Measuring genetic differentiation from Pool-seq data

Valentin Hivert[⋆,†], Raphël Leblois[⋆,†], Eric J. Petit[‡], Mathieu Gautier[⋆,†,§], and Renaud Vitalis[⋆,†,§]

[⋆]CBGP, INRA, CIRAD, IRD, Montpellier SupAgro, Univ Montpellier, Montpellier, France

[†]Institut de Biologie Computationnelle, Univ Montpellier, Montpellier, France

[‡]ESE, Ecology and Ecosystem Health, INRA, Agrocampus Ouest, Rennes, France

[§]These authors are joint senior authors on this work

Running title: Genetic differentiation from pools

Keywords: F_{ST}, genetic differentiation, pool sequencing, population genomics

Corresponding author: Renaud Vitalis

Centre de Biologie pour la Gestion des Populations

Campus International de Baillarguet, CS 30 016

34988 Montferrier-sur-Lez cedex

France

Tel : +33 (0)4 99 62 33 42

Fax : +33 (0)4 99 62 33 45

E-mail: `renaud.vitalis@inra.fr`

Abstract

The advent of high throughput sequencing and genotyping technologies enables the comparison of patterns of polymorphisms at a very large number of markers. While the characterization of genetic structure from individual sequencing data remains expensive for many non-model species, it has been shown that sequencing pools of individual DNAs (Pool-seq) represents an attractive and cost-effective alternative. However, analyzing sequence read counts from a DNA pool instead of individual genotypes raises statistical challenges in deriving correct estimates of genetic differentiation. In this article, we provide a method-of-moments estimator of F_{ST} for Pool-seq data, based on an analysis-of-variance framework. We show, by means of simulations, that this new estimator is unbiased, and outperforms previously proposed estimators. We evaluate the robustness of our estimator to model misspecification, such as sequencing errors and uneven contributions of individual DNAs to the pools. Finally, by reanalyzing published Pool-seq data of different ecotypes of the prickly sculpin *Cottus asper*, we show how the use of an unbiased F_{ST} estimator may question the interpretation of population structure inferred from previous analyses.

INTRODUCTION

It has long been recognized that the subdivision of species into subpopulations, social groups and families fosters genetic differentiation (Wahlund 1928; Wright 1931). Characterizing genetic differentiation as a means to infer unknown population structure is therefore fundamental to population genetics, and finds applications in multiple domains, including conservation biology, invasion biology, association mapping and forensics, among many others. In the late 1940s and early 1950s, Malécot (1948) and Wright (1951) introduced F-statistics to partition genetic variation within and between groups of individuals (Holsinger and Weir 2009; Bhatia et al. 2013). Since then, the estimation of F-statistics has become standard practice (see, e.g., Weir 1996; Weir and Hill 2002; Weir 2012), and the most commonly used estimators of F_{ST} have been developed in an analysis-of-variance framework (Cockerham 1969, 1973; Weir and Cockerham 1984), which can be recast in terms of probabilities of identity of pairs of homologous genes (Cockerham and Weir 1987; Rousset 2007; Weir and Goudet 2017).

Assuming that molecular markers are neutral, estimates of F_{ST} are typically used to quantify genetic structure in natural populations, which is then interpreted as the result of demographic history (Holsinger and Weir 2009): large F_{ST} values are expected for small populations among which dispersal is limited (Wright 1951), or between populations that have long diverged in isolation from each other (Reynolds et al. 1983); when dispersal is spatially restricted, a positive relationship between F_{ST} and the geographical distance for pairs of populations generally holds (Slatkin 1993; Rousset 1997). It has also been proposed to characterize the heterogeneity of F_{ST}

estimates across markers for identifying loci that are targeted by selection (Cavalli-Sforza 1966; Lewontin and Krakauer 1973; Beaumont and Nichols 1996; Vitalis et al. 2001; Akey et al. 2002; Beaumont 2005; Weir et al. 2005; Lotterhos and Whitlock 2014, 2015; Whitlock and Lotterhos 2015).

Next-generation sequencing (NGS) technologies provide unprecedented amounts of polymorphism data in both model and non-model species (Ellegren 2014). Although the sequencing strategy initially involved individually tagged samples in humans (The International HapMap Consortium 2005), whole-genome sequencing of pools of individuals (Pool-seq) is being increasingly popular for population genomic studies (Schlötterer et al. 2014). Because it consists in sequencing libraries of pooled DNA samples and does not require individual tagging of sequences, Pool-seq provides genome-wide polymorphism data at considerably lower cost than sequencing of individuals (Schlötterer et al. 2014). However, non-equimolar amounts of DNA from all individuals in a pool and stochastic variation in the amplification efficiency of individual DNAs have raised concerns with respect to the accuracy of the so-obtained allele frequency estimates, particularly at low sequencing depth and with small pool sizes (Cutler and Jensen 2010; Ellegren 2014; Anderson et al. 2014). Nonetheless, it has been shown that, at equal sequencing effort, Pool-seq provides similar, if not more accurate, allele frequency estimates than individual-based analyses (Futschik and Schlötterer 2010; Gautier et al. 2013). The problem is different for diversity and differentiation parameters, which depend on second moments of allele frequencies or, equivalently, on pairwise measures of genetic identity. With Pool-seq data, however, it is impossible to distinguish pairs of reads that are identical because they were

sequenced from a single gene, from pairs of reads that are identical because they were sequenced from two distinct genes that are identical in state (IIS) (Ferretti et al. 2013).

Appropriate estimators of diversity and differentiation parameters must therefore be sought, to account for both the sampling of individual genes from the pool and the sampling of reads from these genes. There has been several attempts to define estimators for the parameter F_{ST} for Pool-seq data (Kofler et al. 2011; Ferretti et al. 2013), from ratios of heterozygosities (or from probabilities of genetic identity between pairs of reads) within and between pools. In the following, we will argue that these estimators are biased (i.e., they do not converge towards the expected value of the parameter), and that some of them have undesired statistical properties (i.e., the bias depends upon sample size and coverage). Here, following Cockerham (1969), Cockerham (1973), Weir and Cockerham (1984), Weir (1996), Weir and Hill (2002) and Rousset (2007), we define a method-of-moments estimator of the parameter F_{ST} using an analysis-of-variance framework. We then evaluate the accuracy and the precision of this estimator, based on the analysis of simulated datasets, and compare it to estimates defined in the software package PoPoolation2 (Kofler et al. 2011), and in Ferretti et al. (2013). Furthermore, we test the robustness of our estimators to model misspecifications (including unequal contributions of individuals in pools, and sequencing errors). Finally, we reanalyze the prickly sculpin (*Cottus asper*) Pool-seq data (published by Dennenmoser et al. 2017), and show how the use of biased F_{ST} estimators in previous analyses may challenge the interpretation of population structure.

Note that throughout this article, we use the term "gene" to designate a

96 segregating genetic unit (in the sense of the "Mendelian gene" from Orgogozo
97 et al. 2016). We further use the term "read" in a narrow sense, as a sequenced
98 copy of a gene. For the sake of simplicity, we will use the term "Ind-seq" to
99 refer to analyses based on individual data in which we further assume that
100 individual genotypes are called without error.

MODEL

F-statistics may be described as intra-class correlations for the probability of identity in state (IIS) of pairs of genes (Cockerham and Weir 1987; Rousset 1996, 2007), and F_{ST} is best defined as:

$$F_{\mathrm{ST}} \equiv \frac{Q_1 - Q_2}{1 - Q_2} \tag{1}$$

where Q_1 is the IIS probability for genes sampled within subpopulations, and Q_2 is the IIS probability for genes sampled between subpopulations. In the following, we develop an estimator of F_{ST} for Pool-seq data, by decomposing the total variance of gene frequencies in an analysis-of-variance framework. A complete derivation of the model is provided in the Supplemental File S1.

For the sake of clarity, the notation used throughout this article is given in Table 1. We first derive our model for a single locus, and eventually provide a multilocus estimator of F_{ST}. 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 fol-

lows:

$$\begin{aligned}\sum_{i}^{n_{\mathrm{d}}}\sum_{j}^{n_i}\sum_{r}^{c_{ij}}\left(X_{ijr:k}-X_{\cdots:k}\right)^2 &= \sum_{i}^{n_{\mathrm{d}}}\sum_{j}^{n_i}\sum_{r}^{c_{ij}}\left(X_{ijr:k}-X_{ij\cdot:k}\right)^2 \\ &+ \sum_{i}^{n_{\mathrm{d}}}\sum_{j}^{n_i}\sum_{r}^{c_{ij}}\left(X_{ij\cdot:k}-X_{i\cdots:k}\right)^2 \\ &+ \sum_{i}^{n_{\mathrm{d}}}\sum_{j}^{n_i}\sum_{r}^{c_{ij}}\left(X_{i\cdots:k}-X_{\cdots:k}\right)^2 \\ &\equiv SSR_{:k}+SSI_{:k}+SSP_{:k} \qquad (2)\end{aligned}$$

As is shown in the Supplemental File S1, the expected sums of squares depend on the expectation of the allele frequency π_k over all replicate populations sharing the same evolutionary history, as well as on the IIS probability $Q_{1:k}$ that two genes in the same pool are both of type k, and the IIS probability $Q_{2:k}$ that two genes from different pools are both of type k. Taking expectations (see the detailed computations in the Supplemental File S1), one has:

$$\mathbb{E}(SSR_{:k}) = 0 \qquad (3)$$

for reads within individual genes, since we assume that there is no sequencing error, i.e. all the reads sequenced from a single gene are identical and $X_{ijr:k} = X_{ij\cdot:k}$ for all r. For reads between genes within pools, we get:

$$\mathbb{E}(SSI_{:k}) = (C_1 - D_2)(\pi_k - Q_{1:k}) \qquad (4)$$

where $C_1 \equiv \sum_i \sum_j c_{ij} = \sum_j c_{ij}$ is the total number of reads in the full sample (total coverage), C_{1i} is the coverage of the ith pool and $D_2 \equiv \sum_i (C_{1i} + n_i - 1)/n_i$. D_2 arises from the assumption that the distribution of the read counts c_{ij} is multinomial (i.e., that all genes contribute equally to the pool of reads;

see Equation A15 in Supplemental File S1). For reads between genes from different pools, we have:

$$\mathbb{E}(SSP_{:k}) = \left(C_1 - \frac{C_2}{C_1}\right)(Q_{1:k} - Q_{2:k}) + (D_2 - D_2^\star)(\pi_k - Q_{1:k}) \quad (5)$$

where $C_2 \equiv \sum_i \left(\sum_j c_{ij}\right)^2$ and $D_2^\star \equiv \left[\sum_i C_{1i}(C_{1i} + n_i - 1)/n_i\right]/C_1$ (see Equation A16 in Supplemental File S1). Rearranging Equations 4–5, and summing over alleles, we get:

$$Q_1 - Q_2 = \frac{(C_1 - D_2)\,\mathbb{E}(SSP) - (D_2 - D_2^\star)\,\mathbb{E}(SSI)}{(C_1 - D_2)(C_1 - C_2/C_1)} \quad (6)$$

and:

$$1 - Q_2 = \frac{(C_1 - D_2)\,\mathbb{E}(SSP) + (n_c - 1)(D_2 - D_2^\star)\,\mathbb{E}(SSI)}{(C_1 - D_2)(C_1 - C_2/C_1)} \quad (7)$$

where $n_c \equiv (C_1 - C_2/C_1)/(D_2 - D_2^\star)$. Let $MSI \equiv SSI/(C_1 - D_2)$ and $MSP \equiv SSP/(D_2 - D_2^\star)$. Then:

$$F_{ST} \equiv \frac{Q_1 - Q_2}{1 - Q_2} = \frac{\mathbb{E}(MSP) - \mathbb{E}(MSI)}{\mathbb{E}(MSP) + (n_c - 1)\,\mathbb{E}(MSI)} \quad (8)$$

which yields the method-of-moments estimator:

$$\hat{F}_{ST}^{pool} = \frac{MSP - MSI}{MSP + (n_c - 1)\,MSI} \quad (9)$$

where

$$MSI = \frac{1}{C_1 - D_2} \sum_k \sum_i^{n_d} C_{1i}\hat{\pi}_{i:k}(1 - \hat{\pi}_{i:k}) \quad (10)$$

and:

$$MSP = \frac{1}{D_2 - D_2^\star} \sum_k \sum_i^{n_\mathrm{d}} C_{1i} \left(\hat{\pi}_{i:k} - \hat{\pi}_k\right)^2 \tag{11}$$

(see Equations A25 and A26 in Supplemental File S1). In Equations 10 and 11, $\hat{\pi}_{i:k} \equiv X_{i\cdots k}$ is the average frequency of reads of type k within the ith pool, and $\hat{\pi}_k \equiv X_{\cdots:k}$ is the average frequency of reads of type k in the full sample. 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. This is 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). Finally, the full expression of $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ in terms of sample frequencies reads:

$$\hat{F}_{\mathrm{ST}}^{\mathrm{pool}} = \frac{\sum_k \left[(C_1 - D_2) \sum_i^{n_\mathrm{d}} C_{1i} \left(\hat{\pi}_{i:k} - \hat{\pi}_k\right)^2 - (D_2 - D_2^\star) \sum_i^{n_\mathrm{d}} C_{1i} \hat{\pi}_{i:k} \left(1 - \hat{\pi}_{i:k}\right) \right]}{\sum_k \left[(C_1 - D_2) \sum_i^{n_\mathrm{d}} C_{1i} \left(\hat{\pi}_{i:k} - \hat{\pi}_k\right)^2 + (n_\mathrm{c} - 1)(D_2 - D_2^\star) \sum_i^{n_\mathrm{d}} C_{1i} \hat{\pi}_{i:k} \left(1 - \hat{\pi}_{i:k}\right) \right]} \tag{12}$$

If we take the limit case where each gene is sequenced exactly once, we recover the Ind-seq model: assuming $c_{ij} = 1$ for all (i, j), then $C_1 = \sum_i^{n_\mathrm{d}} n_i$, $C_2 = \sum_i^{n_\mathrm{d}} n_i^2$, $D_2 = n_\mathrm{d}$ and $D_2^\star = 1$. Therefore, $n_\mathrm{c} = \left(C_1 - C_2/C_1\right) / \left(n_\mathrm{d} - 1\right)$, and Equation 9 reduces exactly to the estimator of F_{ST} for haploids: see Weir (1996), p. 182, and Rousset (2007), p. 977.

