



**HAL**  
open science

## Measuring Genetic Differentiation from Pool-seq Data

Valentin Hivert, Raphael Leblois, Eric Petit, Mathieu Gautier, Renaud Vitalis

► **To cite this version:**

Valentin Hivert, Raphael Leblois, Eric Petit, Mathieu Gautier, Renaud Vitalis. Measuring Genetic Differentiation from Pool-seq Data. *Genetics*, 2018, 210 (1), pp.315-330. 10.1534/genetics.118.300900 . hal-01936905

**HAL Id: hal-01936905**

**<https://hal.science/hal-01936905v1>**

Submitted on 26 May 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

## Measuring genetic differentiation from Pool-seq data

Valentin Hivert<sup>\*,†</sup>, Raphaël Leblois<sup>\*,†</sup>, Eric J. Petit<sup>‡</sup>, Mathieu  
Gautier<sup>\*,†,§</sup>, and Renaud Vitalis<sup>\*,†,§</sup>

<sup>\*</sup>CBGP, INRA, CIRAD, IRD, Montpellier SupAgro, Univ Montpellier,  
Montpellier, France

<sup>†</sup>Institut de Biologie Computationnelle, Univ Montpellier, Montpellier, France

<sup>‡</sup>ESE, Ecology and Ecosystem Health, INRA, Agrocampus Ouest, Rennes, France

<sup>§</sup>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](mailto: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

21

22 It has long been recognized that the subdivision of species into subpopu-  
23 lations, social groups and families fosters genetic differentiation (Wahlund  
24 1928; Wright 1931). Characterizing genetic differentiation as a means to infer  
25 unknown population structure is therefore fundamental to population genet-  
26 ics, and finds applications in multiple domains, including conservation biol-  
27 ogy, invasion biology, association mapping and forensics, among many others.  
28 In the late 1940s and early 1950s, Malécot (1948) and Wright (1951) intro-  
29 duced  $F$ -statistics to partition genetic variation within and between groups  
30 of individuals (Holsinger and Weir 2009; Bhatia et al. 2013). Since then, the  
31 estimation of  $F$ -statistics has become standard practice (see, e.g., Weir 1996;  
32 Weir and Hill 2002; Weir 2012), and the most commonly used estimators of  
33  $F_{ST}$  have been developed in an analysis-of-variance framework (Cockerham  
34 1969, 1973; Weir and Cockerham 1984), which can be recast in terms of prob-  
35 abilities of identity of pairs of homologous genes (Cockerham and Weir 1987;  
36 Rousset 2007; Weir and Goudet 2017).

37 Assuming that molecular markers are neutral, estimates of  $F_{ST}$  are typ-  
38 ically used to quantify genetic structure in natural populations, which is  
39 then interpreted as the result of demographic history (Holsinger and Weir  
40 2009): large  $F_{ST}$  values are expected for small populations among which  
41 dispersal is limited (Wright 1951), or between populations that have long  
42 diverged in isolation from each other (Reynolds et al. 1983); when dispersal  
43 is spatially restricted, a positive relationship between  $F_{ST}$  and the geograph-  
44 ical distance for pairs of populations generally holds (Slatkin 1993; Rousset  
45 1997). It has also been proposed to characterize the heterogeneity of  $F_{ST}$

46 estimates across markers for identifying loci that are targeted by selection  
47 (Cavalli-Sforza 1966; Lewontin and Krakauer 1973; Beaumont and Nichols  
48 1996; Vitalis et al. 2001; Akey et al. 2002; Beaumont 2005; Weir et al. 2005;  
49 Lotterhos and Whitlock 2014, 2015; Whitlock and Lotterhos 2015).

50 Next-generation sequencing (NGS) technologies provide unprecedented  
51 amounts of polymorphism data in both model and non-model species (Elle-  
52 gren 2014). Although the sequencing strategy initially involved individually  
53 tagged samples in humans (The International HapMap Consortium 2005),  
54 whole-genome sequencing of pools of individuals (Pool-seq) is being increas-  
55 ingly used for population genomic studies (Schlötterer et al. 2014). Because  
56 it consists in sequencing libraries of pooled DNA samples and does not re-  
57 quire individual tagging of sequences, Pool-seq provides genome-wide poly-  
58 morphism data at considerably lower cost than sequencing of individuals  
59 (Schlötterer et al. 2014). However, non-equimolar amounts of DNA from all  
60 individuals in a pool and stochastic variation in the amplification efficiency  
61 of individual DNAs have raised concerns with respect to the accuracy of the  
62 so-obtained allele frequency estimates, particularly at low sequencing depth  
63 and with small pool sizes (Cutler and Jensen 2010; Ellegren 2014; Anderson  
64 et al. 2014). Nonetheless, it has been shown that, at equal sequencing effort,  
65 Pool-seq provides similar, if not more accurate, allele frequency estimates  
66 than individual-based analyses (Futschik and Schlötterer 2010; Gautier et al.  
67 2013). The problem is different for diversity and differentiation parameters,  
68 which depend on second moments of allele frequencies or, equivalently, on  
69 pairwise measures of genetic identity: with Pool-seq data, it is indeed im-  
70 possible to distinguish pairs of reads that are identical because they were

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

74 Appropriate estimators of diversity and differentiation parameters must  
75 therefore be sought, to account for both the sampling of individual genes  
76 from the pool and the sampling of reads from these genes. There has been  
77 several attempts to define estimators for the parameter  $F_{ST}$  for Pool-seq data  
78 (Kofler et al. 2011; Ferretti et al. 2013), from ratios of heterozygosities (or  
79 from probabilities of genetic identity between pairs of reads) within and be-  
80 tween pools. In the following, we will argue that these estimators are biased  
81 (i.e., they do not converge towards the expected value of the parameter),  
82 and that some of them have undesired statistical properties (i.e., the bias  
83 depends upon sample size and coverage). Here, following Cockerham (1969),  
84 Cockerham (1973), Weir and Cockerham (1984), Weir (1996), Weir and Hill  
85 (2002) and Rousset (2007), we define a method-of-moments estimator of the  
86 parameter  $F_{ST}$  using an analysis-of-variance framework. We then evaluate  
87 the accuracy and the precision of this estimator, based on the analysis of sim-  
88 ulated datasets, and compare it to estimates defined in the software package  
89 PoPoolation2 (Kofler et al. 2011), and in Ferretti et al. (2013). Furthermore,  
90 we test the robustness of our estimators to model misspecifications (including  
91 unequal contributions of individuals in pools, and sequencing errors). Finally,  
92 we reanalyze the prickly sculpin (*Cottus asper*) Pool-seq data (published by  
93 Dennenmoser et al. 2017), and show how the use of biased  $F_{ST}$  estimators in  
94 previous analyses may challenge the interpretation of population structure.

95 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, for which we further assume that  
100 individual genotypes are called without error.



