_{2:k}\right) \qquad \text{(A12)}\end{aligned}$$

Taking expectation over all possible samples from all replicate populations sharing the same evolutionary history, we get from Equation A11:

$$\begin{aligned}
\mathbb{E}(SSI_{:k}) &= \sum_{i}^{n_{\mathrm{d}}} \sum_{j}^{n_i} \mathbb{E}\left(\hat{\pi}_{i:k} - \hat{\pi}_{ij:k}\right) \mathbb{E}\left(\frac{c_{ij}^2}{C_{1i}}\right) \\
&+ \sum_{i}^{n_{\mathrm{d}}} \mathbb{E}\left(\hat{\pi}_{i:k} - \hat{Q}_{1i:k}\right) \mathbb{E}\left(C_{1i} - \sum_{j}^{n_i} \frac{c_{ij}^2}{C_{1i}}\right) \\
&= (\pi_k - Q_{1:k}) \left(C_1 - \mathbb{E}\left(\sum_{i}^{n_{\mathrm{d}}} \sum_{j}^{n_i} \frac{c_{ij}^2}{C_{1i}}\right)\right) \qquad \text{(A13)}
\end{aligned}$$

where $Q_{1:k}$ is the expected IIS probability that two genes in the same pool are both of type k. Likewise, from Equation A12:

$$\begin{aligned}
\mathbb{E}(SSP_{:k}) &= \sum_{i}^{n_{\mathrm{d}}} \sum_{j}^{n_i} \mathbb{E}\left(\hat{\pi}_{i:k} - \hat{\pi}_{ij:k}\right) \mathbb{E}\left(\frac{c_{ij}^2}{C_{1i}}\right) + \sum_{i}^{n_{\mathrm{d}}} \sum_{j}^{n_i} \mathbb{E}\left(\hat{\pi}_{k} - \hat{\pi}_{ij:k}\right) \mathbb{E}\left(\frac{c_{ij}^2}{C_{1}}\right) \\
&- \sum_{i}^{n_{\mathrm{d}}} \mathbb{E}\left(\hat{\pi}_{i:k} - \hat{Q}_{1i:k}\right) \mathbb{E}\left(C_{1i} - \sum_{j}^{n_i} \frac{c_{ij}^2}{C_{1i}}\right) \\
&+ \mathbb{E}\left(\hat{\pi}_{k} - \hat{Q}_{1:k}\right) \mathbb{E}\left(\frac{C_2}{C_1} - \sum_{i}^{n_{\mathrm{d}}} \sum_{j}^{n_i} \frac{c_{ij}^2}{C_1}\right) \\
&+ \left(C_1 - \frac{C_2}{C_1}\right) \mathbb{E}\left(\hat{\pi}_{k} - \hat{Q}_{2:k}\right) \\
&= (\pi_k - Q_{1:k}) \left(\frac{C_2}{C_1} - \mathbb{E}\left(\sum_{i}^{n_{\mathrm{d}}} \sum_{j}^{n_i} \frac{c_{ij}^2}{C_1}\right)\right) \\
&- (\pi_k - Q_{1:k}) \left(C_1 - \mathbb{E}\left(\sum_{i}^{n_{\mathrm{d}}} \sum_{j}^{n_i} \frac{c_{ij}^2}{C_{1i}}\right)\right) \\
&+ \left(C_1 - \frac{C_2}{C_1}\right) (\pi_k - Q_{2:k}) \qquad \text{(A14)}
\end{aligned}$$

where $Q_{2:k}$ is the expected IIS probability that two genes from different pools are both of type k. Note that the expected sums $\mathbb{E}\left(\sum_i \sum_j c_{ij}^2\right)/C_{1i}$ and $\mathbb{E}\left(\sum_i \sum_j c_{ij}^2\right)/C_1$ in Equations A13 and A14 depend on the latent variable