As in Reynolds et al. (1983), Weir and Cockerham (1984), Weir (1996) and Rousset (2007), a multilocus estimate is derived as the sum of locus specific numerators over the sum of locus-specific denominators:

$$\hat{F}_{\mathrm{ST}} = \frac{\sum_l MSP_l - MSI_l}{\sum_l MSP_l + (n_\mathrm{c} - 1) MSI_l} \tag{13}$$

where MSI and MSP are subscripted with l to denote the lth locus. For Ind-seq data, Bhatia et al. (2013) refer to this multilocus estimate as a "ratio of averages" by opposition to an "average of ratios", which would consist in averaging single-locus F_{ST} over loci. This approach is justified in the Appendix of Weir and Cockerham (1984) and in Bhatia et al. (2013), who analyzed both estimates by means of coalescent simulations. Note that Equation 13 assumes that the pool size is equal across loci. Also note that the construction of the estimator in Equation 13 is different from Weir and Cockerham's (1984). These authors defined their multilocus estimator as a ratio of sums of components of variance (a, b and c in their notation) over loci, which give the same weight to all loci, whatever the number of sampled genes at each locus. Equation 13 follows GENEPOP's rationale (Rousset 2008), which gives instead more weight to loci that are more intensively covered.

MATERIALS AND METHODS

Simulation study

Generating individual genotypes: we first generated individual genotypes using `ms` (Hudson 2002), assuming an island model of population structure (Wright 1931). For each simulated scenario, we considered 8 demes, each made of $N = 5,000$ haploid individuals. The migration rate (m) was fixed to achieve the desired value of $F_{\rm ST}$ (0.05 or 0.2), using Equation 6 in Rousset (1996) leading, e.g., to $M \equiv 2Nm = 16.569$ for $F_{\rm ST} = 0.05$ and $M = 3.489$ for $F_{\rm ST} = 0.20$. The mutation rate was set at $\mu = 10^{-6}$, giving $\theta \equiv 2N\mu = 0.01$. We considered either fixed, or variable sample sizes across demes. In the latter case, the haploid sample size n was drawn independently for each deme from a Gaussian distribution with mean 100 and standard deviation 30; this number was rounded up to the nearest integer, with min. 20 and max. 300 haploids per deme. We generated a very large number of sequences for each scenario, and sampled independent single nucleotide polymorphisms (SNPs) from sequences with a single segregating site. Each scenario was replicated 50 times (500 times for Figures 3 and S2).

Pool sequencing: for each `ms` simulated dataset, we generated Pool-seq data by drawing reads from a binomial distribution (Gautier et al. 2013). More precisely, we assume that for each SNP, the number $r_{i:k}$ of reads of allelic type k in pool i follows:

$$r_{i:k} \sim \text{Bin}\left(\frac{y_{i:k}}{n_i}, \delta_i\right) \tag{14}$$

where $y_{i:k}$ is the number of genes of type k in the ith pool, n_i is the total number of genes in pool i (haploid pool size), and δ_i is the simulated total coverage for pool i. In the following, we either consider a fixed coverage, with $\delta_i = \Delta$ for all pools and loci, or a varying coverage across pools and loci, with $\delta_i \sim \text{Pois}(\Delta)$.

Sequencing error: we simulated sequencing errors occurring at rate $\mu_e = 0.001$, which is typical of Illumina sequencers (Glenn 2011; Ross et al. 2013). We assumed that each sequencing error modifies the allelic type of a read to one of three other possible states with equal probability (there are therefore four allelic types in total, corresponding to four nucleotides). Note that only biallelic markers are retained in the final datasets. Also note that, since we initiated this procedure with polymorphic markers only, we neglect sequencing errors that would create spurious SNPs from monomorphic sites. However, such SNPs should be rare in real datasets, since markers with a low minimum read count (MRC) are generally filtered out.

Experimental error: non-equimolar amounts of DNA from all individuals in a pool and stochastic variation in the amplification efficiency of individual DNAs are sources of experimental errors in pool sequencing. To simulate experimental errors, we used the model derived by Gautier et al. (2013). In this model, it is assumed that the contribution $\eta_{ij} = c_{ij}/C_{1i}$ of each gene j to the total coverage of the ith pool (C_{1i}) follows a Dirichlet distribution:

$$\{\eta_{ij}\}_{1 \leq j \leq n_i} \sim \text{Dir}\left(\frac{\rho}{n_i}\right) \tag{15}$$

where the parameter ρ controls the dispersion of gene contributions around the value $\eta_{ij} = 1/n_i$, expected if all genes contributed equally to the pool of reads. For convenience, we define the experimental error ϵ as the coefficient of variation of η_{ij}, i.e.: $\epsilon \equiv \sqrt{\mathbb{V}(\eta_{ij})}/\mathbb{E}(\eta_{ij}) = \sqrt{(n_i - 1)/(\rho + 1)}$ (see Gautier et al. 2013). When ϵ tends toward 0 (or equivalently when ρ tends to infinity), all individuals contribute equally to the pool, and there is no experimental error. We tested the robustness of our estimates to values of ϵ comprised between 0.05 and 0.5. The case $\epsilon = 0.5$ could correspond, for example, to a situation where (for $n_i = 10$) 5 individuals contribute 2.8× more reads than the other 5 individuals.

Other estimators

For the sake of clarity, a summary of the notation of the F_{ST} estimators used throughout this article is given in Table 2.

$\mathrm{PP2_d}$: this estimator of F_{ST} is implemented by default in the software package PoPoolation2 (Kofler et al. 2011). It is based on a definition of the parameter F_{ST} as the overall reduction in average heterozygosity relative to the total combined population (see, e.g., Nei and Chesser 1983):

$$\mathrm{PP2_d} \equiv \frac{\hat{H}_{\mathrm{T}} - \hat{H}_{\mathrm{S}}}{\hat{H}_{\mathrm{T}}} \tag{16}$$

where $\hat{H}_{\mathrm{S}}$ is the average heterozygosity within subpopulations, and $\hat{H}_{\mathrm{T}}$ is the average heterozygosity in the total population (obtained by pooling together all subpopulation to form a single virtual unit). In PoPoolation2, $\hat{H}_{\mathrm{S}}$ is

the unweighted average of within-subpopulation heterozygosities:

$$\hat{H}_{\mathrm{S}} = \frac{1}{n_{\mathrm{d}}} \sum_{i}^{n_{\mathrm{d}}} \left(\frac{n_i}{n_i - 1} \right) \left(\frac{C_{1i}}{C_{1i} - 1} \right) \left(1 - \sum_{k} \hat{\pi}_{i:k}^2 \right) \tag{17}$$

(using the notation from Table 1). Note that in PoPoolation2, $\mathrm{PP2_d}$ is restricted to the case of two subpopulations only ($n_{\mathrm{d}} = 2$). The two ratios in the right-hand side of Equation 17 are presumably borrowed from Nei (1978) to provide an unbiased estimate, although we found no formal justification for the expression in Equation 17 for Pool-seq data. The total heterozygosity is computed as (using the notation from Table 1):

$$\hat{H}_{\mathrm{T}} = \left(\frac{\min_i(n_i)}{\min_i(n_i) - 1} \right) \left(\frac{\min_i(C_{1i})}{\min_i(C_{1i}) - 1} \right) \left(1 - \sum_{k} \hat{\pi}_{k}^2 \right) \tag{18}$$

$\mathrm{PP2_a}$: this is the alternative estimator of F_{ST} provided in the software package PoPoolation2. It is based on an interpretation by Kofler et al. (2011) of Karlsson et al.'s (2007) estimator of F_{ST}, as:

$$\mathrm{PP2_a} \equiv \frac{\hat{Q}_1^{\mathrm{r}} - \hat{Q}_2^{\mathrm{r}}}{1 - \hat{Q}_2^{\mathrm{r}}} \tag{19}$$

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 estimates of Q_1^{r}, the IIS probability for two reads in the same pool (whether they are sequenced from the same gene or not) and Q_2^{r}, the IIS probability for two reads in different pools. Note that the IIS probabiliy Q_1^{r} is different from Q_1 in Equation 1, which, from our definition, represents the IIS probability between distinct genes in the same pool. This approach therefore confounds pairs of reads within pools that are identical because

they were sequenced from a single gene, from pairs of reads that are identical because they were sequenced from distinct, yet IIS genes.

FRP_{13} : this estimator of F_{ST} was developed by Ferretti et al. (2013) (see their Equations 3 and 10–13). Ferretti et al. (2013) use the same definition of F_{ST} as in Equation 16 above, although they estimate heterozygosities within and between pools as "average pairwise nucleotide diversities", which, from their definitions, are formally equivalent to IIS probabilities. In particular, they estimate the average heterozygosity within pools as (using the notation from Table 1):

$$\hat{H}_{\mathrm{S}} = \frac{1}{n_{\mathrm{d}}} \sum_{i}^{n_{\mathrm{d}}} \left(\frac{n_i}{n_i - 1} \right) \left(1 - \hat{Q}^{\mathrm{r}}_{1i} \right) \tag{20}$$

and the total heterozygosity among the n_{d} populations as:

$$\hat{H}_{\mathrm{T}} = \frac{1}{n_{\mathrm{d}}^2} \left[\sum_{i}^{n_{\mathrm{d}}} \left(\frac{n_i}{n_i - 1} \right) \left(1 - \hat{Q}^{\mathrm{r}}_{1i} \right) + \sum_{i \neq i'}^{n_{\mathrm{d}}} \left(1 - \hat{Q}^{\mathrm{r}}_{2ii'} \right) \right] \tag{21}$$

Analyses of Ind-seq data:

For the comparison of Ind-seq and Pool-seq datasets, we computed F_{ST} on subsamples of 5,000 loci. These subsamples were defined so that only those loci that were polymorphic in all coverage conditions were retained, and the same loci were used for the analysis of the corresponding Ind-seq data. For the latter, we used either the Nei and Chesser's (1983) estimator based on a ratio of heterozygosity (see Equation 16 above), hereafter denoted by NC_{83}, or the analysis-of-variance estimator developed by Weir and Cockerham (1984), hereafter denoted by WC_{84}.

All the estimators were computed using custom functions in the R soft-

ware environment for statistical computing, version 3.3.1 (R Core Team 2017). All these functions were carefully checked against available software packages, to ensure that they provided strictly identical results.

Application example: *Cottus asper*

Dennenmoser et al. (2017) investigated the genomic basis of adaption to osmotic conditions in the prickly sculpin (*Cottus asper*), an abundant euryhaline fish in northwestern North America. To do so, they sequenced the whole-genome of pools of individuals from two estuarine populations (CR, Capilano River Estuary; FE, Fraser River Estuary) and two freshwater populations (PI, Pitt Lake and HZ, Hatzic Lake) in southern British Columbia (Canada). We downloaded the four corresponding BAM files from the Dryad Digital Repository (doi: 10.5061/dryad.2qg01) and combined them into a single mpileup file using `SAMtools` version 0.1.19 (Li et al. 2009) with default options, except the maximum depth per BAM that was set to 5,000 reads. The resulting file was further processed using a custom `awk` script, to call SNPs and compute read counts, after discarding bases with a Base Alignment Quality (BAQ) score lower than 25. A position was then considered as a SNP if: (*i*) only two different nucleotides with a read count > 1 were observed (nucleotides with ≤ 1 read being considered as a sequencing error); (*ii*) the coverage was comprised between 10 and 300 in each of the four alignment files; (*iii*) the minor allele frequency, as computed from read counts, was ≥ 0.01 in the four populations. The final data set consisted of 608,879 SNPs.