102  $F$ -statistics may be described as intra-class correlations for the probability of  
 103 identity in state (IIS) of pairs of genes (Cockerham and Weir 1987; Rousset  
 104 1996, 2007), and  $F_{ST}$  is best defined as:

$$F_{ST} \equiv \frac{Q_1 - Q_2}{1 - Q_2} \quad (1)$$

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

110 For the sake of clarity, the notation used throughout this article is given in  
 111 Table 1. We first derive our model for a single locus, and eventually provide  
 112 a multilocus estimator of  $F_{ST}$ . Consider a sample of  $n_d$  subpopulations, each  
 113 of which is made of  $n_i$  genes ( $i = 1, \dots, n_d$ ) sequenced in pools (hence  $n_i$  is  
 114 the haploid sample size of the  $i$ th pool). We define  $c_{ij}$  as the number of reads  
 115 sequenced from gene  $j$  ( $j = 1, \dots, n_i$ ) in subpopulation  $i$  at the locus consid-  
 116 ered. Note that  $c_{ij}$  is a latent variable, that cannot be directly observed from  
 117 the data. Let  $X_{ijr:k}$  be an indicator variable for read  $r$  ( $r = 1, \dots, c_{ij}$ ) from  
 118 gene  $j$  in subpopulation  $i$ , such that  $X_{ijr:k} = 1$  if the  $r$ th read from the  $j$ th  
 119 gene in the  $i$ th deme is of type  $k$ , and  $X_{ijr:k} = 0$  otherwise. In the following,  
 120 we use standard dot notations for sample averages, i.e.:  $X_{ij:k} \equiv \sum_r X_{ijr:k}/c_{ij}$ ,  
 121  $X_{i:k} \equiv \sum_j \sum_r X_{ijr:k}/\sum_j c_{ij}$  and  $X_{:k} \equiv \sum_i \sum_j \sum_r X_{ijr:k}/\sum_i \sum_j c_{ij}$ . The  
 122 analysis of variance is based on the computation of sums of squares, as fol-

123 lows:

$$\begin{aligned}
\sum_i^{n_d} \sum_j^{n_i} \sum_r^{c_{ij}} (X_{ijr:k} - X_{\dots:k})^2 &= \sum_i^{n_d} \sum_j^{n_i} \sum_r^{c_{ij}} (X_{ijr:k} - X_{ij:k})^2 \\
&+ \sum_i^{n_d} \sum_j^{n_i} \sum_r^{c_{ij}} (X_{ij:k} - X_{i\cdot:k})^2 \\
&+ \sum_i^{n_d} \sum_j^{n_i} \sum_r^{c_{ij}} (X_{i\cdot:k} - X_{\dots:k})^2 \\
&\equiv SSR_{:k} + SSI_{:k} + SSP_{:k} \tag{2}
\end{aligned}$$

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

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

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

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

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

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

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

139 where  $C_2 \equiv \sum_i C_{1i}^2$  and  $D_2^* \equiv [\sum_i C_{1i} (C_{1i} + n_i - 1) / n_i] / C_1$  (see Equa-  
 140 tion A16 in Supplemental File S1). Rearranging Equations 4–5, and summing  
 141 over alleles, we get:

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

142 and:

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

143 where  $n_c \equiv (C_1 - C_2/C_1) / (D_2 - D_2^*)$ . Let  $MSI \equiv SSI / (C_1 - D_2)$  and  
 144  $MSP \equiv SSP / (D_2 - D_2^*)$ . Then, using the definition of  $F_{ST}$  from Equation 1,  
 145 we have:

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

146 which yields the method-of-moments estimator:

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

147 where

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

148 and:

$$MSP = \frac{1}{D_2 - D_2^*} \sum_k \sum_i^{n_d} C_{1i} (\hat{\pi}_{i:k} - \hat{\pi}_k)^2 \quad (11)$$

149 (see Equations A25 and A26 in Supplemental File S1). In Equations 10  
 150 and 11,  $\hat{\pi}_{i:k} \equiv X_{i\cdots:k}$  is the average frequency of reads of type  $k$  within the  $i$ th  
 151 pool, and  $\hat{\pi}_k \equiv X_{\cdots:k}$  is the average frequency of reads of type  $k$  in the full sam-  
 152 ple. Note that from the definition of  $X_{\cdots:k}$ ,  $\hat{\pi}_k \equiv \sum_i \sum_j \sum_r X_{ijr:k} / \sum_i \sum_j c_{ij} =$   
 153  $\sum_i C_{1i} \hat{\pi}_{i:k} / \sum_i C_{1i}$  is the weighted average of the sample frequencies with  
 154 weights equal to the pool coverage. This is equivalent to the weighted  
 155 analysis-of-variance in Cockerham (1973) (see also Weir and Cockerham 1984;  
 156 Weir 1996; Weir and Hill 2002; Rousset 2007; Weir and Goudet 2017). Fi-  
 157 nally, the full expression of  $\hat{F}_{ST}^{\text{pool}}$  in terms of sample frequencies reads:

$$\hat{F}_{ST}^{\text{pool}} = \frac{\sum_k [(C_1 - D_2) \sum_i^{n_d} C_{1i} (\hat{\pi}_{i:k} - \hat{\pi}_k)^2 - (D_2 - D_2^*) \sum_i^{n_d} C_{1i} \hat{\pi}_{i:k} (1 - \hat{\pi}_{i:k})]}{\sum_k [(C_1 - D_2) \sum_i^{n_d} C_{1i} (\hat{\pi}_{i:k} - \hat{\pi}_k)^2 + (n_c - 1) (D_2 - D_2^*) \sum_i^{n_d} C_{1i} \hat{\pi}_{i:k} (1 - \hat{\pi}_{i:k})]} \quad (12)$$

158 If we take the limit case where each gene is sequenced exactly once, we  
 159 recover the Ind-seq model: assuming  $c_{ij} = 1$  for all  $(i, j)$ , then  $C_1 = \sum_i^{n_d} n_i$ ,  
 160  $C_2 = \sum_i^{n_d} n_i^2$ ,  $D_2 = n_d$  and  $D_2^* = 1$ . Therefore,  $n_c = (C_1 - C_2/C_1) / (n_d - 1)$ ,  
 161 and Equation 9 reduces exactly to the estimator of  $F_{ST}$  for haploids: see Weir  
 162 (1996), p. 182, and Rousset (2007), p. 977.

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

$$\hat{F}_{ST} = \frac{\sum_l MSP_l - MSI_l}{\sum_l MSP_l + (n_c - 1) MSI_l} \quad (13)$$

166 where  $MSI$  and  $MSP$  are subscripted with  $l$  to denote the  $l$ th locus. For  
167 Ind-seq data, Bhatia et al. (2013) refer to this multilocus estimate as a “ratio  
168 of averages” by opposition to an “average of ratios”, which would consist in av-  
169 eraging single-locus  $F_{ST}$  over loci. This approach is justified in the Appendix  
170 of Weir and Cockerham (1984) and in Bhatia et al. (2013), who analyzed  
171 both estimates by means of coalescent simulations. Note that Equation 13  
172 assumes that the pool size is equal across loci. Also note that the construc-  
173 tion of the estimator in Equation 13 is different from Weir and Cockerham’s  
174 (1984). These authors defined their multilocus estimator as a ratio of sums  
175 of components of variance ( $a$ ,  $b$  and  $c$  in their notation) over loci, which give  
176 the same weight to all loci, whatever the number of sampled genes at each lo-  
177 cus. Equation 13 follows GENEPOP’s rationale (Rousset 2008) instead, which  
178 gives instead more weight to loci that are more intensively covered.