c_{ij}, that cannot be directly observed from the data. Therefore, we must make an assumption on the distribution of the c_{ij}'s to proceed. In the following, we assume that for each pool i, c_{ij} follows a multinomial distribution with parameter C_{1i} (the number of trials, i.e. the total number of reads in the ith pool) and probabilities $(1/n_i, \ldots, 1/n_i)$ for the n_i individuals in the pool. Then:

$$\begin{aligned}
\mathbb{E}\left(\sum_i^{n_\mathrm{d}}\sum_j^{n_i}\frac{c_{ij}^2}{C_{1i}}\right) &= \sum_i^{n_\mathrm{d}}\frac{1}{C_{1i}}\sum_j^{n_i}\mathbb{E}\left(c_{ij}^2\right) \\
&= \sum_i^{n_\mathrm{d}}\frac{1}{C_{1i}}\sum_j^{n_i}\left(\mathbb{E}\left(c_{ij}\right)^2+\mathbb{V}\left(c_{ij}\right)\right) \\
&= \sum_i^{n_\mathrm{d}}\frac{1}{C_{1i}}\sum_j^{n_i}\left(\left(\frac{C_{1i}}{n_i}\right)^2+\frac{C_{1i}}{n_i}\left(\frac{n_i-1}{n_i}\right)\right) \\
&= \sum_i^{n_\mathrm{d}}\left(\frac{C_{1i}}{n_i}+\left(\frac{n_i-1}{n_i}\right)\right) \equiv D_2 \qquad \text{(A15)}
\end{aligned}$$

and:

$$\begin{aligned}
\mathbb{E}\left(\sum_i^{n_\mathrm{d}}\sum_j^{n_i}\frac{c_{ij}^2}{C_1}\right) &= \frac{1}{C_1}\sum_i^{n_\mathrm{d}}\sum_j^{n_i}\mathbb{E}\left(c_{ij}^2\right) \\
&= \frac{1}{C_1}\sum_i^{n_\mathrm{d}}C_{1i}\left[\frac{C_{1i}}{n_i}+\left(\frac{n_i-1}{n_i}\right)\right] \equiv D_2^\star \qquad \text{(A16)}
\end{aligned}$$

Hence, from Equations A13 and A15, we have:

$$\mathbb{E}(SSI_{:k}) = (C_1 - D_2)\left(\pi_k - Q_{1:k}\right) \qquad \text{(A17)}$$

and from Equations A14 and A16:

$$\begin{aligned}\mathbb{E}(SSP_{:k}) &= \left(\frac{C_2}{C_1} - D_2^\star\right)(\pi_k - Q_{1:k}) - (C_1 - D_2)(\pi_k - Q_{1:k}) \\ &+ \left(C_1 - \frac{C_2}{C_1}\right)(\pi_k - Q_{2:k}) \\ &= \left(C_1 - \frac{C_2}{C_1}\right)(Q_{1:k} - Q_{2:k}) \\ &+ (D_2 - D_2^\star)(\pi_k - Q_{1:k}) \end{aligned} \tag{A18}$$

Summing over alleles, we get the following expressions for the expected sums of squares for genes between individuals within pools:

$$\mathbb{E}(SSI) = \sum_k \mathbb{E}(SSI_{:k}) = (C_1 - D_2)(1 - Q_1) \tag{A19}$$

and for genes between individuals from different pools:

$$\begin{aligned}\mathbb{E}(SSP) &= \sum_k \mathbb{E}(SSP_{:k}) \\ &= \left(C_1 - \frac{C_2}{C_1}\right)(Q_1 - Q_2) + (D_2 - D_2^\star)(1 - Q_1) \end{aligned} \tag{A20}$$