Our aim here was to compare the population structure inferred from pairwise estimates of F_{ST}, using the estimator $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ on the one hand, and $\mathrm{PP2_d}$

on the other hand. Then, to conclude on which of the two estimators performs better, we compared the population structure inferred from $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ and $\mathrm{PP2_d}$ to that inferred from the Bayesian hierarchical model implemented in the software package BayPass (Gautier 2015). BayPass allows indeed the robust estimation of the scaled covariance matrix of allele frequencies across populations for Pool-seq data, which is known to be informative about population history (Pickrell and Pritchard 2012). The elements of the estimated matrix can be interpreted as pairwise and population-specific estimates of differentiation (Coop et al. 2010), and therefore provide a comprehensive description of population structure that makes full use of the available data.

Data availability

The authors state that all data necessary for confirming the conclusions presented in this article are fully represented within the article, figures, and tables. Supplemental Tables S1–S3 and Figures S1–S4 are available at FigShare, along with a complete derivation of the model in the Supplemental File S1 at FigShare.

RESULTS

Comparing Ind-seq and Pool-seq estimates of F_{ST}

Single-locus estimates $\hat{F}_{ST}^{pool}$ are highly correlated with the classical estimates WC_{84} (Weir and Cockerham 1984) computed on the individual data that were used to generate the pools in our simulations (see Figure 1). The variance of $\hat{F}_{ST}^{pool}$ across independent replicates decreases as the coverage increases. The correlation between $\hat{F}_{ST}^{pool}$ and WC_{84} is stronger for multilocus estimates (see Figure S1A).

Comparing Pool-seq estimators of F_{ST}

We found that our estimator $\hat{F}_{ST}^{pool}$ has extremely low bias ($< 0.5\%$ over all scenarios tested: see Tables 3 and S1-S3). In other words, the average estimates across multiple loci and replicates closely equals the expected value of the F_{ST} parameter, as given by Equation 6 in Rousset (1996), which is based on the computation of IIS probabilities in an island model of population structure. In all the situations examined, the bias did neither depend on the sample size (i.e., the size of each pool) nor on the coverage (see Figure 2). Only the variance of the estimator across independent replicates decreases as the sample size increases and/or as the coverage increases. At high coverage, the mean and root mean squared error (RMSE) of $\hat{F}_{ST}^{pool}$ over independent replicates are virtually indistinguishable from that of the WC_{84} estimator (see Table S1).

Figure 3 shows the RMSE of F_{ST} estimates for a wide range of pool sizes and coverage. The RMSE decreases as the pool size and/or the coverage increases. The F_{ST} estimates are more precise and accurate when differen-

tiation is low. Figure 3 provides some clues to evaluate the pool size and the coverage that is necessary to achieve the same RMSE than for Ind-seq data. Consider, for example, the case of samples of $n = 20$ haploids. For $F_{ST} \leq 0.05$ (in the conditions of our simulations), the RMSE of F_{ST} estimates based on Pool-seq data tends to the RMSE of F_{ST} estimates based on Ind-seq data either by sequencing pools of ca. 200 haploids at 20X, or by sequencing pools of 20 haploids at ca. 200X. However, the same precision and accuracy are achieved by sequencing ca. 50 haploids at ca. 50X.

Conversely, we found that $PP2_d$ (the default estimator of F_{ST} implemented in the software package PoPoolation2) is biased when compared to the expected value of the parameter. We observed that the bias depends on both the sample size, and the coverage (see Figure 2). We note that, as the coverage and the sample size increase, $PP2_d$ converges to the estimator NC_{83} (Nei and Chesser 1983) computed from individual data (see Figure S1B). This argument was used by Kofler et al. (2011) to validate the approach, even though the estimates $PP2_d$ depart from the true value of the parameter (Figure S1B–C).

The second of the two estimators of F_{ST} implemented in PoPoolation2, that we refer to as $PP2_a$, is also biased (see Figure 2). We note that the bias decreases as the sample size increases. However, the bias does not depend on the coverage (only the variance over independent replicates does). The estimator developed by Ferretti et al. (2013), that we refer to as FRP_{13}, is also biased (see Figure 2). However, the bias does neither depend on the pool size, nor on the coverage (only the variance over independent replicates does). FRP_{13} converges to the estimator NC_{83}, computed from individual data (see

Figure 2). At high coverage, the mean and RMSE over independent replicates are virtually indistinguishable from that of the NC_{83} estimator.

Last, we stress out that our estimator $\hat{F}_{ST}^{pool}$ provides estimates for multiple populations, and is therefore not restricted to pairwise analyses, contrary to PoPoolation2's estimators. We show that, even at low sample size and low coverage, Pool-seq estimates of differentiation are virtually indistinguishable from classical estimates for Ind-seq data (see Table 3).

Robustness to unbalanced pool sizes and variable sequencing coverage

We evaluated the accuracy and the precision of the estimator $\hat{F}_{ST}^{pool}$ when sample sizes differ across pools, and when the coverage varies across pools and loci (see Figure 4). We found that, at low coverage, unequal sampling or variable coverage causes a negligible departure from the median of WC_{84} estimates computed on individual data, which vanishes as the coverage increases. At 100X coverage, the distribution of $\hat{F}_{ST}^{pool}$ estimates is almost indistinguishable from that of WC_{84} (see Figure 4 and Tables S2–S3).

Robustness to sequencing and experimental errors

Figure 5 shows that sequencing errors cause a negligible negative bias for $\hat{F}_{ST}^{pool}$ estimates. Filtering (using a minimum read count of 4) improves estimation slightly, but only at high coverage (Figure 6B). It must be noted, though, that filtering increases the bias in the absence of sequencing error, especially at low coverage (Figure 6A). With experimental error, i.e., when individuals do not contribute evenly to the final set of reads, we observed a positive bias for $\hat{F}_{ST}^{pool}$ estimates (Figure 5). We note that the bias decreases

as the size of the pools increases. Figure S2 shows the RMSE of F_{ST} estimates for a wider range of pool sizes, coverage and experimental error rate. For $\epsilon \geq 0.25$, increasing the coverage cannot improve the quality of the inference, if the pool size is too small. When Pool-seq experiments are prone to large experimental error rates, increasing the size of pools is the only way to improve the estimation of F_{ST}. Filtering (using a minimum read count of 4) does not improve estimation (Figure 6C).

Application example

The reanalysis of the prickly sculpin data revealed larger pairwise estimates of multilocus F_{ST} using $\mathrm{PP2_d}$ estimator, as compared to $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ (see Figure 7A). Furthermore, we found that $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ estimates are smaller for within-ecotype pairwise comparisons as compared to between-ecotype comparisons. Therefore, the inferred relationships between samples based on pairwise $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ estimates show a clear-cut structure, separating the two estuarine samples from the freshwater ones (see Figure 7C). We did not recover the same structure using $\mathrm{PP2_d}$ estimates (see Figure 7B). Supportingly, the scaled covariance matrix of allele frequencies across samples is consistent with the structure inferred from $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ estimates (see Figure 7D).

DISCUSSION

Whole-genome sequencing of pools of individuals is being increasingly popular for population genomic research on both model and non-model species (Schlötterer et al. 2014). The development of dedicated software packages (reviewed in Schlötterer et al. 2014) has undoubtedly something to do with the breadth of research questions that have been tackled using pool-sequencing. Yet, the analysis of population structure from Pool-seq data is complicated by the double sampling process of genes from the pool and sequence reads from those genes (Ferretti et al. 2013).

The naive approach that consists in computing F_{ST} from read counts, as if they were allele counts (e.g., as in Chen et al. 2016), ignores the extra variance brought by the random sampling of reads from the gene pool during Pool-seq experiments. Furthermore, such computation fails to consider the actual number of lineages in the pool (haploid pool size). Altogether, these limits may result in severely biased estimates of differentiation when the pool size is low (see Figure S3). A possible alternative is to compute F_{ST} from allele counts imputed from read counts using a maximum-likelihood approach conditional on the haploid size of the pools (e.g., as in Smadja et al. 2012; Leblois et al. 2018), or from allele frequencies estimated using a model-based method that accounts for the sampling effects and the sequencing error probabilities inherent to pooled NGS experiments (see Fariello et al. 2017). However, these latter approaches may only be accurate in situations where the coverage is much larger than pool size, allowing to reduce sampling variance of reads (see Figure S3).

Here, we therefore developed a new estimator of the parameter F_{ST} for

Pool-seq data, in an analysis-of-variance framework (Cockerham 1969, 1973). The accuracy of this estimator is barely distinguishable from that of the Weir and Cockerham's (1984) estimator for individual data. Furthermore, does neither depend on the pool size nor on the coverage, and is robust to unequal pool sizes and varying coverage across demes and loci. In our analysis the frequency of reads within pools is a weighted average of the sample frequencies with weights equal to the pool coverage. Therefore, our approach follows Cockerham's (1973) one, which he referred to as a weighted analysis-of-variance (see also Weir and Cockerham 1984; Weir 1996; Weir and Hill 2002; Weir and Goudet 2017).

With unequal pool sizes, weighted and unweighted analyses differ. As discussed recently in Weir and Goudet (2017), the unweighted approach seems appropriate when the between component exceed the within component, i.e. when F_{ST} is large (Tukey 1957). It turns out that optimal weighting depends upon the parameter to be estimated (Cockerham 1973) and is only efficient at lower levels of differentiation (Robertson 1962). In a likelihood analysis of the island model, Rousset (2007) derived asymptotically efficient weights that are proportional to n_i^2 for the sum of squares of different samples (i.e., as in Robertson 1962). To the best of our knowledge, such optimal weighting has never been considered in the literature. Nevertheless, if these arguments are true for estimators of variance components, they do not necessarily apply to estimates of intra-class correlations (Cockerham 1973).

Analysis of variance and probabilities of identity

In the analysis-of-variance framework, F_{ST} is defined in Equation 1 as an intraclass correlation for the probability of identity in state (Cockerham and

Weir 1987; Rousset 1996). Extensive statistical literature is available on estimators of intraclass correlations. Beside analysis-of-variance estimators, introduced in population genetics by Cockerham (1969, 1973), estimators based on the computation of probabilities of identical response within and between groups have been proposed (see, e.g., Fleiss 1971; Fleiss and Cuzick 1979; Mak 1988; Ridout et al. 1999; Wu et al. 2012), which were originally referred to as kappa-type statistics (Fleiss 1971; Landis and Koch 1977). These estimators have later been endorsed in population genetics, where the "probability of identical response" was then interpreted as the frequency with which the genes are alike (Cockerham 1973; Cockerham and Weir 1987; Weir 1996; Rousset 2007; Weir and Goudet 2017).

This suggests that, with Pool-seq data, another strategy could consist in computing F_{ST} from IIS probabilities between (unobserved) pairs of genes, which requires that unbiased estimates of such quantities are derived from read count data. We have done so in the second section of the Supplemental File S1, and we provide alternative estimators of F_{ST} for Pool-seq data (see Equations A44 and A48 in Supplemental File S1). These estimators (denoted by $\hat{F}_{\mathrm{ST}}^{\mathrm{pool-PID}}$ and $\tilde{F}_{\mathrm{ST}}^{\mathrm{pool-PID}}$) have exactly the same form as the analysis-of-variance estimator if the pools have all the same size and if the number of reads per pool is constant (Equation A33). 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 provide identical estimates when sample sizes are equal (see Equation A28 and also Cockerham and Weir 1987; Weir 1996; Karlsson et al. 2007). With unbalanced samples, we

found that analysis-of-variance estimates have better precision and accuracy than IIS-based estimates, particularly for low levels of differentiation (see Figure S4). Interestingly, we found that IIS-based estimates of F_{ST} for Pool-seq data have generally lower bias and variance if the overall estimates of IIS probabilities within and between pools are computed as unweighted averages of population-specific or pairwise estimates (see Equations A39 and A43), as compared to weighted averages. Equation A28 further shows that our estimator may be rewritten as a function close to $\left(\hat{Q}_1 - \hat{Q}_2\right) / \left(1 - \hat{Q}_2\right)$, except that it also depends on the sums $\sum_i \left(\hat{Q}_{1i} - \hat{Q}_1\right)$ in both the numerator and the denominator. This suggests that if the Q_{1i}'s differ among subpopulations, then our estimator provides an estimate of an average of population-specific F_{ST} (Weir and Hill 2002; Weir and Goudet 2017).

It follows from the derivations in the Supplemental File S1 that the estimator $\mathrm{PP2_a}$ (Equation 19) is biased, because the IIS probability between pairs of reads within a pool $\left(\hat{Q}_1^{\mathrm{r}}\right)$ is a biased estimator of the IIS probability between pairs of distinct genes in that pool (see Equation A34 in Supplemental File S1). 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 that are identical because they were sequenced from distinct, yet IIS genes.