180 **Simulation study**

181 *Generating individual genotypes:* we first generated individual genotypes us-  
 182 ing `ms` (Hudson 2002), assuming an island model of population structure  
 183 (Wright 1931). For each simulated scenario, we considered 8 demes, each  
 184 made of  $N = 5,000$  haploid individuals. The migration rate ( $m$ ) was fixed  
 185 to achieve the desired value of  $F_{ST}$  (0.05 or 0.2), using Equation 6 in Rousset  
 186 (1996) leading, e.g., to  $M \equiv 2Nm = 16.569$  for  $F_{ST} = 0.05$  and  $M = 3.489$  for  
 187  $F_{ST} = 0.20$ . The mutation rate was set at  $\mu = 10^{-6}$ , giving  $\theta \equiv 2N\mu = 0.01$ .  
 188 We considered either fixed, or variable sample sizes across demes. In the lat-  
 189 ter case, the haploid sample size  $n$  was drawn independently for each deme  
 190 from a Gaussian distribution with mean 100 and standard deviation 30; this  
 191 number was rounded up to the nearest integer, with min. 20 and max. 300  
 192 haploids per deme. We generated a very large number of sequences for each  
 193 scenario, and sampled independent single nucleotide polymorphisms (SNPs)  
 194 from sequences with a single segregating site. Each scenario was replicated  
 195 50 times (500 times for Figures 3 and S2).

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

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

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

205 *Sequencing error:* we simulated sequencing errors occurring at rate  $\mu_e =$   
206 0.001, which is typical of Illumina sequencers (Glenn 2011; Ross et al. 2013).  
207 We assumed that each sequencing error modifies the allelic type of a read to  
208 one of three other possible states with equal probability (there are therefore  
209 four allelic types in total, corresponding to four nucleotides). Note that  
210 only biallelic markers are retained in the final datasets. Also note that,  
211 since we initiated this procedure with polymorphic markers only, we neglect  
212 sequencing errors that would create spurious SNPs from monomorphic sites.  
213 However, such SNPs should be rare in real datasets, since markers with a  
214 low minimum read count (MRC) are generally filtered out.

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

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

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

### 231 **Other estimators**

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

234 PP2<sub>d</sub> : this estimator of  $F_{ST}$  is implemented by default in the software  
235 package POPOOLATION2 (Kofler et al. 2011). It is based on a definition of  
236 the parameter  $F_{ST}$  as the overall reduction in average heterozygosity relative  
237 to the total combined population (see, e.g., Nei and Chesser 1983):

$$\text{PP2}_d \equiv \frac{\hat{H}_T - \hat{H}_S}{\hat{H}_T} \quad (16)$$

238 where  $\hat{H}_S$  is the average heterozygosity within subpopulations, and  $\hat{H}_T$  is the  
239 average heterozygosity in the total population (obtained by pooling together  
240 all subpopulation to form a single virtual unit). In POPOOLATION2,  $\hat{H}_S$  is



241 the unweighted average of within-subpopulation heterozygosities:

$$\hat{H}_S = \frac{1}{n_d} \sum_i^{n_d} \left( \frac{n_i}{n_i - 1} \right) \left( \frac{C_{1i}}{C_{1i} - 1} \right) \left( 1 - \sum_k \hat{\pi}_{i:k}^2 \right) \quad (17)$$

242 (using the notation from Table 1). Note that in POPOOLATION2, PP2<sub>d</sub> is  
243 restricted to the case of two subpopulations only ( $n_d = 2$ ). The two ratios in  
244 the right-hand side of Equation 17 are presumably borrowed from Nei (1978)  
245 to provide an unbiased estimate, although we found no formal justification  
246 for the expression in Equation 17 for Pool-seq data. The total heterozygosity  
247 is computed as (using the notation from Table 1):

$$\hat{H}_T = \left( \frac{\min_i(n_i)}{\min_i(n_i) - 1} \right) \left( \frac{\min_i(C_{1i})}{\min_i(C_{1i}) - 1} \right) \left( 1 - \sum_k \hat{\pi}_k^2 \right) \quad (18)$$

248 PP2<sub>a</sub> : this is the alternative estimator of  $F_{ST}$  provided in the software  
249 package POPOOLATION2. It is based on an interpretation by Kofler et al.  
250 (2011) of Karlsson et al.'s (2007) estimator of  $F_{ST}$ , as:

$$\text{PP2}_a \equiv \frac{\hat{Q}_1^r - \hat{Q}_2^r}{1 - \hat{Q}_2^r} \quad (19)$$

251 where  $\hat{Q}_1^r$  and  $\hat{Q}_2^r$  are the frequencies of identical pairs of reads within and  
252 between pools, respectively, computed by simple counting of IIS pairs. These  
253 are estimates of  $Q_1^r$ , the IIS probability for two reads in the same pool  
254 (whether they are sequenced from the same gene or not) and  $Q_2^r$ , the IIS  
255 probability for two reads in different pools. Note that the IIS probability  $Q_1^r$   
256 is different from  $Q_1$  in Equation 1, which, from our definition, represents  
257 the IIS probability between distinct genes in the same pool. This approach  
258 therefore confounds pairs of reads within pools that are identical because

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

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

$$\hat{H}_S = \frac{1}{n_d} \sum_i^{n_d} \left( \frac{n_i}{n_i - 1} \right) (1 - \hat{Q}_{1i}^r) \quad (20)$$

268 and the total heterozygosity among the  $n_d$  populations as:

$$\hat{H}_T = \frac{1}{n_d^2} \left[ \sum_i^{n_d} \left( \frac{n_i}{n_i - 1} \right) (1 - \hat{Q}_{1i}^r) + \sum_{i \neq i'}^{n_d} (1 - \hat{Q}_{2ii'}^r) \right] \quad (21)$$

### 269 **Analyses of Ind-seq data:**

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

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

279 ware environment for statistical computing, version 3.3.1 (R Core Team  
280 2017). All these functions were carefully checked against available software  
281 packages, to ensure that they provided strictly identical estimates.

### 282 **Application example: *Cottus asper***

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

302 Our aim here was to compare the population structure inferred from pair-  
303 wise estimates of  $F_{ST}$ , using the estimator  $\hat{F}_{ST}^{pool}$  on the one hand (Equa-

304 tion 12), and PP2<sub>d</sub> on the other hand. Then, to conclude on which of the  
305 two estimators performs better, we compared the population structure in-  
306 ferred from  $\hat{F}_{ST}^{pool}$  and PP2<sub>d</sub> to that inferred from the Bayesian hierarchical  
307 model implemented in the software package BAYPASS (Gautier 2015). BAY-  
308 PASS allows the robust estimation of the scaled covariance matrix of allele  
309 frequencies across populations for Pool-seq data, which is known to be infor-  
310 mative about population history (Pickrell and Pritchard 2012). The elements  
311 of the estimated matrix can be interpreted as pairwise and population-specific  
312 estimates of differentiation (Coop et al. 2010), and therefore provide a com-  
313 prehensive description of population structure that makes full use of the  
314 available data.

### 315 **Data availability**

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

## RESULTS

321

### 322 **Comparing Ind-seq and Pool-seq estimates of $F_{ST}$**

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

### 329 **Comparing Pool-seq estimators of $F_{ST}$**

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

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

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

353 Conversely, we found that PP2<sub>d</sub> (the default estimator of  $F_{ST}$  imple-  
354 mented in the software package POPOOLATION2) is biased when compared  
355 to the expected value of the parameter. We observed that the bias depends  
356 on both the sample size, and the coverage (see Figure 2). We note that, as the  
357 coverage and the sample size increase, PP2<sub>d</sub> converges to the estimator NC<sub>83</sub>  
358 (Nei and Chesser 1983) computed from individual data (see Figure S1B).  
359 This argument was used by Kofler et al. (2011) to validate the approach,  
360 even though the estimates PP2<sub>d</sub> depart from the true value of the parameter  
361 (Figure S1B–C).