Rearranging Equations A19–A20, we get:

$$Q_1 - Q_2 = \frac{(C_1 - D_2)\,\mathbb{E}(SSP) - (D_2 - D_2^\star)\,\mathbb{E}(SSI)}{(C_1 - D_2)(C_1 - C_2/C_1)} \tag{A21}$$

and:

$$1 - Q_2 = \frac{(C_1 - D_2)\,\mathbb{E}(SSP) + (n_\mathrm{c} - 1)(D_2 - D_2^\star)\,\mathbb{E}(SSI)}{(C_1 - D_2)(C_1 - C_2/C_1)} \tag{A22}$$

where $n_\mathrm{c} \equiv (C_1 - C_2/C_1)/(D_2 - D_2^\star)$. Let $MSI \equiv SSI/(C_1 - D_2)$ and $MSP \equiv SSP/(D_2 - D_2^\star)$. Then, using the definition of F_{ST} from Equation 1

in the main text, and rearranging Equations A21–A22, we get:

$$F_{\mathrm{ST}} \equiv \frac{Q_1 - Q_2}{1 - Q_2} = \frac{\mathbb{E}(MSP) - \mathbb{E}(MSI)}{\mathbb{E}(MSP) + (n_{\mathrm{c}} - 1)\,\mathbb{E}(MSI)} \tag{A23}$$

which yields the method-of-moments estimator:

$$\hat{F}_{\mathrm{ST}}^{\mathrm{pool}} = \frac{MSP - MSI}{MSP + (n_{\mathrm{c}} - 1)\,MSI} \tag{A24}$$

Since SSI (Equation A3) and SSP (Equation A4) may be rewritten in terms of sample frequencies as:

$$\begin{aligned} SSI &= \sum_k SSI_{:k} = \sum_k \sum_i^{n_{\mathrm{d}}} \sum_j^{n_i} \sum_r^{c_{ij}} \left(X_{ij:k} - X_{i\cdot:k}\right)^2 \\ &= \sum_k \sum_i^{n_{\mathrm{d}}} C_{1i} \hat{\pi}_{i:k} \left(1 - \hat{\pi}_{i:k}\right) \end{aligned} \tag{A25}$$

and:

$$\begin{aligned} SSP &= \sum_k SSP_k = \sum_k \sum_i^{n_{\mathrm{d}}} \sum_j^{n_i} \sum_r^{c_{ij}} \left(X_{i\cdot:k} - X_{\cdot\cdot:k}\right)^2 \\ &= \sum_k \sum_i^{n_{\mathrm{d}}} C_{1i} \left(\hat{\pi}_{i:k} - \hat{\pi}_k\right)^2 \end{aligned} \tag{A26}$$

our estimator then takes the form:

$$\hat{F}_{\mathrm{ST}}^{\mathrm{pool}} = \frac{\sum_k \left[(C_1 - D_2) \sum_i^{n_{\mathrm{d}}} C_{1i} \left(\hat{\pi}_{i:k} - \hat{\pi}_k\right)^2 - (D_2 - D_2^{\star}) \sum_i^{n_{\mathrm{d}}} C_{1i} \hat{\pi}_{i:k} \left(1 - \hat{\pi}_{i:k}\right)\right]}{\sum_k \left[(C_1 - D_2) \sum_i^{n_{\mathrm{d}}} C_{1i} \left(\hat{\pi}_{i:k} - \hat{\pi}_k\right)^2 + (n_{\mathrm{c}} - 1)\,(D_2 - D_2^{\star}) \sum_i^{n_{\mathrm{d}}} C_{1i} \hat{\pi}_{i:k} \left(1 - \hat{\pi}_{i:k}\right)\right]} \tag{A27}$$