A more justified estimator of F_{ST} has been proposed by Ferretti et al. (2013), based on previous developments by Futschik and Schlötterer (2010). Note that, although they defined F_{ST} as a ratio of functions of heterozygosities, they actually worked with IIS probabilities (see Equations 20 and 21). However, although their Equation 20 is strictly identical to our Equation A34

in Supplemental File S1, we note that they computed the total heterozygosity by integrating over pairs of genes sampled both within and between populations (see Equation 21), which may explain the observed bias (see Figure 2).

Comparison with alternative estimators

An alternative framework to Weir and Cockerham's (1984) analysis-of-variance has been developed by Masatoshi Nei and coworkers to estimate F_{ST} from gene diversities (Nei 1973, 1977; Nei and Chesser 1983; Nei 1986). The estimator $\mathrm{PP2_d}$ (see Equations 16–18) implemented in the software package PoPoolation2 (Kofler et al. 2011) follows this logic. However, it has long been recognized that both frameworks are fundamentally different in that the analysis-of-variance approach considers both statistical and genetic (or evolutionary) sampling, whereas Nei and coworkers' approach do not (Weir and Cockerham 1984; Excoffier 2007; Holsinger and Weir 2009). Furthermore, the expectation of Nei and coworkers' estimators depend upon the number of sampled populations, with a larger bias for lower numbers of sampled populations (Goudet 1993; Excoffier 2007; Weir and Goudet 2017). This is so, because the computation of the total diversity in Equations 18 and 21 includes the comparison of pairs of genes from the same subpopulation, whereas the computation of IIS probabilities between subpopulations do not (see, e.g., Excoffier 2007). Therefore, we do not recommend using the estimator $\mathrm{PP2_d}$ implemented in the software package PoPoolation2 (Kofler et al. 2011).

Applications in evolutionary ecology studies

Pool-seq is being increasingly used in many application domains (Schlötterer et al. 2014), such as conservation genetics (see, e.g., Fuentes-Pardo 2017),

invasion biology (see, e.g., Dexter et al. 2018) and evolutionary biology in a broader sense (see, e.g., Collet et al. 2016). These studies use a large range of methods, which aim at characterizing fine-scaled population structure (see, e.g., Fischer et al. 2017), reconstructing past demography (see, e.g., Chen et al. 2016; Leblois et al. 2018), or identifying footprints of natural or artificial selection (see, e.g., Chen et al. 2016; Fariello et al. 2017; Leblois et al. 2018).

Here, we reanalyzed the Pool-seq data produced by Dennenmoser et al. (2017), who investigated the adaptive genomic divergence between freshwater and brackish-water ecotypes of the prickly sculpin *C. asper*, an abundant euryhaline fish in northwestern North America. Measuring pairwise genetic differentiation between samples using $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$, we found a clear-cut structure separating the freshwater from the brackish-water ecotypes. Such genetic strucure supports the hypothesis that populations are locally adaptated to osmotic conditions in these two contrasted habitats, as discussed in Dennenmoser et al. (2017). This structure, which is at odds with that inferred from $\mathrm{PP2_d}$ estimates, is not only supported by the scaled covariance matrix of allele frequencies, but also by previous microsatellite-based studies, who showed that populations were genetically more differentiated between ecotypes than within ecotypes (Dennenmoser et al. 2014, 2015).

Limits of the model and perspectives

We have shown that the stronger source of bias for the $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ estimate is unequal contributions of individuals in pools. This is so, because we assume in our model that the read counts are multinomially distributed, which supposes that all genes contribute equally to the pool of reads (Gautier et al. 2013), i.e. that there is no variation in DNA yield across individuals and that all

genes have equal sequencing coverage (Rode et al. 2018). Because the effect of unequal contribution is expected to be stronger with small pool sizes, it has been recommended to use pool-seq with at least 50 diploid individuals per pool (Lynch et al. 2014; Schlötterer et al. 2014). However, this limit may be overly conservative for allele frequency estimates (Rode et al. 2018), and we have shown here that we can achieve very good precision and accuracy of F_{ST} estimates with smaller pool sizes. Furthermore, because genotypic information is lost during Pool-seq experiments, we assume in our derivations that pools are haploid (and therefore that F_{IS} is nil). Analyzing non-random mating populations (e.g., in selfing species) is therefore problematic.

Finally, our model, as in Weir and Cockerham (1984), formally assumes that all populations provide independent replicates of some evolutionary process (Excoffier 2007; Holsinger and Weir 2009). This may be unrealistic in many natural populations, which motivated Weir and Hill (2002) to derive a population-specific estimator of F_{ST} for Ind-seq data (see also Vitalis et al. 2001). Even though the use of Weir and Hill's (2002) estimator is still scarce in the literature (but see Weir et al. 2005; Vitalis 2012), Weir and Goudet (2017) recently proposed a re-interpretation of population-specific estimates of F_{ST} in terms of allelic matching proportions, which are strictly equivalent to IIS probabilities between pairs of genes. It would therefore be straightforward to extend Weir and Goudet's (2017) estimator of population-specific F_{ST} for the analysis of Pool-seq data, using the unbiased estimates of IIS probabilies provided in the Supplemental File S1.

DATA ACCESSIBILITY

A R package, called `poolfstat`, which impletements F_{ST} estimates for Pool-seq data, is available at the Comprehensive R Archive Network (CRAN): `https://cran.r-project.org/web/packages/poolfstat/index.html`.

ACKNOWLEDGEMENTS

We thank Alexandre Dehne-Garcia for his assistance in using computer farms. Analyses were performed on the genotoul bioinformatics platform Toulouse Midi-Pyrénées (`bioinfo.genotoul.fr`) and the CBGP HPC computational platform. This work is part of Valentin Hivert's Ph.D., who was supported by a grant from the INRA's Plant Health and Environment (SPE) Division, and by the BiodivERsA project EXOTIC (ANR-13-EBID-0001). Part of this work was supported by the ANR project SWING (ANR-16-CE02-0015) of the French National Research Agency, and by the CORBAM project of the French region Hauts-de-France. We thank two anonymous reviewers for their positive comments and suggestions.

Literature Cited

Akey, J. M., Zhang, G., Jin, L., and Shriver, M. D. (2002). Interrogating a high-density SNP map for signatures of natural selection. *Genome Res.*, 12:1805–1814.

Anderson, E. C., Skaug, H. J., and Barshis, D. J. (2014). Next-generation sequencing for molecular ecology: a caveat regarding pooled samples. *Mol. Ecol.*, 23:502–512.

Beaumont, M. A. (2005). Adaptation and speciation: what can F_{ST} tell us? *Trends Ecol. Evol.*, 20:435–440.

Beaumont, M. A. and Nichols, R. A. (1996). Evaluating loci for use in the genetic analysis of population structure. *Proc. R. Soc. Lond. B*, 263:1619–1626.

Bhatia, G., Patterson, N., Sankararaman, S., and Price, A. L. (2013). Estimating and interpreting F_{ST}: the impact of rare variants. *Genome Res.*, 23:1514–1521.

Cavalli-Sforza, L. (1966). Population structure and human evolution. *Proc. R. Soc. Lond., B, Biol. Sci.*, 164:362–379.

Chen, J., Källman, T., Ma, X.-F., Zaina, G., Morgante, M., and Lascoux, M. (2016). Identifying genetic signatures of natural selection using pooled populations sequencing in *Picea abies*. *G3*, 6:1979–1989.

Cockerham, C. C. (1969). Variance of gene frequencies. *Evolution*, 23:72–84.

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.

Collet, J. M., Fuentes, S., Hesketh, J., Hill, M. S., Innocenti, P., Morrow, E. H., Fowler, K., and Reuter, M. (2016). Rapid evolution of the intersexual genetic correlation for fitness in *Drosophila melanogaster*. *Evolution*, 70:781–795.

Coop, G., Witonsky, D., Di Rienzo, A., and Pritchard, J. K. (2010). Using environmental correlations to identify loci underlying local adaptation. *Genetics*, 185:1411–1423.

Cutler, D. J. and Jensen, J. D. (2010). To pool, or not to pool? *Genetics*, 186:41–43.

Dennenmoser, S., Nolte, A. W., Vamosi, S. M., and Rogers S, M. (2015). Phylogeography of the prickly sculpin (*Cottus asper*) in north-western North America reveals parallel phenotypic evolution across multiple coastal-inland colonizations. *J. Biogeogr.*, 42:1626–1638.

Dennenmoser, S., Rogers, S. M., and Vamosi, S. M. (2014). Genetic population structure in prickly sculpin (*Cottus asper*) reflects isolation-by-environment between two life-history ecotypes. *Biol. J. Linnean Soc.*, 113:943–957.

Dennenmoser, S., Vamosi, S. M., Nolte, S. W., and Rogers, S. M. (2017). Adaptive genomic divergence under high gene flow between freshwater and brackish-water ecotypes of prickly sculpin (*Cottus asper*) revealed by Pool-Seq. *Mol. Ecol.*, 26:25–42.

642 Dexter, E., Bollens, S. M., Cordell, J., Soh, H. Y., Rollwagen-Bollens, G.,
643 Pfeifer, S. P., Goudet, J., and Vuilleumier, S. (2018). A genetic reconstruc-
644 tion of the invasion of the calanoid copepod *Pseudodiaptomus inopinus*
645 across the North American Pacific Coast. *Biol. Invasions*, 20:1577–1595.

646 Ellegren, H. (2014). Genome sequencing and population genomics in non-
647 model organisms. *Trends Ecol. Evol.*, 29:51–63.

648 Excoffier, L. (2007). Analysis of population subdivision. In Balding, D. J.,
649 Bishop, M., and Cannings, C., editors, *Handbook of Statistical Genetics*,
650 pages 980–1020, Chichester. John Wiley & Sons, Ltd.

651 Fariello, M. I. a nd Boitard, S., Mercier, S., Robelin, D., Faraut, T., Arnould,
652 C., Recoquillay, J., Bouchez, O., Salin, G., Dehais, P., Gourichon, D., Ler-
653 oux, S., Pitel, F., Leterrier, C., and SanCristobal, M. (2017). Accounting
654 for Linkage Disequilibrium in genome scans for selection without individual
655 genotypes : the local score approach. *Mol. Ecol.*, 26:3700–3714.

656 Ferretti, L., Ramos Onsins, S., and Pérez-Enciso, M. (2013). Population
657 genomics from pool sequencing. *Mol. Ecol.*, 22:5561–5576.

658 Fischer, M. C., Rellstab, C., Leuzinger, M., Roumet, M., Gugerli, F.,
659 Shimizu, K. K., Holderegger, R., and Widmer, A. (2017). Estimating ge-
660 nomic diversity and population differentiation – an empirical comparison
661 of microsatellite and SNP variation in *Arabidopsis halleri*. *BMC Genomics*,
662 18:69.

663 Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters.
664 *Psychol. Bull.*, 76:378–382.

Fleiss, J. L. and Cuzick, J. (1979). The reliability of dichotomous judgements: Unequal numbers of judges per subject. *Appl. Psychol. Meas.*, 3:537–542.

Fuentes-Pardo, A. P.and Ruzzente, D. E. (2017). Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations. *Mol. Ecol.*, 26:5369–5406.

Futschik, A. and Schlötterer, C. (2010). The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics*, 186:207–218.

Gautier, M. (2015). Genome-wide scan for adaptive divergence and association with population-specific covariates. *Genetics*, 201:1555–1579.

Gautier, M., Gharbi, K., Cezaerd, T., Galan, M., Loiseau, A., Thomson, M., Pudlo, P., Kerdelhué, C., and Estoup, A. (2013). Estimation of population allele frequencies from next-generation sequencing data: pool-versus individual-based genotyping. *Mol. Ecol.*, 22:3766–3779.

Glenn, T. C. (2011). Field guide to next-generation DNA sequencers. *Mol. Ecol. Resour.*, 11:759–769.

Goudet, J. (1993). *The genetics of geographically structured populations*. PhD thesis, University of Wales, Bangor.

Holsinger, K. S. and Weir, B. S. (2009). Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. *Nat. Rev. Genet.*, 10:639–650.

Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18:337–338.

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

694 Kofler, R., Pandey, R. V., and Schlötterer, C. (2011). PoPoolation2: identi-
695 fying differentiation between populations using sequencing of pooled DNA
696 samples (Pool-Seq). *Bioinformatics*, 27:3435–3436.

697 Landis, J. R. and Koch, G. G. (1977). A one-way components of variance
698 model for categorical data. *Biometrics*, 33:671–679.

699 Leblois, R., Gautier, M., Rohfritsch, A., Foucaud, J., Burban, C., Galan, M.,
700 Loiseau, A., Sauné, L., Branco, M., Gharbi, K., Vitalis, R., and Kerdelhué,
701 C. (2018). Deciphering the demographic history of allochronic differentia-
702 tion in the pine processionary moth *Thaumetopoea pityocampa*. *Mol. Ecol.*,
703 27:264–278.