362 The second of the two estimators of  $F_{ST}$  implemented in POPOOLATION2,  
363 that we refer to as PP2<sub>a</sub>, is also biased (see Figure 2). We note that the bias  
364 decreases as the sample size increases. However, the bias does not depend  
365 on the coverage (only the variance over independent replicates does). The  
366 estimator developed by Ferretti et al. (2013), that we refer to as FRP<sub>13</sub>, is  
367 also biased (see Figure 2). However, the bias does neither depend on the pool  
368 size, nor on the coverage (only the variance over independent replicates does).  
369 FRP<sub>13</sub> converges to the estimator NC<sub>83</sub>, computed from individual data (see

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

372 Last, we stress out that our estimator  $\hat{F}_{ST}^{pool}$  provides estimates for multiple  
373 populations, and is therefore not restricted to pairwise analyses, contrary to  
374 POPOOLATION2's estimators. We show that, even at low sample size and low  
375 coverage, Pool-seq estimates of differentiation are virtually indistinguishable  
376 from classical estimates for Ind-seq data (see Table 3).

### 377 **Robustness to unbalanced pool sizes and variable sequencing cov-** 378 **erage**

379 We evaluated the accuracy and the precision of the estimator  $\hat{F}_{ST}^{pool}$  when sam-  
380 ple sizes differ across pools, and when the coverage varies across pools and loci  
381 (see Figure 4). We found that, at low coverage, unequal sampling or variable  
382 coverage causes a negligible departure from the median of  $WC_{84}$  estimates  
383 computed on individual data, which vanishes as the coverage increases. At  
384 100X coverage, the distribution of  $\hat{F}_{ST}^{pool}$  estimates is almost indistinguishable  
385 from that of  $WC_{84}$  (see Figure 4 and Tables S2–S3).

### 386 **Robustness to sequencing and experimental errors**

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

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

#### 401 **Application example**

402 The reanalysis of the prickly sculpin data revealed larger pairwise estimates of  
403 multilocus  $F_{ST}$  using PP2<sub>d</sub> estimator, as compared to  $\hat{F}_{ST}^{pool}$  (see Figure 7A).  
404 Furthermore, we found that  $\hat{F}_{ST}^{pool}$  estimates are smaller for within-ecotype  
405 pairwise comparisons as compared to between-ecotype comparisons. There-  
406 fore, the inferred relationships between samples based on pairwise  $\hat{F}_{ST}^{pool}$  esti-  
407 mates show a clear-cut structure, separating the two estuarine samples from  
408 the freshwater ones (see Figure 7C). We did not recover the same structure  
409 using PP2<sub>d</sub> estimates (see Figure 7B). Supportingly, the scaled covariance  
410 matrix of allele frequencies across samples is consistent with the structure  
411 inferred from  $\hat{F}_{ST}^{pool}$  estimates (see Figure 7D).



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

421 The naive approach that consists in computing  $F_{ST}$  from read counts, as  
422 if they were allele counts (e.g., as in Chen et al. 2016), ignores the extra  
423 variance brought by the random sampling of reads from the gene pool dur-  
424 ing Pool-seq experiments. Furthermore, such computation fails to consider  
425 the actual number of lineages in the pool (haploid pool size). Altogether,  
426 these limits may result in severely biased estimates of differentiation when  
427 the pool size is low (see Figure S3). A possible alternative is to compute  $F_{ST}$   
428 from allele counts imputed from read counts using a maximum-likelihood  
429 approach conditional on the haploid size of the pools (e.g., as in Smadja  
430 et al. 2012; Leblois et al. 2018), or from allele frequencies estimated using a  
431 model-based method that accounts for the sampling effects and the sequenc-  
432 ing error probabilities inherent to pooled NGS experiments (see Fariello et al.  
433 2017). However, these latter approaches may only be accurate in situations  
434 where the coverage is much larger than pool size, allowing to reduce sam-  
435 pling variance of reads (see Figure S3). Here, we therefore developed a new  
436 estimator of the parameter  $F_{ST}$  for Pool-seq data, in an analysis-of-variance

437 framework (Cockerham 1969, 1973). The accuracy of this estimator is barely  
438 distinguishable from that of the Weir and Cockerham's (1984) estimator for  
439 individual data. Furthermore, it does neither depend on the pool size nor on  
440 the coverage, and is robust to unequal pool sizes and varying coverage across  
441 demes and loci.

442 In our analysis, the frequency of reads within pools is a weighted av-  
443 erage of the sample frequencies, with weights equal to the pool coverage.  
444 Therefore, our approach follows Cockerham's (1973) one, which he referred  
445 to as a weighted analysis-of-variance (see also Weir and Cockerham 1984;  
446 Weir 1996; Weir and Hill 2002; Weir and Goudet 2017). With unequal pool  
447 sizes, weighted and unweighted analyses differ. As discussed recently in Weir  
448 and Goudet (2017), the unweighted approach seems appropriate when the  
449 between component exceeds the within component, i.e. when  $F_{ST}$  is large  
450 (Tukey 1957). It turns out that optimal weighting depends upon the param-  
451 eter to be estimated (Cockerham 1973) and is only efficient at lower levels of  
452 differentiation (Robertson 1962). In a likelihood analysis of the island model,  
453 Rousset (2007) derived asymptotically efficient weights that are proportional  
454 to  $n_i^2$  for the sum of squares of different samples (see also Robertson 1962). To  
455 the best of our knowledge, such optimal weighting has never been considered  
456 in the literature.

## 457 **Analysis of variance and probabilities of identity**

458 In the analysis-of-variance framework,  $F_{ST}$  is defined in Equation 1 as an  
459 intraclass correlation for the probability of identity in state (Cockerham and  
460 Weir 1987; Rousset 1996). Extensive statistical literature is available on  
461 estimators of intraclass correlations. Beside analysis-of-variance estimators,

462 introduced in population genetics by Cockerham (1969, 1973), estimators  
463 based on the computation of probabilities of identical response within and  
464 between groups have been proposed (see, e.g., Fleiss 1971; Fleiss and Cuzick  
465 1979; Mak 1988; Ridout et al. 1999; Wu et al. 2012), which were originally  
466 referred to as kappa-type statistics (Fleiss 1971; Landis and Koch 1977).  
467 These estimators have later been endorsed in population genetics, where the  
468 “probability of identical response” was then interpreted as the frequency with  
469 which the genes are alike (Cockerham 1973; Cockerham and Weir 1987; Weir  
470 1996; Rousset 2007; Weir and Goudet 2017).

471 This suggests that, with Pool-seq data, another strategy could consist in  
472 computing  $F_{ST}$  from IIS probabilities between (unobserved) pairs of genes,  
473 which requires that unbiased estimates of such quantities are derived from  
474 read count data. We have done so in the second section of the Supplemental  
475 File S1, and we provide alternative estimators of  $F_{ST}$  for Pool-seq data (see  
476 Equations A44 and A48 in Supplemental File S1). These estimators (denoted  
477 by  $\hat{F}_{ST}^{\text{pool-PID}}$  and  $\tilde{F}_{ST}^{\text{pool-PID}}$ ) have exactly the same form as the analysis-of-  
478 variance estimator if the pools have all the same size and if the number of  
479 reads per pool is constant (Equation A33). This echoes the derivations by  
480 Rousset (2007) for Ind-seq data, who showed that the analysis-of-variance  
481 approach (Weir and Cockerham 1984) and the simple strategy of estimat-  
482 ing IIS probabilities by counting identical pairs of genes provide identical  
483 estimates when sample sizes are equal (see Equation A28 and also Cock-  
484 erham and Weir 1987; Weir 1996; Karlsson et al. 2007). With unbalanced  
485 samples, we found that analysis-of-variance estimates have better precision  
486 and accuracy than IIS-based estimates, particularly for low levels of differ-