The estimator in Equation A24 can also be expressed as a function of the

frequencies of identical pairs of genes $\hat{Q}_1 = \sum_k \hat{Q}_{1:k}$ and $\hat{Q}_2 = \sum_k \hat{Q}_{2:k}$, as:

$$\hat{F}_{\mathrm{ST}}^{\mathrm{pool}} = \frac{\left(\hat{Q}_1 - \hat{Q}_2\right)\alpha + \left(C_1 - \sum_i \sum_j \frac{c_{ij}^2}{C_1}\right)\beta}{\left(1 - \hat{Q}_2\right)\alpha + \left(C_2/C_1 - \sum_i \sum_j \frac{c_{ij}^2}{C_1}\right)\beta} \tag{A28}$$

where:

$$\alpha \equiv \left(C_1 - \sum_i \sum_j \frac{c_{ij}^2}{C_{1i}}\right)\left(C_1 - \frac{C_2}{C_1}\right) \tag{A29}$$

and:

$$\beta \equiv \sum_i \left(C_{1i} - \sum_j \frac{c_{ij}^2}{C_{1i}}\right)\left(\hat{Q}_{1i} - \hat{Q}_1\right) \tag{A30}$$

If we take the limit case where the number of sequenced reads per gene is constant, i.e. if $C_{1i} = C$, for all $i \in (1, \ldots, n_{\mathrm{d}})$, then it can be shown that Equation A28 reduces exactly to Equations 28A29–28A30 in Rousset (2007), p. 977. Furthermore, if the pools have all the same size, i.e. if $n_i = n$ for all $i \in (1, \ldots, n_{\mathrm{d}})$, then $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}} = \left(\hat{Q}_1 - \hat{Q}_2\right) / \left(1 - \hat{Q}_2\right)$.

If the pools have all the same size and if the number of reads per pool is constant, then one can also show that Equations A25–A26 reduce to:

$$SSI = n_{\mathrm{d}}(C - 1)\left(1 - \hat{Q}_1^{\mathrm{r}}\right) \tag{A31}$$

and:

$$SSP = C(n_{\mathrm{d}} - 1)\left(1 - \hat{Q}_2^{\mathrm{r}}\right) - (n_{\mathrm{d}} - 1)(C - 1)\left(1 - \hat{Q}_1^{\mathrm{r}}\right) \tag{A32}$$

where $\hat{Q}_1^{\mathrm{r}}$ and $\hat{Q}_2^{\mathrm{r}}$ are the frequencies of identical pairs of reads within and between pools, respectively, computed by simple counting of IIS pairs. These

are (unweighted) averages of the population-specific estimates $\hat{Q}^{\mathrm{r}}_{1i}$ (Equation A34) and the pairwise estimates $\hat{Q}^{\mathrm{r}}_{2ii'}$ (Equation A40), respectively. Then, from Equation A24, we get:

$$\hat{F}^{\mathrm{pool}}_{\mathrm{ST}} = 1 - \left(\frac{1-\hat{Q}^{\mathrm{r}}_{1}}{1-\hat{Q}^{\mathrm{r}}_{2}}\right)\left(\frac{n}{n-1}\right) \tag{A33}$$

IIS probabilities for Pool-seq data

In this Appendix, we provide unbiased estimates of IIS probabilies between pairs of genes, computed from read count data. Let $r_{i:k} = \sum_j \sum_r X_{ijr:k}$ be the number of reads of type k in the ith pool. A straightforward estimate of the IIS probability between pairs of reads in the ith pool is given by:

$$\hat{Q}^{\mathrm{r}}_{1i} \equiv \frac{\sum_k r_{i:k}\left(r_{i:k}-1\right)}{C_{1i}\left(C_{1i}-1\right)} \tag{A34}$$

where $C_{1i} = \sum_k r_{i:k}$. As above (see Equations A15 and A16), we assume that in each pool, the conditional distribution of the read counts $r_{i:k}$, given the (unobserved) allele counts $y_{i:k}$, is binomial, i.e.: $r_{i:k} \mid y_{i:k} \sim \mathrm{Bin}\left(y_{i:k}/n_i, C_{1i}\right)$. The conditional expectation of the number of reads is therefore given by: $\mathbb{E}(r_{i:k} \mid y_{i:k}) = C_{1i}\left(y_{i:k}/n_i\right)$, and the conditional expectation of the squared number of reads by: $\mathbb{E}(r^2_{i:k} \mid y_{i:k}) = C_{1i}(C_{1i}-1)\left(y_{i:k}/n_i\right)^2 + C_{1i}\left(y_{i:k}/n_i\right)$. Therefore, the conditional expectation of the IIS probability between pairs of reads in the ith pool reads:

$$\mathbb{E}\left(\hat{Q}^{\mathrm{r}}_{1i} \mid y_{i:k}\right) = \frac{\sum_k \mathbb{E}\left(r^2_{i:k} - r_{i:k}\right)}{C_{1i}\left(C_{1i}-1\right)} = \sum_k \left(\frac{y_{i:k}}{n_i}\right)^2 \tag{A35}$$

Since

$$\hat{Q}_{1i} \equiv \frac{\sum_k y_{i:k}\left(y_{i:k}-1\right)}{n_i\left(n_i-1\right)} \tag{A36}$$

is an unbiased estimate of the IIS probability between pairs of distinct genes in the ith pool, Equation A35 implies that $\hat{Q}^{\mathrm{r}}_{1i}$ (Equation A34) is a biased estimate of that quantity (i.e., the IIS probability between pairs of reads within a pool is a biased estimate of the IIS probability between pairs of

distinct genes in that pool). This is so, because the former confounds pairs of reads that are identical because they were sequenced from a single gene copy, from pairs of reads (from distinct gene copies) that are identical because they share a common ancestor. However, inspection of Equation A35 suggests that an unbiased estimate of $\hat{Q}_{1i}$ may be given by:

$$\hat{Q}_{1i}^{\text{pool}} \equiv 1 - \frac{n_i}{n_i - 1}\left(1 - \hat{Q}_{1i}^{\text{r}}\right) \tag{A37}$$

Taking expectation of Equation A37, we get indeed:

$$\begin{aligned}
\mathbb{E}\left(\hat{Q}_{1i}^{\text{pool}} \mid y_{i:k}\right) &= \frac{n_i}{n_i - 1}\mathbb{E}\left(\hat{Q}_{1i}^{\text{r}}\right) - \frac{1}{n_i - 1} \\
&= \frac{n_i}{n_i - 1}\sum_k \left(\frac{y_{i:k}}{n_i}\right)^2 - \frac{n_i}{n_i(n_i - 1)} \\
&= \frac{\sum_k y_{i:k}^2}{n_i(n_i - 1)} - \frac{\sum_k y_{i:k}}{n_i(n_i - 1)} \\
&= \frac{\sum_k y_{i:k}(y_{i:k} - 1)}{n_i(n_i - 1)} \equiv \hat{Q}_{1i}
\end{aligned} \tag{A38}$$

Following Weir and Goudet (2017), we define the overall IIS probability between pairs of genes within pools as the unweighted average of population-specific estimates, leading to:

$$\hat{Q}_1^{\text{pool}} \equiv \frac{\sum_i \hat{Q}_{1i}^{\text{pool}}}{n_{\text{d}}} = 1 - \frac{1}{n_{\text{d}}}\sum_i \frac{n_i}{n_i - 1}\left(1 - \hat{Q}_{1i}^{\text{r}}\right) \tag{A39}$$

A straightforward estimate of the IIS probability between pairs of reads taken in different pools i and i' is given by:

$$\hat{Q}_{2ii'}^{\text{r}} \equiv \frac{\sum_k r_{i:k} r_{i':k}}{C_{1i} C_{1i'}} \tag{A40}$$

Since we assume that pools are conditionally independent, taking expectation gives:

$$\begin{aligned}\mathbb{E}\left(\hat{Q}^{\mathrm{r}}_{2ii'} \mid y_{i:k}, y_{i':k}\right) &= \frac{\sum_k \mathbb{E}(r_{i:k})\mathbb{E}(r_{i':k})}{C_{1i}C_{1i'}} \\ &= \sum_k \left(\frac{y_{i:k}y_{i':k}}{n_i n_{i'}}\right) \equiv \hat{Q}_{2ii'} \end{aligned} \tag{A41}$$