704 Lewontin, R. C. and Krakauer, J. (1973). Distribution of gene frequency as
705 a test of the theory of the selective neutrality of polymorphism. *Genetics*,
706 74:175–195.

707 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth,
708 G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing
709 Subgroup (2009). The Sequence Alignment/Map format and SAMtools.
710 *Bioinformatics*, 25:2078–2079.

711 Lotterhos, K. E. and Whitlock, M. C. (2014). Evaluation of demographic
712 history and neutral parameterization on the performance of F_{ST} outlier
713 tests. *Mol. Ecol.*, 23:2178–2192.

714 Lotterhos, K. E. and Whitlock, M. C. (2015). The relative power of genome
715 scans to detect local adaptation depends on sampling design and statistical
716 method. *Mol. Ecol.*, 24:1031–1046.

717 Lynch, M., Bost, D., Wilson, S., Maruki, T., and Harrison, S. (2014).
718 Population-genetic inference from pooled-sequencing data. *Genome Biol.*
719 *Evol.*, 6:1210–1218.

720 Mak, T. K. (1988). Analysing intraclass correlation for dichotomous vari-
721 ables. *J. R. Stat. Soc. Ser. C Appl. Stat.*, 37:344–352.

722 Malécot, G. (1948). *Les Mathématiques de l'Hérédité*. Masson, Paris.

723 Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proc.*
724 *Natl. Acad. Sci. USA*, 70:3321–3323.

725 Nei, M. (1977). F-statistics and analysis of gene diversity in subdivided
726 populations. *Ann. Hum. Genet.*, 41:225–233.

727 Nei, M. (1978). Estimation of average heterozygosity and genetic distance
728 from a small number of individuals. *Genetics*, 89:583–590.

729 Nei, M. (1986). Definition and estimation of fixation indices. *Evolution*,
730 40:643–645.

731 Nei, M. and Chesser, R. K. (1983). Estimation of fixation indices and gene
732 diversities. *Ann. Hum. Genet.*, 47:253–259.

733 Nychka, D., Furrer, R., Paige, J., and Sain, S. (2017). fields: Tools for spatial
734 data. R package version 9.6.

735 Orgogozo, V., Peluffo, A. E., and Morizot, B. (2016). The "mendelian gene"
736 and the "molecular gene": two relevant concepts of genetic units. In Or-
737 gogozo, V., editor, *Genes and Evolution*, volume 119 of *Current Topics in*
738 *Developmental Biology*, pages 1–26. Academic Press.

739 Pickrell, J. K. and Pritchard, J. K. (2012). Inference of population splits
740 and mixtures from genome-wide allele frequency data. *PLoS Genet.*,
741 8(11):e1002967.

742 R Core Team (2017). *R: A Language and Environment for Statistical Com-*
743 *puting*. R Foundation for Statistical Computing, Vienna, Austria.

744 Reynolds, J., Weir, B. S., and Cockerham, C. C. (1983). Estimation of the
745 coancestry coefficient: basis for a short-term genetic distance. *Genetics*,
746 105:767–779.

747 Ridout, M. S., Demktrio, C. G. B., and Firth, D. (1999). Estimating intra-
748 class correlation for binary data. *Biometrics*, 55:137–148.

749 Robertson, A. (1962). Weighting in the estimation of variance components
750 in the unbalanced single classification. *Biometrics*, 18:413–417.

751 Rode, N. O., Holtz, Y., Loridon, K., Santoni, S., Ronfort, J., and Gay, J.
752 (2018). How to optimize the precision of allele and haplotype frequency
753 estimates using pooled-sequencing data. *Mol. Ecol. Resour.*, 18:194–203.

754 Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty,

R., Nusbaum, C., and Jaffe, D. B. (2013). Characterizing and measuring bias in sequence data. *Genome Biol.*, 14:R51.

Rousset, F. (1996). Equilibrium values of measures of population subdivision for stepwise mutation processes. *Genetics*, 142:1357–1362.

Rousset, F. (1997). Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics*, 145:1219–1228.

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.

Rousset, F. (2008). genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Mol. Ecol. Resour.*, 8:103–106.

Schlötterer, C., Tobler, R., Kofler, R., and Nolte, V. (2014). Sequencing pools of individuals – mining genome-wide polymorphism data without big funding. *Nat. Rev. Genet.*, 15:749–763.

Slatkin, M. (1993). Isolation by distance in equilibrium and non-equilibrium populations. *Evolution*, 47:264–279.

Smadja, C. M., Canbäck, B., Vitalis, R., Gautier, M., Ferrari, J., Zhou, J.-J., and Butlin, R. K. (2012). Large-scale candidate gene scan reveals the role of chemoreceptor genes in host plant specialization and speciation in the pea aphid. *Evolution*, 66:2723–2738.

The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature*, 437:1299–1320.

777 Tukey, J. W. (1957). Variances of variance components: II. The unbalanced
778 single classification. *Ann. Math. Statist.*, 28:43–56.

779 Vitalis, R. (2012). DetSel: An R-Package to detect marker loci responding
780 to selection. In Pompanon, F. and Bonin, A., editors, *Data Production and*
781 *Analysis in Population Genomics: Methods and Protocols*, volume 888 of
782 *Methods in Molecular Biology*, pages 277–293, New York. Humana Press.

783 Vitalis, R., Boursot, P., and Dawson, K. (2001). Interpretation of variation
784 across marker loci as evidence of selection. *Genetics*, 158:1811–1823.

785 Wahlund, S. (1928). Zusammens etzung von populationen und korrelation-
786 serscheinungen vom standpunkt der vererbungslehre aus betrachtet. *Hered-*
787 *itas*, 11:65–106.

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

790 Weir, B. S. (2012). Estimating F-statistics: A historical view. *Philos. Sci.*,
791 79:637–643.

792 Weir, B. S., Cardon, L. R., Anderson, A. D., Nielsen, D. M., and Hill, W. G.
793 (2005). Measures of human population structure show heterogeneity among
794 genomic regions. *Genome Res.*, 15:1468–1476.

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

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

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

Whitlock, M. C. and Lotterhos, K. E. (2015). Reliable detection of loci responsible for local adaptation: inference of a null model through trimming the distribution of F_{ST}. *Am. Nat.*, 186:S24–S36.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, 16:97–159.

Wright, S. (1951). The genetical structure of populations. *Ann. Eugen.*, 15:323–354.

Wu, S., Crespi, C. M., and Wong, W. K. (2012). Comparison of methods for estimating the intraclass correlation coefficient for binary responses in cancer prevention cluster randomized trials. *Contemp. Clin. Trials*, 33:869–880.

Table 1 Summary of main notations

Notation	Parameter definition
$X_{ijr:k}$	Indicator variable: $X_{ijr:k} = 1$ if the rth read from the jth individual in the ith pool is of type k, and $X_{ijr:k} = 0$ otherwise
$r_{i:k} = \sum_j \sum_r X_{ijr:k}$	Number of reads of type k in the ith pool
c_{ij}	Number of reads sequenced from individual j in subpopulation i (unobserved individual coverage)
$C_{1i} \equiv \sum_j c_{ij}$	Total number of reads in the ith pool (pool coverage)
$C_1 \equiv \sum_i C_{1i}$	Total number of reads in the full sample (total coverage)
$C_2 \equiv \sum_i C_{1i}^2$	Squared number of reads in the full sample
n_i	Total number of genes the ith pool (haploid pool size)
$y_{i:k}$	(Unobserved) number of genes of type k in the ith pool
$\pi_k \equiv \mathbb{E}(X_{ijr:k})$	Expected frequency of reads of type k in the full sample
$\hat{\pi}_{ij:k} \equiv X_{ij\cdot:k}$	(Unobserved) average frequency of reads of type k for individual j in the ith pool
$\hat{\pi}_{i:k} \equiv X_{i\cdots:k}$	Average frequency of reads of type k in the ith pool
$\hat{\pi}_k \equiv X_{\cdots:k}$	Average frequency of reads of type k in the full sample
Q_1 (resp. Q_2)	IIS probability for two genes sampled within (resp. between) pools
Q_1^r (resp. Q_2^r)	IIS probability for two reads sampled within (resp. between) pools
$\hat{Q}_1^{\text{pool}}$ (resp. $\hat{Q}_2^{\text{pool}}$)	Unbiased estimator of the IIS probability for genes sampled within (resp. between) populations

Table 2 Definition of the F_{ST} estimators used in the text

Notation	Definition
$\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$	Equation 9
FRP_{13}	Ferretti et al. (2013) and Equations 16,20–21
NC_{83}	Nei and Chesser (1983)
$\mathrm{PP2}_{\mathrm{d}}$	Kofler et al. (2011) and Equations 16–18
$\mathrm{PP2}_{\mathrm{a}}$	Kofler et al. (2011) and Equation 19
WC_{84}	Weir and Cockerham (1984)

Table 3 Overall F_{ST} estimates from multiple pools

		Pool-seq		Ind-seq
F_{ST}	n	Cov.	$\hat{F}_{ST}^{pool}$	WC_{84}
0.05	10	20×	0.050 (0.002)	
0.05	10	50×	0.051 (0.002)	0.050 (0.002)
0.05	10	100×	0.050 (0.002)	
0.05	100	20×	0.050 (0.001)	
0.05	100	50×	0.050 (0.001)	0.051 (0.001)
0.05	100	100×	0.050 (0.001)	
0.20	10	20×	0.200 (0.002)	
0.20	10	50×	0.201 (0.002)	0.201 (0.002)
0.20	10	100×	0.201 (0.002)	
0.20	100	20×	0.201 (0.003)	
0.20	100	50×	0.202 (0.003)	0.203 (0.003)
0.20	100	100×	0.203 (0.003)	

Overall F_{ST} was estimated for various conditions of expected F_{ST}, pool size (n) and coverage (Cov.). For Pool-seq data, we computed our estimator $\hat{F}_{ST}^{pool}$ (Equation 13). The mean (RMSE) over 50 independent replicates of the `ms` simulations are provided, for all populations ($n_d = 8$). For comparison, we computed WC_{84} from allele count data inferred from individual genotypes (Ind-seq).

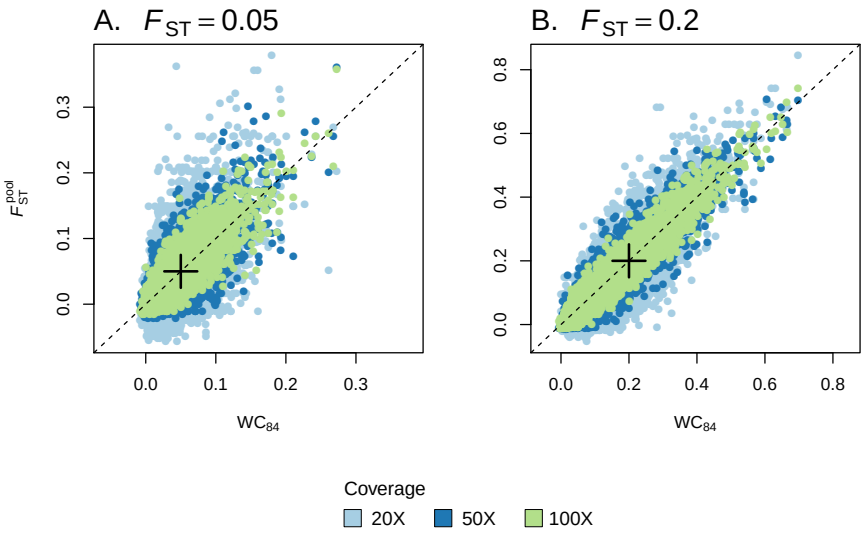

Figure 1 Single-locus estimates of F_{ST}. We compared single-locus estimates of F_{ST} based on allele count data inferred from individual genotypes (Ind-seq), using the WC_{84} estimator, to $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ estimates from Pool-seq data. We simulated 5,000 SNPs using `ms` in an island model with $n_{\mathrm{d}} = 8$ demes. We used two migration rates corresponding to $F_{\mathrm{ST}} = 0.05$ (A) and $F_{\mathrm{ST}} = 0.20$ (B). The size of each pool was fixed to 100. We show the results for different coverages (20X, 50X and 100X). In each graph, the cross indicates the simulated value of F_{ST}.