487 entiation (see Figure S4). Interestingly, we found that IIS-based estimates  
488 of  $F_{ST}$  for Pool-seq data have generally lower bias and variance if the over-  
489 all estimates of IIS probabilities within and between pools are computed as  
490 unweighted averages of population-specific or pairwise estimates (see Equa-  
491 tions A39 and A43), as compared to weighted averages (Equations A46–A47).  
492 Equation A28 further shows that our estimator may be rewritten as a func-  
493 tion close to  $(\hat{Q}_1 - \hat{Q}_2) / (1 - \hat{Q}_2)$ , except that it also depends on the sum  
494  $\sum_i (\hat{Q}_{1i} - \hat{Q}_1)$  in both the numerator and the denominator. This suggests  
495 that if the  $Q_{1i}$ 's differ among subpopulations, then our estimator provides an  
496 estimate of an average of population-specific  $F_{ST}$  (Weir and Hill 2002; Weir  
497 and Goudet 2017).

498 It follows from the derivations in the Supplemental File S1 that the es-  
499 timator PP2<sub>a</sub> (Equation 19) is biased because the IIS probability between  
500 pairs of reads within a pool ( $\hat{Q}_1^r$ ) is a biased estimator of the IIS probability  
501 between pairs of distinct genes in that pool (see Equations A34–A36 in Sup-  
502 plemental File S1). This is so, because the former confounds pairs of reads  
503 that are identical because they were sequenced from a single gene, from pairs  
504 of reads that are identical because they were sequenced from distinct, yet IIS  
505 genes.

506 A more justified estimator of  $F_{ST}$  has been proposed by Ferretti et al.  
507 (2013), based on previous developments by Futschik and Schlötterer (2010).  
508 Note that, although they defined  $F_{ST}$  as a ratio of functions of heterozygosi-  
509 ties, they actually worked with IIS probabilities (see Equations 20 and 21).  
510 However, although Equation 20 is strictly identical to Equation A39 in Sup-  
511 plemental File S1, we note that they computed the total heterozygosity by

512 integrating over pairs of genes sampled both within and between subpopula-  
513 tions (compare Equation 21 with A43), which may explain the observed bias  
514 (see Figure 2).

### 515 **Comparison with alternative estimators**

516 An alternative framework to Weir and Cockerham's (1984) analysis-of-variance  
517 has been developed by Masatoshi Nei and coworkers to estimate  $F_{ST}$  from  
518 gene diversities (Nei 1973, 1977; Nei and Chesser 1983; Nei 1986). The es-  
519 timator  $PP2_d$  (see Equations 16–18) implemented in the software package  
520 `POPOOLATION2` (Kofler et al. 2011) follows this logic. However, it has long  
521 been recognized that both frameworks are fundamentally different in that the  
522 analysis-of-variance approach considers both statistical and genetic (or evo-  
523 lutionary) sampling, whereas Nei and coworkers' approach do not (Weir and  
524 Cockerham 1984; Excoffier 2007; Holsinger and Weir 2009). Furthermore,  
525 the expectation of Nei and coworkers' estimators depend upon the number  
526 of sampled populations, with a larger bias for lower numbers of sampled pop-  
527 ulations (Goudet 1993; Excoffier 2007; Weir and Goudet 2017). This is so,  
528 because the computation of the total diversity in Equations 18 and 21 includes  
529 the comparison of pairs of genes from the same subpopulation, whereas the  
530 computation of IIS probabilities between subpopulations do not (see, e.g.,  
531 Excoffier 2007). Therefore, we do not recommend using the estimator  $PP2_d$   
532 implemented in the software package `POPOOLATION2` (Kofler et al. 2011).

### 533 **Applications in evolutionary ecology studies**

534 Pool-seq is being increasingly used in many application domains (Schlöt-  
535 terer et al. 2014), such as conservation genetics (see, e.g., Fuentes-Pardo and

536 Ruzzente 2017), invasion biology (see, e.g., Dexter et al. 2018) and evolution-  
537 ary biology in a broader sense (see, e.g., Collet et al. 2016). These studies  
538 use a large range of methods, which aim at characterizing fine-scaled popula-  
539 tion structure (see, e.g., Fischer et al. 2017), reconstructing past demography  
540 (see, e.g., Chen et al. 2016; Leblois et al. 2018), or identifying footprints of  
541 natural or artificial selection (see, e.g., Chen et al. 2016; Fariello et al. 2017;  
542 Leblois et al. 2018).

543 Here, we reanalyzed the Pool-seq data produced by Dennenmoser et al.  
544 (2017), who investigated the adaptive genomic divergence between freshwa-  
545 ter and brackish-water ecotypes of the prickly sculpin *C. asper*, an abundant  
546 euryhaline fish in northwestern North America. Measuring pairwise genetic  
547 differentiation between samples using  $\hat{F}_{ST}^{pool}$ , we found a clear-cut structure  
548 separating the freshwater from the brackish-water ecotypes. Such genetic  
549 structure supports the hypothesis that populations are locally adapted to  
550 osmotic conditions in these two contrasted habitats, as discussed in Den-  
551 nenmoser et al. (2017). This structure, which is at odds with that inferred  
552 from PP2<sub>d</sub> estimates, is not only supported by the scaled covariance ma-  
553 trix of allele frequencies, but also by previous microsatellite-based studies,  
554 who showed that populations were genetically more differentiated between  
555 ecotypes than within ecotypes (Dennenmoser et al. 2014, 2015).

## 556 **Limits of the model and perspectives**

557 We have shown that the stronger source of bias for the  $\hat{F}_{ST}^{pool}$  estimate is un-  
558 equal contributions of individuals in pools. This is so, because we assume in  
559 our model that the read counts are multinomially distributed, which supposes  
560 that all genes contribute equally to the pool of reads (Gautier et al. 2013),

561 i.e. that there is no variation in DNA yield across individuals and that all  
562 genes have equal sequencing coverage (Rode et al. 2018). Because the effect  
563 of unequal contribution is expected to be stronger with small pool sizes, it  
564 has been recommended to use pool-seq with at least 50 diploid individuals  
565 per pool (Lynch et al. 2014; Schlötterer et al. 2014). However, this limit may  
566 be overly conservative for allele frequency estimates (Rode et al. 2018), and  
567 we have shown here that we can achieve very good precision and accuracy  
568 of  $F_{ST}$  estimates with smaller pool sizes. Furthermore, because genotypic in-  
569 formation is lost during Pool-seq experiments, we assume in our derivations  
570 that pools are haploid (and therefore that  $F_{IS}$  is nil). Analyzing non-random  
571 mating populations (e.g., in selfing species) is therefore problematic.

572 Finally, our model, as in Weir and Cockerham (1984), formally assumes  
573 that all populations provide independent replicates of some evolutionary pro-  
574 cess (Excoffier 2007; Holsinger and Weir 2009). This may be unrealistic in  
575 many natural populations, which motivated Weir and Hill (2002) to derive a  
576 population-specific estimator of  $F_{ST}$  for Ind-seq data (see also Vitalis et al.  
577 2001). Even though the use of Weir and Hill's (2002) estimator is still scarce  
578 in the literature (but see Weir et al. 2005; Vitalis 2012), Weir and Goudet  
579 (2017) recently proposed a re-interpretation of population-specific estimates  
580 of  $F_{ST}$  in terms of allelic matching proportions, which are strictly equivalent  
581 to IIS probabilities between pairs of genes. It would therefore be straight-  
582 forward to extend Weir and Goudet's (2017) estimator of population-specific  
583  $F_{ST}$  for the analysis of Pool-seq data, using the unbiased estimates of IIS  
584 probabilities provided in the Supplemental File S1.