Therefore, the IIS probability between pairs of reads sampled in different pools is an unbiased estimate of the IIS probability between pairs of genes in these pools, and an unbiased estimate of the IIS probabilitiy of genes sampled from different pools is given by:

$$\hat{Q}^{\mathrm{pool}}_{2ii'} \equiv \hat{Q}^{\mathrm{r}}_{2ii'} \tag{A42}$$

As above, we define the overall IIS probability between pairs of genes sampled from different pools as the unweighted average of pairwise estimates, i.e.:

$$\hat{Q}^{\mathrm{pool}}_{2} \equiv \frac{\sum_{i\neq i'} \hat{Q}^{\mathrm{pool}}_{2ii'}}{n_{\mathrm{d}}(n_{\mathrm{d}}-1)} = 1 - \frac{1}{n_{\mathrm{d}}(n_{\mathrm{d}}-1)} \sum_{i\neq i'} \left(1 - \hat{Q}^{\mathrm{r}}_{2ii'}\right) \tag{A43}$$

We can then derive an IIS-based estimator of F_{ST}, as:

$$\begin{aligned}\hat{F}^{\mathrm{pool-PID}}_{\mathrm{ST}} &\equiv \frac{\hat{Q}^{\mathrm{pool}}_{1} - \hat{Q}^{\mathrm{pool}}_{2}}{1 - \hat{Q}^{\mathrm{pool}}_{2}} = 1 - \frac{1 - \hat{Q}^{\mathrm{pool}}_{1}}{1 - \hat{Q}^{\mathrm{pool}}_{2}} \\ &= 1 - \frac{\sum_i \left[\left(1 - \hat{Q}^{\mathrm{r}}_{1i}\right) n_i / \left(n_i - 1\right)\right]}{\sum_{i\neq i'} \left(1 - \hat{Q}^{\mathrm{r}}_{2ii'}\right) / \left(n_{\mathrm{d}} - 1\right)} \end{aligned} \tag{A44}$$

which, to the extent that we may take the expectation of a ratio to be the ratio of expectations, is unbiased. If the pools have all the same size (i.e., if

$n_i = n$ for all i), then Equation A44 reduces to:

$$\hat{F}_{\mathrm{ST}}^{\mathrm{pool-PID}} = 1 - \left(\frac{1-\hat{Q}_1^{\mathrm{r}}}{1-\hat{Q}_2^{\mathrm{r}}}\right)\left(\frac{n}{n-1}\right) \tag{A45}$$

where $\hat{Q}_1^{\mathrm{r}} \equiv \sum_i \hat{Q}_{1i}^{\mathrm{r}}/n_{\mathrm{d}}$ and $\hat{Q}_2^{\mathrm{r}} \equiv \sum_{i\neq i'} \hat{Q}_{2ii'}^{\mathrm{r}}/\left[n_{\mathrm{d}}(n_{\mathrm{d}}-1)\right]$. Note that Equation A45 is strictly identical to Equation A33. Therefore, if the pools have all the same size and if the number of reads per pool is constant, the analysis-of-variance estimator $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ is strictly equivalent to the estimator $\hat{F}_{\mathrm{ST}}^{\mathrm{pool-PID}}$ based on the computation of IIS probabilities between pairs of reads, with appropriate bias correction (see Equation A37). This echoes the derivations by Rousset (2007) for Ind-seq data, who showed that the analysis-of-variance approach (Weir and Cockerham 1984) and the simple strategy of estimating IIS probabilities by counting identical pairs of genes provides identical estimates when sample sizes are equal (see also Cockerham and Weir 1987; Karlsson et al. 2007).