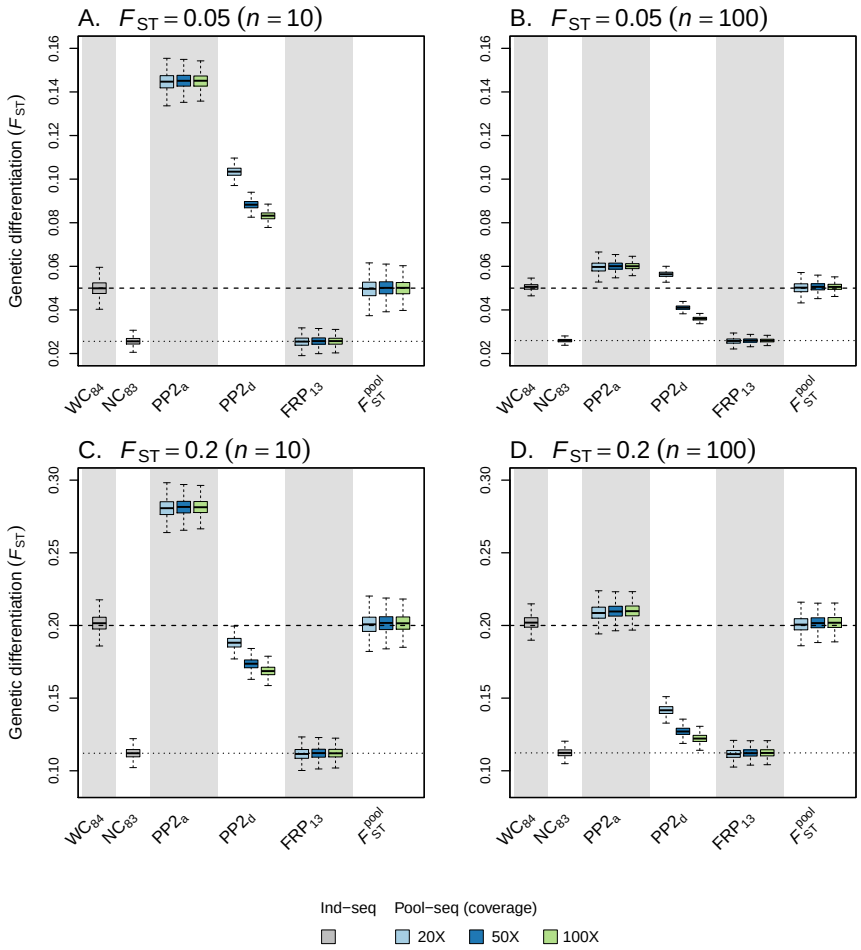

Figure 2 Precision and accuracy of pairwise estimators of F_{ST}. We considered two estimators based on allele count data inferred from individual genotypes (Ind-seq): WC_{84} and NC_{83}. For pooled data, we computed the two estimators implemented in the software package PoPoolation2, that we refer to as $\mathrm{PP2}_{\mathrm{d}}$ and $\mathrm{PP2}_{\mathrm{a}}$, as well as the FRP_{13} estimator and our estimator $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ (Equation 13). Each boxplot represents the distribution of multilocus F_{ST} estimates across all pairwise comparisons in an island model with $n_{\mathrm{d}} = 8$ demes, and across 50 independent replicates of the `ms` simulations. We used two migration rates, corresponding to $F_{\mathrm{ST}} = 0.05$ (A–B) or $F_{\mathrm{ST}} = 0.20$ (C–D). The size of each pool was either fixed to 10 (A and C) or to 100 (B and D). For Pool-seq data, we show the results for different coverages (20X, 50X and 100X). In each graph, the dashed line indicates the simulated value of F_{ST} and the dotted line indicates the median of the distribution of NC_{83} estimates.

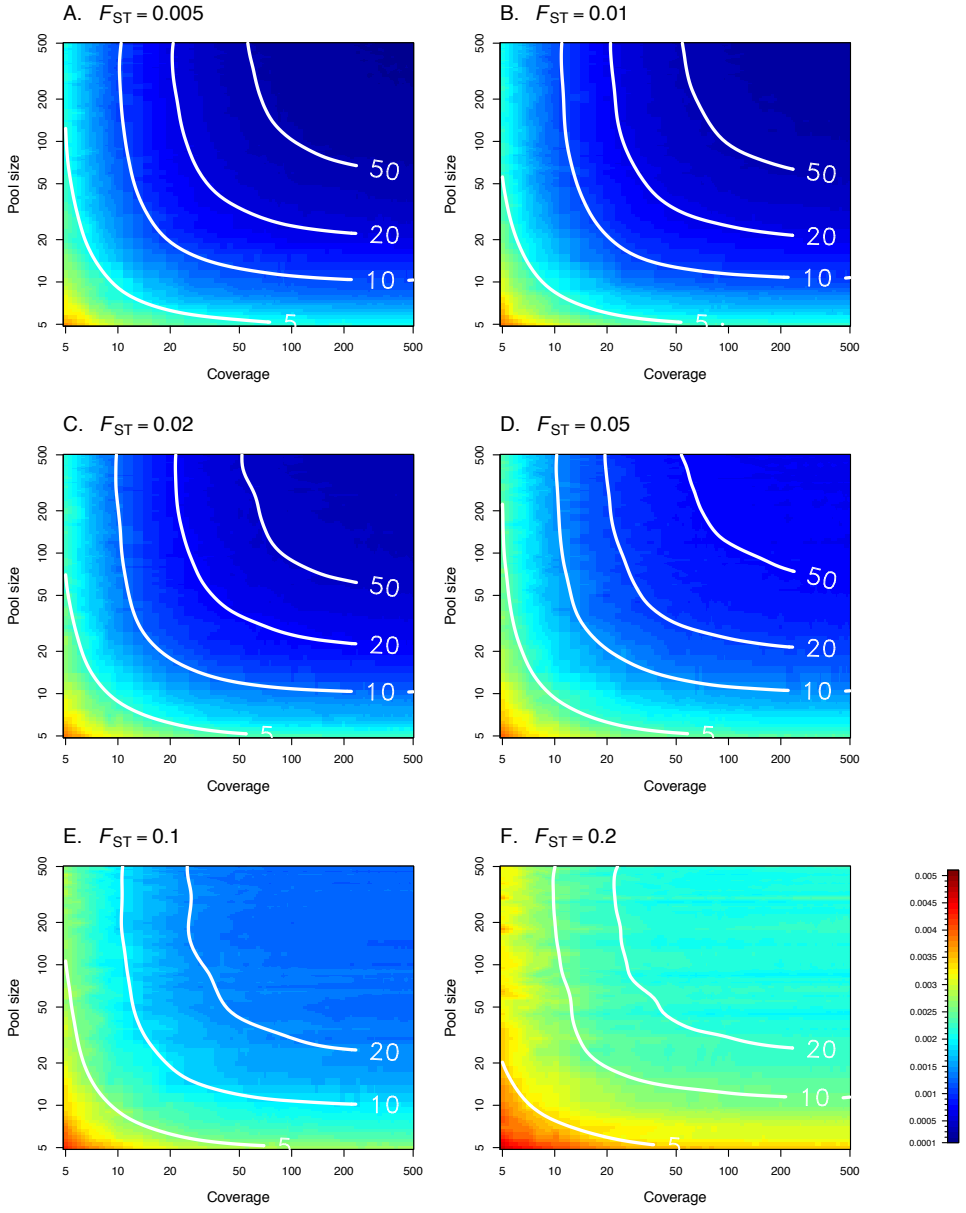

Figure 3 Root mean squared error (RMSE) of F_{ST} estimates for a wide range of pool sizes and coverage, with F_{ST} varying from 0.005 to 0.2 (A–F). Each density plot gives the RMSE of our estimator $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$, using simple linear interpolation from a set of 44×44 pairs of pool size and coverage values. For each pool size and coverage, 500 replicates of 5,000 markers were simulated. Plain white isolines represent the RMSE of the WC_{84} estimator computed from Ind-seq data, for various sample sizes ($n = 5$, 10, 20, and 50). Each isoline was fitted using a thin plate spline regression with smoothing parameter $\lambda = 0.005$, implemented in the `fields` package for R (Nychka et al. 2017).

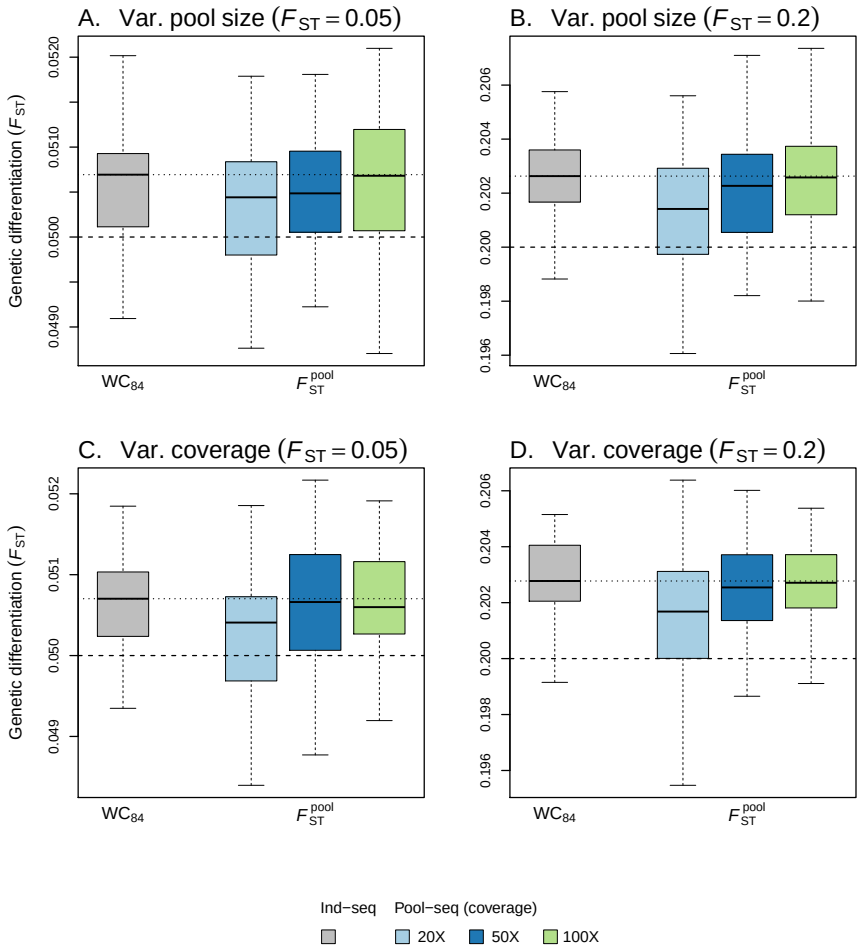

Figure 4 Precision and accuracy of F_{ST} estimates with varying pool size or varying coverage. Our estimator $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ (Equation 13) was calculated from Pool-seq data over all loci and demes and compared to the estimator WC_{84}, computed from allele count data inferred from individual genotypes (Ind-seq). Each boxplot represents the distribution of multilocus F_{ST} estimates across 50 independent replicates of the `ms` simulations. We used two migration rates, corresponding to $F_{\mathrm{ST}} = 0.05$ (A and C) or $F_{\mathrm{ST}} = 0.20$ (B and D). In A–B the pool size was variable across demes, with haploid sample size n drawn independently for each deme from a Gaussian distribution with mean 100 and standard deviation 30; n was rounded up to the nearest integer, with min. 20 and max. 300 haploids per deme. In C–D, the pool size was fixed ($n = 100$), and the coverage (δ_i) was varying across demes and loci, with $\delta_i \sim \mathrm{Pois}(\Delta)$ where $\Delta \in \{20, 50, 100\}$. For Pool-seq data, we show the results for different coverages (20X, 50X and 100X). In each graph, the dashed line indicates the simulated value of F_{ST} and the dotted line indicates the median of the distribution of WC_{84} estimates.