585

## DATA ACCESSIBILITY

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

589

## ACKNOWLEDGEMENTS

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



LITERATURE CITED

600

601 Akey, J. M., Zhang, G., Jin, L., and Shriver, M. D. (2002). Interrogating a  
602 high-density SNP map for signatures of natural selection. *Genome Res.*,  
603 12:1805–1814.

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

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

609 Beaumont, M. A. and Nichols, R. A. (1996). Evaluating loci for use in the  
610 genetic analysis of population structure. *Proc. R. Soc. Lond. B*, 263:1619–  
611 1626.

612 Bhatia, G., Patterson, N., Sankararaman, S., and Price, A. L. (2013). Esti-  
613 mating and interpreting  $F_{ST}$ : the impact of rare variants. *Genome Res.*,  
614 23:1514–1521.

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

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

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

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

- 622 Cockerham, C. C. and Weir, B. S. (1987). Analyses of gene frequencies. *Proc.*  
623 *Natl. Acad. Sci. USA*, 84:8512–8514.
- 624 Collet, J. M., Fuentes, S., Hesketh, J., Hill, M. S., Innocenti, P., Morrow,  
625 E. H., Fowler, K., and Reuter, M. (2016). Rapid evolution of the intersex-  
626 ual genetic correlation for fitness in *Drosophila melanogaster*. *Evolution*,  
627 70:781–795.
- 628 Coop, G., Witonsky, D., Di Rienzo, A., and Pritchard, J. K. (2010). Us-  
629 ing environmental correlations to identify loci underlying local adaptation.  
630 *Genetics*, 185:1411–1423.
- 631 Cutler, D. J. and Jensen, J. D. (2010). To pool, or not to pool? *Genetics*,  
632 186:41–43.
- 633 Dennenmoser, S., Nolte, A. W., Vamosi, S. M., and Rogers S, M. (2015). Phy-  
634 logeography of the prickly sculpin (*Cottus asper*) in north-western North  
635 America reveals parallel phenotypic evolution across multiple coastal-  
636 inland colonizations. *J. Biogeogr.*, 42:1626–1638.
- 637 Dennenmoser, S., Rogers, S. M., and Vamosi, S. M. (2014). Genetic pop-  
638 ulation structure in prickly sculpin (*Cottus asper*) reflects isolation-by-  
639 environment between two life-history ecotypes. *Biol. J. Linnean Soc.*,  
640 113:943–957.
- 641 Dennenmoser, S., Vamosi, S. M., Nolte, S. W., and Rogers, S. M. (2017).  
642 Adaptive genomic divergence under high gene flow between freshwater and  
643 brackish-water ecotypes of prickly sculpin (*Cottus asper*) revealed by Pool-  
644 Seq. *Mol. Ecol.*, 26:25–42.

- 645 Dexter, E., Bollens, S. M., Cordell, J., Soh, H. Y., Rollwagen-Bollens, G.,  
646 Pfeifer, S. P., Goudet, J., and Vuilleumier, S. (2018). A genetic reconstruc-  
647 tion of the invasion of the calanoid copepod *Pseudodiaptomus inopinus*  
648 across the North American Pacific Coast. *Biol. Invasions*, 20:1577–1595.
- 649 Ellegren, H. (2014). Genome sequencing and population genomics in non-  
650 model organisms. *Trends Ecol. Evol.*, 29:51–63.
- 651 Excoffier, L. (2007). Analysis of population subdivision. In Balding, D. J.,  
652 Bishop, M., and Cannings, C., editors, *Handbook of Statistical Genetics*,  
653 pages 980–1020, Chichester. John Wiley & Sons, Ltd.
- 654 Fariello, M. I., Boitard, S., Mercier, S., Robelin, D., Faraut, T., Arnould, C.,  
655 Recoquillay, J., Bouchez, O., Salin, G., Dehais, P., Gourichon, D., Leroux,  
656 S., Pitel, F., Leterrier, C., and SanCristobal, M. (2017). Accounting for  
657 Linkage Disequilibrium in genome scans for selection without individual  
658 genotypes : the local score approach. *Mol. Ecol.*, 26:3700–3714.
- 659 Ferretti, L., Ramos Onsins, S., and Pérez-Enciso, M. (2013). Population  
660 genomics from pool sequencing. *Mol. Ecol.*, 22:5561–5576.
- 661 Fischer, M. C., Rellstab, C., Leuzinger, M., Roumet, M., Gugerli, F.,  
662 Shimizu, K. K., Holderegger, R., and Widmer, A. (2017). Estimating ge-  
663 nomic diversity and population differentiation – an empirical comparison  
664 of microsatellite and SNP variation in *Arabidopsis halleri*. *BMC Genomics*,  
665 18:69.
- 666 Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters.  
667 *Psychol. Bull.*, 76:378–382.

- 668 Fleiss, J. L. and Cuzick, J. (1979). The reliability of dichotomous judgements:  
669 Unequal numbers of judges per subject. *Appl. Psychol. Meas.*, 3:537–542.
- 670 Fuentes-Pardo, A. P. and Ruzzente, D. E. (2017). Whole-genome sequencing  
671 approaches for conservation biology: Advantages, limitations and practical  
672 recommendations. *Mol. Ecol.*, 26:5369–5406.
- 673 Futschik, A. and Schlötterer, C. (2010). The next generation of molecu-  
674 lar markers from massively parallel sequencing of pooled DNA samples.  
675 *Genetics*, 186:207–218.
- 676 Gautier, M. (2015). Genome-wide scan for adaptive divergence and associa-  
677 tion with population-specific covariates. *Genetics*, 201:1555–1579.
- 678 Gautier, M., Gharbi, K., Cezaerd, T., Galan, M., Loiseau, A., Thomson, M.,  
679 Pudlo, P., Kerdelhué, C., and Estoup, A. (2013). Estimation of popula-  
680 tion allele frequencies from next-generation sequencing data: pool-versus  
681 individual-based genotyping. *Mol. Ecol.*, 22:3766–3779.
- 682 Glenn, T. C. (2011). Field guide to next-generation DNA sequencers. *Mol.*  
683 *Ecol. Resour.*, 11:759–769.
- 684 Goudet, J. (1993). *The genetics of geographically structured populations*. PhD  
685 thesis, University of Wales, Bangor.
- 686 Holsinger, K. S. and Weir, B. S. (2009). Genetics in geographically structured  
687 populations: defining, estimating and interpreting  $F_{ST}$ . *Nat. Rev. Genet.*,  
688 10:639–650.
- 689 Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral  
690 model of genetic variation. *Bioinformatics*, 18:337–338.