Alternatively, the overall IIS probability between pairs of genes within pools may be defined as the weighted average of population-specific estimates, with weights equal to the number of pairs of genes in each pool (see Rousset 2007), i.e.:

$$\tilde{Q}_1^{\mathrm{pool}} \equiv \frac{\sum_i n_i(n_i-1)\hat{Q}_{1i}^{\mathrm{pool}}}{\sum_i n_i(n_i-1)} \tag{A46}$$

Likewise, the overall IIS probability between pairs of genes sampled from different pools may be defined as the weighted average of pairwise estimates, with weights equal to the number of pairs of genes sampled between pools,

i.e.:

$$\tilde{Q}_2^{\text{pool}} \equiv \frac{\sum_{i \neq i'} n_i n_{i'} \hat{Q}_{2ii'}^{\text{pool}}}{\sum_{i \neq i'} n_i n_{i'}} \tag{A47}$$

We can then derive an IIS-based estimator of F_{ST}, using weighted IIS probabilities, as:

$$\begin{aligned} \tilde{F}_{\text{ST}}^{\text{pool-PID}} &\equiv \frac{\tilde{Q}_1^{\text{pool}} - \tilde{Q}_2^{\text{pool}}}{1 - \tilde{Q}_2^{\text{pool}}} = 1 - \frac{1 - \tilde{Q}_1^{\text{pool}}}{1 - \tilde{Q}_2^{\text{pool}}} \\ &= 1 - \frac{\sum_i \left[n_i^2 \left(1 - \hat{Q}_{1i}^{\text{r}} \right) \right] / \sum_i n_i (n_i - 1)}{\sum_{i \neq i'} n_i n_{i'} \left(1 - \hat{Q}_{2ii'}^{\text{r}} \right) / \sum_{i \neq i'} n_i n_{i'}} \end{aligned} \tag{A48}$$

If the pools have all the same size (i.e., if $n_i = n$ for all i), then Equation A48 reduces to Equation A45, and $\tilde{F}_{\text{ST}}^{\text{pool-PID}} = \hat{F}_{\text{ST}}^{\text{pool-PID}}$. With unbalanced samples, simulation analyses show that $\tilde{F}_{\text{ST}}^{\text{pool-PID}}$ has larger bias and variance than $\hat{F}_{\text{ST}}^{\text{pool-PID}}$, in particular for low levels of differentiation (see Figure S4).

LITERATURE CITED

Cockerham, C. C. (1973). Analyses of gene frequencies. *Genetics*, 74:679–700.

Cockerham, C. C. and Weir, B. S. (1987). Analyses of gene frequencies. *Proc. Natl. Acad. Sci. USA*, 84:8512–8514.

Karlsson, E. K., Baranowska, I., Wade, C. M., Salmon Hillbertz, N. H. C., Zody, M. C., Anderson, N., Biagi, T. M., Patterson, N., Pielberg, G. R., Kulbokas, E. J., Comstock, K. E., Keller, E. T., Mesirov, J. P., von Euler, H., Kämpe, O., Hedhammar, A., Lander, E. S., Andersson, G., Andersson, L., and Lindblad-Toh, K. (2007). Efficient mapping of Mendelian traits in dogs through genome-wide association. *Nat. Genet.*, 39:1321–1328.

Rousset, F. (2007). Inferences from spatial population genetics. In Balding, D. J., Bishop, M., and Cannings, C., editors, *Handbook of Statistical Genetics*, pages 945–979, Chichester. John Wiley & Sons, Ltd.

Weir, B. S. (1996). *Genetic Data Analysis II*. Sinauer Associates, Inc., Sunderland, MA.

Weir, B. S. and Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, 38:1358–1370.

Weir, B. S. and Goudet, J. (2017). An unified characterization of population structure and relatedness. *Genetics*, 206:2085–2103.

Weir, B. S. and Hill, W. G. (2002). Estimating F-statistics. *Annu. Rev. Genet.*, 36:721–750.