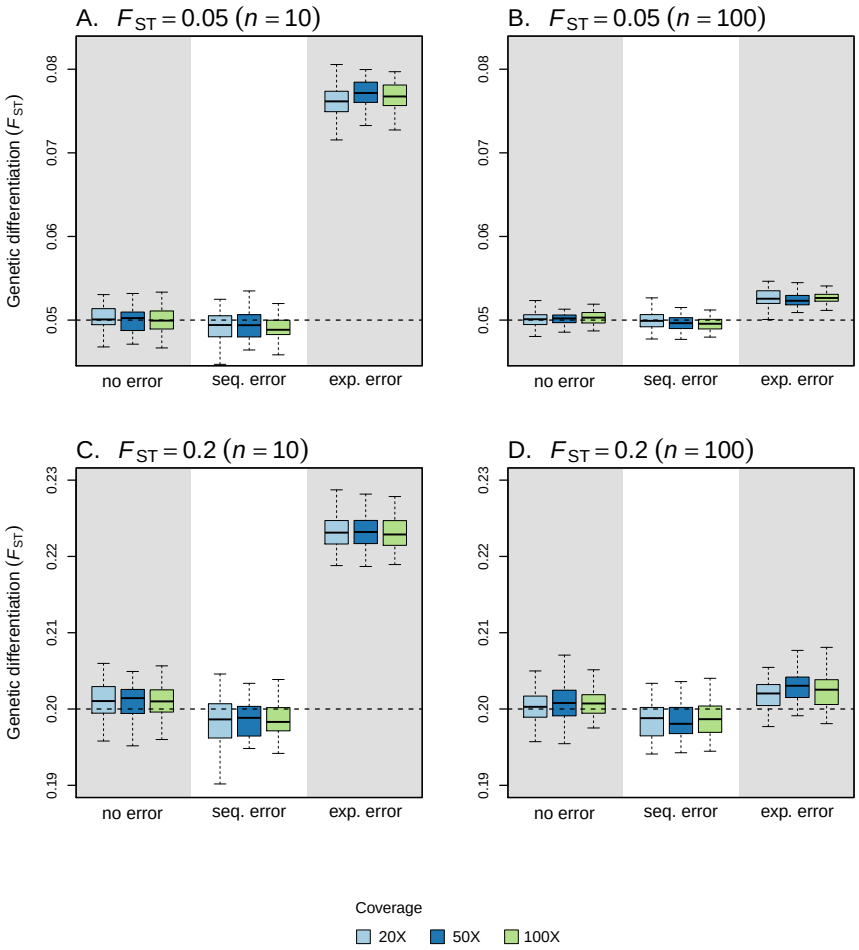

Figure 5 Precision and accuracy of F_{ST} estimates with sequencing and experimental errors. Our estimator $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ (Equation 13) was computed from Pool-seq data over all loci and demes without error, with sequencing error (occurring at rate $\mu_{\mathrm{e}} = 0.001$), and with experimental error ($\epsilon = 0.5$). Each boxplot represents the distribution of multilocus F_{ST} estimates across 50 independent replicates of the `ms` simulations. We used two migration rates, corresponding to $F_{\mathrm{ST}} = 0.05$ (A–B) or $F_{\mathrm{ST}} = 0.20$ (C–D). The size of each pool was either fixed to 10 (A and C) or to 100 (B and D). For Pool-seq data, we show the results for different coverages (20X, 50X and 100X). In each graph, the dashed line indicates the simulated value of F_{ST}.

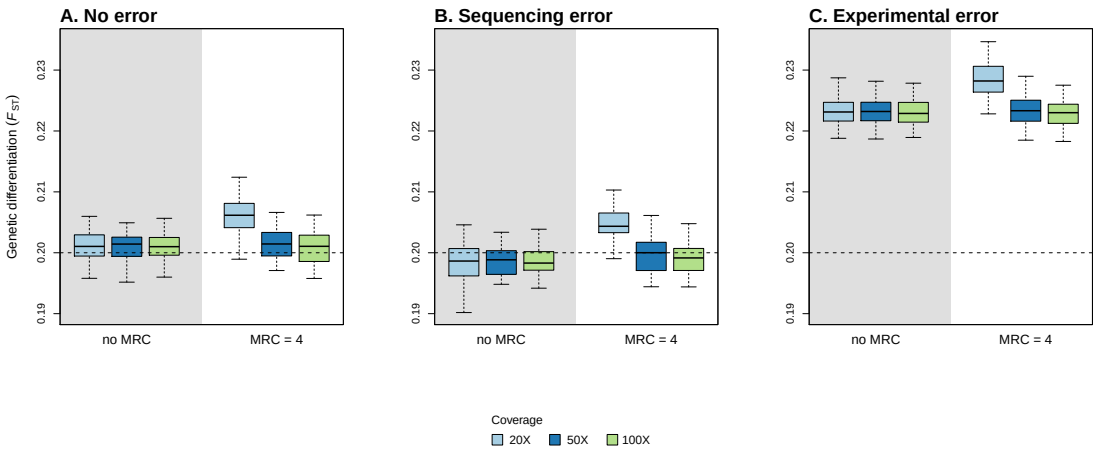

Figure 6 Precision and accuracy of F_{ST} estimates with and without filtering. Our estimator $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ (Equation 13) was computed from Pool-seq data over all loci and demes without error (A), with sequencing error (B) and with experimental error (C) (see the legend of Figure 5 for further details). For each case, we computed F_{ST} without filtering (no MRC) and with filtering (using a minimum read count MRC = 4). Each boxplot represents the distribution of multilocus F_{ST} estimates across 50 independent replicates of the `ms` simulations. We used a migration rate corresponding to $F_{\mathrm{ST}} = 0.20$, and pool size $n = 10$. We show the results for different coverages (20X, 50X and 100X). In each graph, the dashed line indicates the simulated value of F_{ST}.

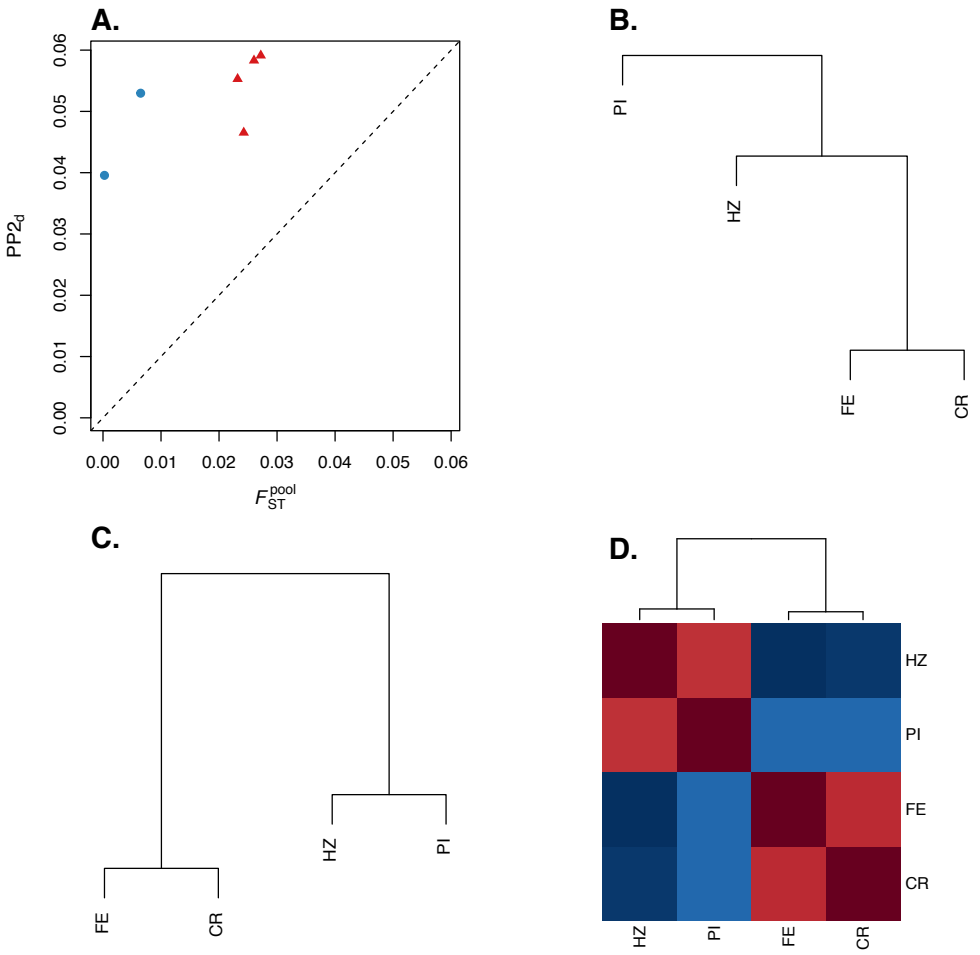

Figure 7 Analysis of the prickly sculpin (*Cottus asper*) Pool-seq data. In (A) we compare the pairwise F_{ST} estimates $\mathrm{PP2_d}$, and $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ (Equation 13) for all pairs of populations from the estuarine (CR and FE) and freshwater samples (PI and HZ). Within-ecotype comparisons are depicted as blue dots, and between-ecotype comparisons as red triangles. In (B–C) we show a UPGMA hierarchical cluster analyses based on $\mathrm{PP2_d}$ (B) and $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ (C) pairwise estimates. In (D), we show a heatmap representation of the scaled covariance matrix among the four *C. asper* populations, inferred from the Bayesian hierarchical model implemented in the software package BayPass.

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 $C_1 \equiv \sum_i \sum_j c_{ij}$ 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 frequency 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:

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

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 $C_2 \equiv \sum_i C_{1i}^2$.

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

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

856 Hence, developing the first term in the right-hand side of Equation A3 using

857 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:

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

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:

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

879 and from Equations A14 and A16:

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

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

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

882 and for genes between individuals from different pools:

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

883 Rearranging Equations A19–A20, we get:

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

884 and:

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

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, 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:

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

and:

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

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:

$$\hat{Q}_1^{\text{pool}} \equiv \frac{\sum_i \hat{Q}_{1i}^{\text{pool}}}{n_{\text{d}}} \tag{A39}$$

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

939 gives:

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

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

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

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

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

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

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

947 which, to the extent that we may take the expectation of a ratio to be the
948 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).

Table S1 Comparison of pairwise F_{ST} estimates

		Pool-seq		Ind-seq
F_{ST}	n	Cov.	$\hat{F}_{ST}^{pool}$	WC_{84}
0.05	10	20×	0.051 (0.004)	
0.05	10	50×	0.051 (0.004)	0.051 (0.003)
0.05	10	100×	0.051 (0.003)	
0.05	100	20×	0.051 (0.003)	
0.05	100	50×	0.051 (0.003)	0.051 (0.002)
0.05	100	100×	0.051 (0.002)	
0.20	10	20×	0.203 (0.007)	
0.20	10	50×	0.202 (0.006)	0.202 (0.007)
0.20	10	100×	0.201 (0.006)	
0.20	100	20×	0.201 (0.006)	
0.20	100	50×	0.201 (0.006)	0.201 (0.005)
0.20	100	100×	0.202 (0.005)	

F_{ST} was estimated for various conditions of expected F_{ST}, pool size (n) and coverage (Cov.). For Pool-seq data, we computed our estimator $\hat{F}_{ST}^{pool}$ (Equation 13). The mean (RMSE) over 50 independent replicates of the `ms` simulations are provided for a single pair of populations. For comparison, we computed WC_{84} from allele count data inferred from individual genotypes (Ind-seq).

Table S2 Effect of unequal sampling on pairwise F_{ST} estimates

		Pool-seq		Ind-seq
F_{ST}	n	Cov.	$\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$	WC_{84}
0.05	$\mathcal{N}(100, 30)$	20×	0.051 (0.003)	
0.05	$\mathcal{N}(100, 30)$	50×	0.052 (0.003)	0.051 (0.002)
0.05	$\mathcal{N}(100, 30)$	100×	0.051 (0.002)	
0.20	$\mathcal{N}(100, 30)$	20×	0.202 (0.007)	
0.20	$\mathcal{N}(100, 30)$	50×	0.202 (0.006)	0.202 (0.006)
0.20	$\mathcal{N}(100, 30)$	100×	0.202 (0.006)	

Pairwise F_{ST} was estimated for various conditions of expected F_{ST} and coverage (Cov.). The pool size (n) was variable across demes, with haploid sample size n drawn independently for each deme from a Gaussian distribution with mean 100 and standard deviation 30; n was rounded up to the nearest integer, with min. 20 and max. 300 haploids per deme. For Pool-seq data, we computed our estimator $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ (Equation 13). The mean (RMSE) over 50 independent replicates of the `ms` simulations are provided, for a single pair of populations. For comparison, we computed WC_{84} (Weir and Cockerham 1984) from allele count data inferred from individual genotypes (Ind-seq).

Table S3 Effect of variable coverage on pairwise F_{ST} estimates

		Pool-seq		Ind-seq
F_{ST}	n	Δ	$\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$	WC_{84}
0.05	10	20	0.050 (0.006)	
0.05	10	50	0.050 (0.004)	0.050 (0.004)
0.05	10	100	0.050 (0.004)	
0.05	100	20	0.051 (0.003)	
0.05	100	50	0.051 (0.002)	0.051 (0.002)
0.05	100	100	0.051 (0.002)	
0.20	10	20	0.200 (0.007)	
0.20	10	50	0.200 (0.007)	0.200 (0.007)
0.20	10	100	0.200 (0.007)	
0.20	100	20	0.202 (0.006)	
0.20	100	50	0.203 (0.006)	0.203 (0.005)
0.20	100	100	0.203 (0.005)	

Pairwise F_{ST} was estimated for various conditions of expected F_{ST} and pool size (n). The coverage (δ_i) was varying across demes and loci, with $\delta_i \sim \mathrm{Pois}(\Delta)$. For Pool-seq data, we computed our estimator $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ (Equation 13). The mean (RMSE) over 50 independent replicates of the `ms` simulations are provided, for a single pair of populations. For comparison, we computed WC_{84} from allele count data inferred from individual genotypes (Ind-seq).