- 691 Karlsson, E. K., Baranowska, I., Wade, C. M., Salmon Hillbertz, N. H. C.,  
692 Zody, M. C., Anderson, N., Biagi, T. M., Patterson, N., Pielberg, G. R.,  
693 Kulbokas, E. J., Comstock, K. E., Keller, E. T., Mesirov, J. P., von Euler,  
694 H., Kämpe, O., Hedhammar, A., Lander, E. S., Andersson, G., Andersson,  
695 L., and Lindblad-Toh, K. (2007). Efficient mapping of Mendelian traits in  
696 dogs through genome-wide association. *Nat. Genet.*, 39:1321–1328.
- 697 Kofler, R., Pandey, R. V., and Schlötterer, C. (2011). PoPoolation2: identi-  
698 fying differentiation between populations using sequencing of pooled DNA  
699 samples (Pool-Seq). *Bioinformatics*, 27:3435–3436.
- 700 Landis, J. R. and Koch, G. G. (1977). A one-way components of variance  
701 model for categorical data. *Biometrics*, 33:671–679.
- 702 Leblois, R., Gautier, M., Rohfritsch, A., Foucaud, J., Burban, C., Galan, M.,  
703 Loiseau, A., Sauné, L., Branco, M., Gharbi, K., Vitalis, R., and Kerdelhué,  
704 C. (2018). Deciphering the demographic history of allochronic differentia-  
705 tion in the pine processionary moth *Thaumetopoea pityocampa*. *Mol. Ecol.*,  
706 27:264–278.
- 707 Lewontin, R. C. and Krakauer, J. (1973). Distribution of gene frequency as  
708 a test of the theory of the selective neutrality of polymorphism. *Genetics*,  
709 74:175–195.
- 710 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth,  
711 G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing  
712 Subgroup (2009). The Sequence Alignment/Map format and SAMtools.  
713 *Bioinformatics*, 25:2078–2079.

- 714 Lotterhos, K. E. and Whitlock, M. C. (2014). Evaluation of demographic  
715 history and neutral parameterization on the performance of  $F_{ST}$  outlier  
716 tests. *Mol. Ecol.*, 23:2178–2192.
- 717 Lotterhos, K. E. and Whitlock, M. C. (2015). The relative power of genome  
718 scans to detect local adaptation depends on sampling design and statistical  
719 method. *Mol. Ecol.*, 24:1031–1046.
- 720 Lynch, M., Bost, D., Wilson, S., Maruki, T., and Harrison, S. (2014).  
721 Population-genetic inference from pooled-sequencing data. *Genome Biol.*  
722 *Evol.*, 6:1210–1218.
- 723 Mak, T. K. (1988). Analysing intraclass correlation for dichotomous vari-  
724 ables. *J. R. Stat. Soc. Ser. C Appl. Stat.*, 37:344–352.
- 725 Malécot, G. (1948). *Les Mathématiques de l’Hérédité*. Masson, Paris.
- 726 Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proc.*  
727 *Natl. Acad. Sci. USA*, 70:3321–3323.
- 728 Nei, M. (1977).  $F$ -statistics and analysis of gene diversity in subdivided  
729 populations. *Ann. Hum. Genet.*, 41:225–233.
- 730 Nei, M. (1978). Estimation of average heterozygosity and genetic distance  
731 from a small number of individuals. *Genetics*, 89:583–590.
- 732 Nei, M. (1986). Definition and estimation of fixation indices. *Evolution*,  
733 40:643–645.
- 734 Nei, M. and Chesser, R. K. (1983). Estimation of fixation indices and gene  
735 diversities. *Ann. Hum. Genet.*, 47:253–259.

- 736 Nychka, D., Furrer, R., Paige, J., and Sain, S. (2017). *fields*: Tools for spatial  
737 data. R package version 9.6.
- 738 Orgogozo, V., Peluffo, A. E., and Morizot, B. (2016). The “mendelian gene”  
739 and the “molecular gene”: two relevant concepts of genetic units. In Or-  
740 gogozo, V., editor, *Genes and Evolution*, volume 119 of *Current Topics in*  
741 *Developmental Biology*, pages 1–26. Academic Press.
- 742 Pickrell, J. K. and Pritchard, J. K. (2012). Inference of population splits  
743 and mixtures from genome-wide allele frequency data. *PLoS Genet.*,  
744 8(11):e1002967.
- 745 R Core Team (2017). *R: A Language and Environment for Statistical Com-*  
746 *puting*. R Foundation for Statistical Computing, Vienna, Austria.
- 747 Reynolds, J., Weir, B. S., and Cockerham, C. C. (1983). Estimation of the  
748 coancestry coefficient: basis for a short-term genetic distance. *Genetics*,  
749 105:767–779.
- 750 Ridout, M. S., Demktrio, C. G. B., and Firth, D. (1999). Estimating intra-  
751 class correlation for binary data. *Biometrics*, 55:137–148.
- 752 Robertson, A. (1962). Weighting in the estimation of variance components  
753 in the unbalanced single classification. *Biometrics*, 18:413–417.
- 754 Rode, N. O., Holtz, Y., Loridon, K., Santoni, S., Ronfort, J., and Gay, J.  
755 (2018). How to optimize the precision of allele and haplotype frequency  
756 estimates using pooled-sequencing data. *Mol. Ecol. Resour.*, 18:194–203.
- 757 Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty,

- 758 R., Nusbaum, C., and Jaffe, D. B. (2013). Characterizing and measuring  
759 bias in sequence data. *Genome Biol.*, 14:R51.
- 760 Rousset, F. (1996). Equilibrium values of measures of population subdivision  
761 for stepwise mutation processes. *Genetics*, 142:1357–1362.
- 762 Rousset, F. (1997). Genetic differentiation and estimation of gene flow from  
763  $F$ -statistics under isolation by distance. *Genetics*, 145:1219–1228.
- 764 Rousset, F. (2007). Inferences from spatial population genetics. In Bald-  
765 ing, D. J., Bishop, M., and Cannings, C., editors, *Handbook of Statistical*  
766 *Genetics*, pages 945–979, Chichester. John Wiley & Sons, Ltd.
- 767 Rousset, F. (2008). genepop'007: a complete re-implementation of the  
768 genepop software for Windows and Linux. *Mol. Ecol. Resour.*, 8:103–106.
- 769 Schlötterer, C., Tobler, R., Kofler, R., and Nolte, V. (2014). Sequencing  
770 pools of individuals – mining genome-wide polymorphism data without  
771 big funding. *Nat. Rev. Genet.*, 15:749–763.
- 772 Slatkin, M. (1993). Isolation by distance in equilibrium and non-equilibrium  
773 populations. *Evolution*, 47:264–279.
- 774 Smadja, C. M., Canbäck, B., Vitalis, R., Gautier, M., Ferrari, J., Zhou, J.-J.,  
775 and Butlin, R. K. (2012). Large-scale candidate gene scan reveals the role  
776 of chemoreceptor genes in host plant specialization and speciation in the  
777 pea aphid. *Evolution*, 66:2723–2738.
- 778 The International HapMap Consortium (2005). A haplotype map of the  
779 human genome. *Nature*, 437:1299–1320.



- 780 Tukey, J. W. (1957). Variances of variance components: II. The unbalanced  
781 single classification. *Ann. Math. Statist.*, 28:43–56.
- 782 Vitalis, R. (2012). DETSEL: An R-Package to detect marker loci responding  
783 to selection. In Pompanon, F. and Bonin, A., editors, *Data Production and*  
784 *Analysis in Population Genomics: Methods and Protocols*, volume 888 of  
785 *Methods in Molecular Biology*, pages 277–293, New York. Humana Press.
- 786 Vitalis, R., Boursot, P., and Dawson, K. (2001). Interpretation of variation  
787 across marker loci as evidence of selection. *Genetics*, 158:1811–1823.
- 788 Wahlund, S. (1928). Zusammensetzung von populationen und korrelation-  
789 serscheinungen vom standpunkt der vererbungslehre aus betrachtet. *Hered-*  
790 *itas*, 11:65–106.
- 791 Weir, B. S. (1996). *Genetic Data Analysis II*. Sinauer Associates, Inc.,  
792 Sunderland, MA.
- 793 Weir, B. S. (2012). Estimating  $F$ -statistics: A historical view. *Philos. Sci.*,  
794 79:637–643.
- 795 Weir, B. S., Cardon, L. R., Anderson, A. D., Nielsen, D. M., and Hill, W. G.  
796 (2005). Measures of human population structure show heterogeneity among  
797 genomic regions. *Genome Res.*, 15:1468–1476.
- 798 Weir, B. S. and Cockerham, C. C. (1984). Estimating  $F$ -statistics for the  
799 analysis of population structure. *Evolution*, 38:1358–1370.
- 800 Weir, B. S. and Goudet, J. (2017). An unified characterization of population  
801 structure and relatedness. *Genetics*, 206:2085–2103.