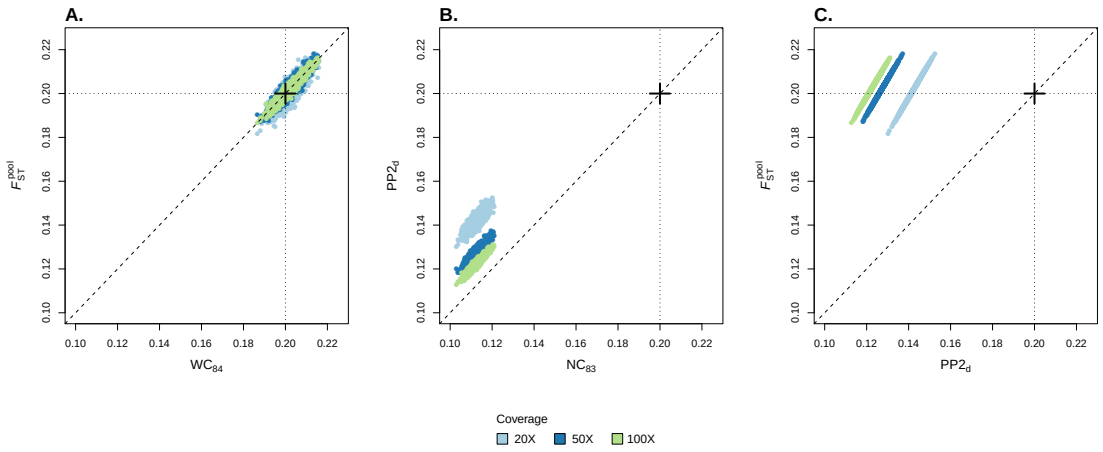

Figure S1 Pairwise estimators of F_{ST}. A. Multilocus estimates $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ (computed using Equation 13) as a function of WC_{84} estimates computed from allele count data inferred from individual genotypes. B. Multilocus estimates $\mathrm{PP2_d}$, as a function of NC_{83} estimates computed from allele count data inferred from individual genotypes. C. Multilocus estimates $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ as a function of $\mathrm{PP2_d}$ estimates. In each graph, the dots represent multilocus estimates of F_{ST} across all pairs of subpopulations from an 8-island model, and across 50 replicate `ms` simulations. We specified the migration rate corresponding to $F_{\mathrm{ST}} = 0.20$. The size of each pool was fixed to 100. The results are shown for different coverages (20X, 50X and 100X). The cross indicates the simulated value of the parameter F_{ST}.

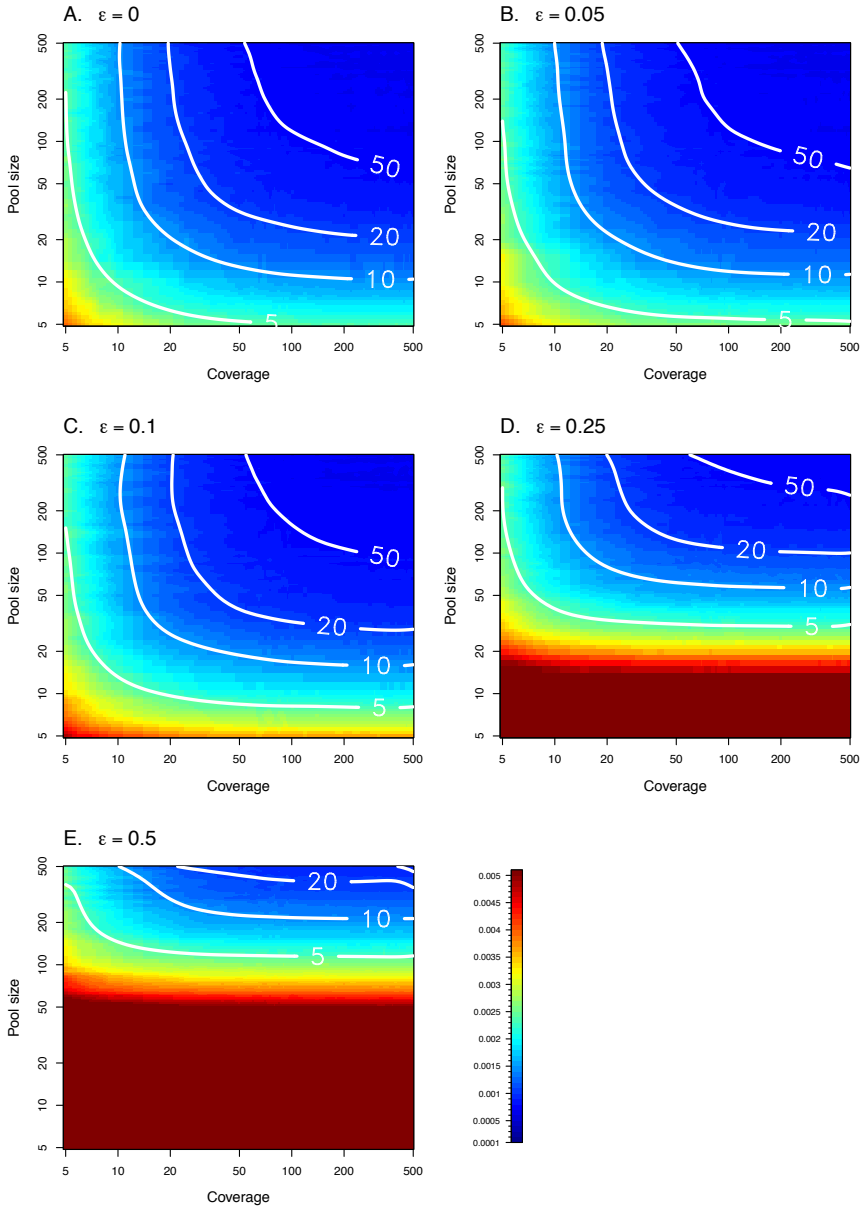

Figure S2 Root mean squared error (RMSE) of F_{ST} estimates for a wide range of pool sizes and coverage, with experimental error rate ϵ varying from 0 to 0.5 (A–E). Each density plot gives the RMSE of our estimator $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$, using simple linear interpolation from a set of 44×44 pairs of pool size and coverage values. For each pool size and coverage, 500 replicates of 5,000 markers were simulated. Plain white isolines represent the RMSE of the WC_{84} estimator computed from Ind-seq data, for various sample sizes (n = 5, 10, 20, and 50). Each isoline was fitted using a thin plate spline regression with smoothing parameter $\lambda = 0.005$, implemented in the `fields` package for R (Nychka et al. 2017).

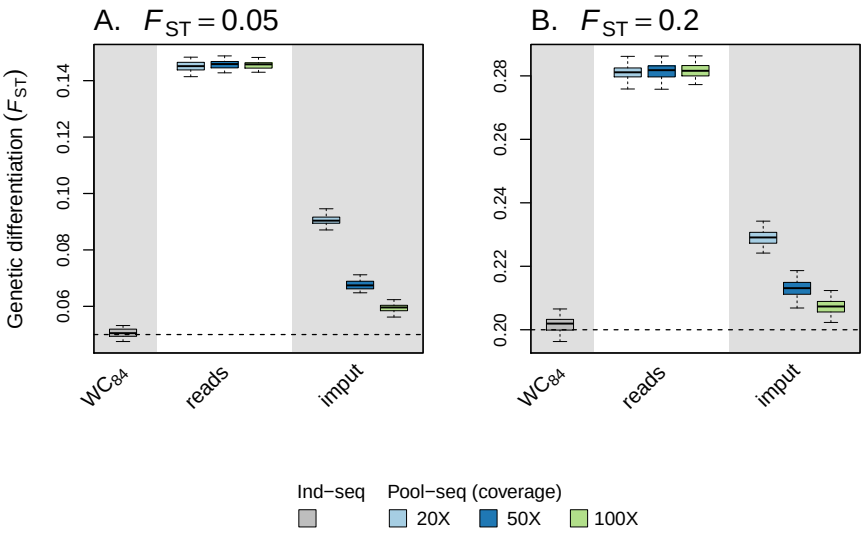

Figure S3 Global estimators of F_{ST}. We considered one estimator based on allele count data inferred from individual genotypes (Ind-seq): WC_{84}. For pooled data, we computed F_{ST} using the WC_{84} estimator: (i) directly from read counts, as if they were allele counts ("reads"); (ii) from allele counts imputed by maximum-likelihood ("imput"), as in Leblois et al. (2018). Each boxplot represents the distribution of multilocus F_{ST} estimates across all demes comparisons in an 8-island model, and across 50 independent replicates of the `ms` simulations. We used two migration rates, corresponding to $F_{\mathrm{ST}} = 0.05$ (A) or $F_{\mathrm{ST}} = 0.20$ (B). The size of each pool was fixed to 10. For Pool-seq data, we show the results for different coverages (20X, 50X and 100X). In each graph, the dashed line indicates the simulated value of F_{ST}.

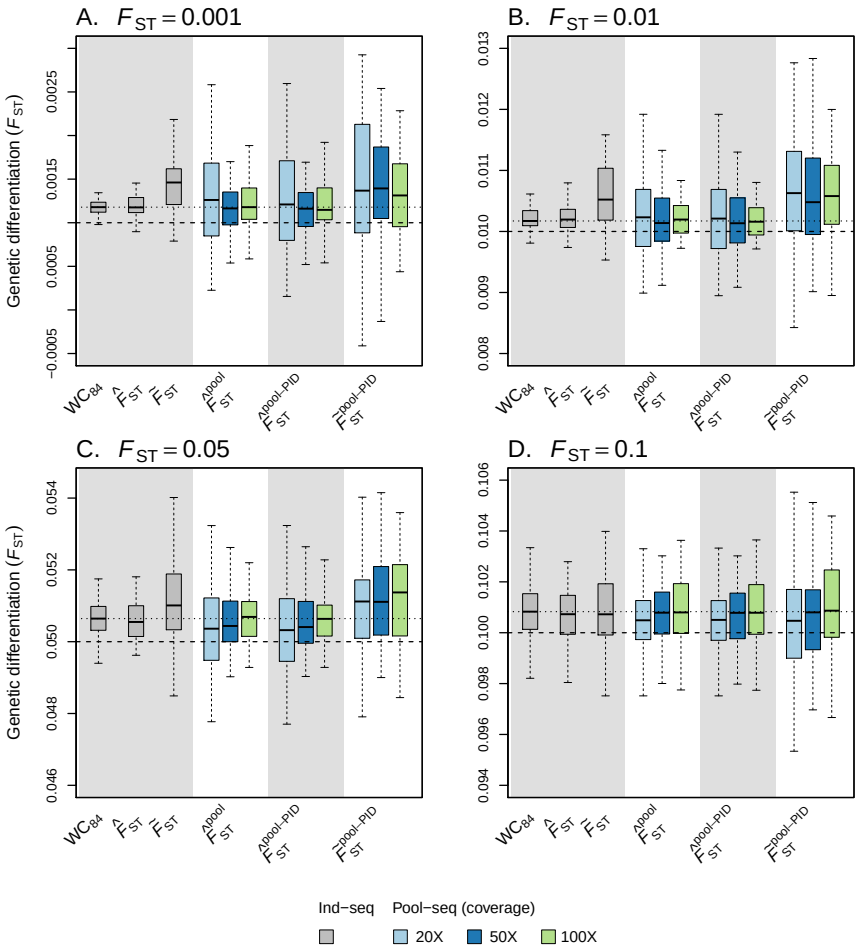

Figure S4 Precision and accuracy of alternative estimators of F_{ST} with varying pool size, for various levels of differentiation (A–D). The haploid pool size n drawn independently for each deme from a Gaussian distribution with mean 100 and standard deviation 30; n was rounded up to the nearest integer, with min. 20 and max. 300 haploids per deme. We considered three estimators based on allele count data inferred from individual genotypes (Ind-seq): WC_{84}, $\hat{F}_{\mathrm{ST}} \equiv \left(\hat{Q}_1 - \hat{Q}_2\right) / \left(1 - \hat{Q}_2\right)$ (where $\hat{Q}_1$ and $\hat{Q}_2$ are the weighted frequencies of identical pairs of genes within and between subpopulations, respectively, with weights equal to the number of pairs of genes) and $\tilde{F}_{\mathrm{ST}} \equiv \left(\tilde{Q}_1 - \tilde{Q}_2\right) / \left(1 - \tilde{Q}_2\right)$ (where $\tilde{Q}_1$ and $\tilde{Q}_2$ are the unweighted frequencies of identical pairs of genes within and between subpopulations, respectively. For Pool-seq data, we considered the estimators $\hat{F}^{\mathrm{pool}}_{\mathrm{ST}}$ (Equation 12), $\hat{F}^{\mathrm{pool-PID}}_{\mathrm{ST}}$ (Equation A44) and $\tilde{F}^{\mathrm{pool-PID}}_{\mathrm{ST}}$ (Equation A45). Each boxplot represents the distribution of multilocus F_{ST} across 50 independent replicates of the `ms` simulations. For Pool-seq data, we show the results for different coverages (20X, 50X and 100X). In each graph, the dashed line indicates the simulated value of F_{ST} and the dotted line indicates the median of the distribution of WC_{84} estimates.