- 802 Weir, B. S. and Hill, W. G. (2002). Estimating  $F$ -statistics. *Annu. Rev.*  
803 *Genet.*, 36:721–750.
- 804 Whitlock, M. C. and Lotterhos, K. E. (2015). Reliable detection of loci re-  
805 sponsible for local adaptation: inference of a null model through trimming  
806 the distribution of  $F_{ST}$ . *Am. Nat.*, 186:S24–S36.
- 807 Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, 16:97–159.
- 808 Wright, S. (1951). The genetical structure of populations. *Ann. Eugen.*,  
809 15:323–354.
- 810 Wu, S., Crespi, C. M., and Wong, W. K. (2012). Comparison of methods  
811 for estimating the intraclass correlation coefficient for binary responses in  
812 cancer prevention cluster randomized trials. *Contemp. Clin. Trials*, 33:869–  
813 880.

**Table 1 Summary of main notations**

Notation	Parameter definition
$X_{ijr:k}$	Indicator variable: $X_{ijr:k} = 1$ if the $r$ th read from the $j$ th individual in the $i$ th 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 $i$ th pool
$c_{ij}$	Number of reads sequenced from individual $j$ in sub-population $i$ (unobserved individual coverage)
$C_{1i} \equiv \sum_j c_{ij}$	Total number of reads in the $i$ th 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 $i$ th pool (haploid pool size)
$y_{i:k}$	(Unobserved) number of genes of type $k$ in the $i$ th 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 $i$ th pool
$\hat{\pi}_{i:k} \equiv X_{i\cdot\cdot:k}$	Average frequency of reads of type $k$ in the $i$ th pool
$\hat{\pi}_k \equiv X_{\dots:k}$	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

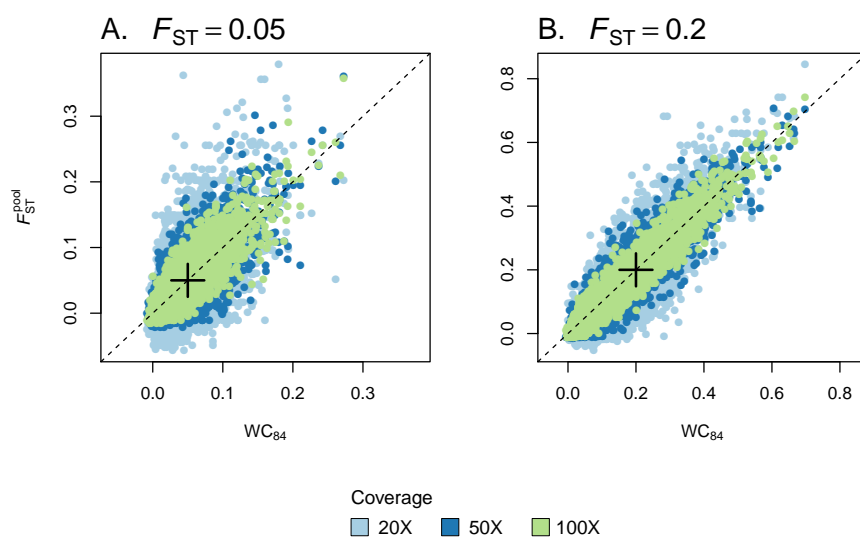
**Table 2** Definition of the  $F_{ST}$  estimators used in the text

Notation	Definition
$\hat{F}_{ST}^{pool}$	Equation 12
FRP <sub>13</sub>	Ferretti et al. (2013) and Equations 16,20–21
NC <sub>83</sub>	Nei and Chesser (1983)
PP2 <sub>d</sub>	Kofler et al. (2011) and Equations 16–18
PP2 <sub>a</sub>	Kofler et al. (2011) and Equation 19
WC <sub>84</sub>	Weir and Cockerham (1984)

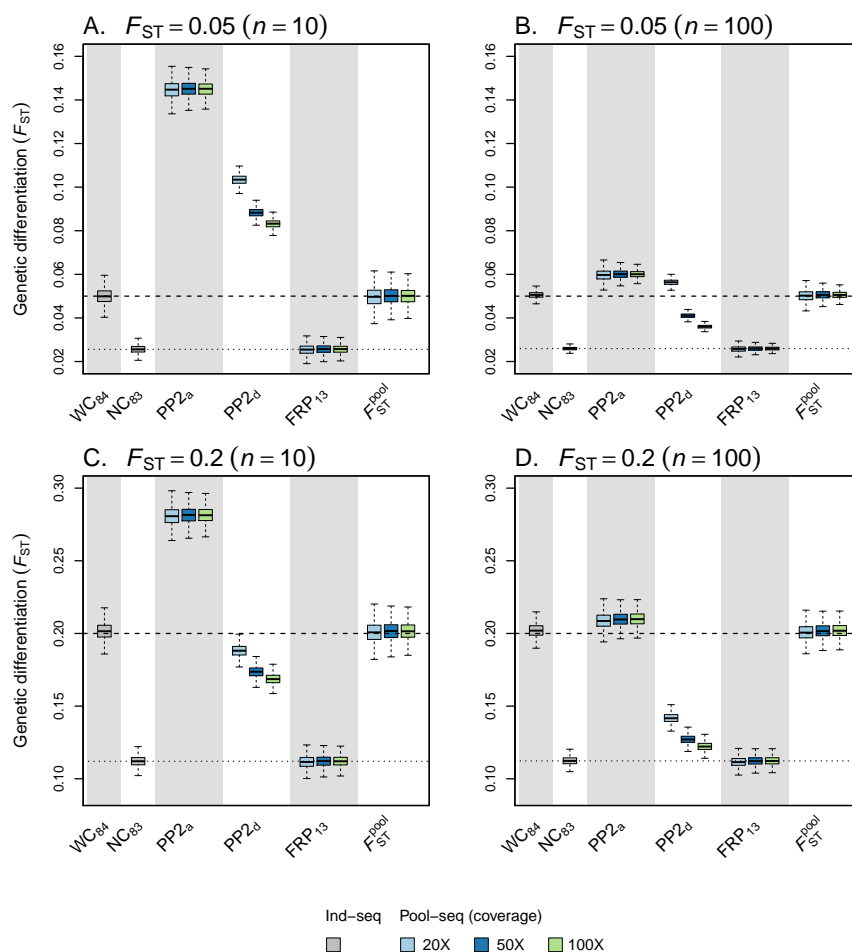
**Table 3 Overall  $F_{ST}$  estimates from multiple pools**

$F_{ST}$	$n$	Pool-seq		Ind-seq
		Cov.	$\hat{F}_{ST}^{pool}$	WC <sub>84</sub>
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)	

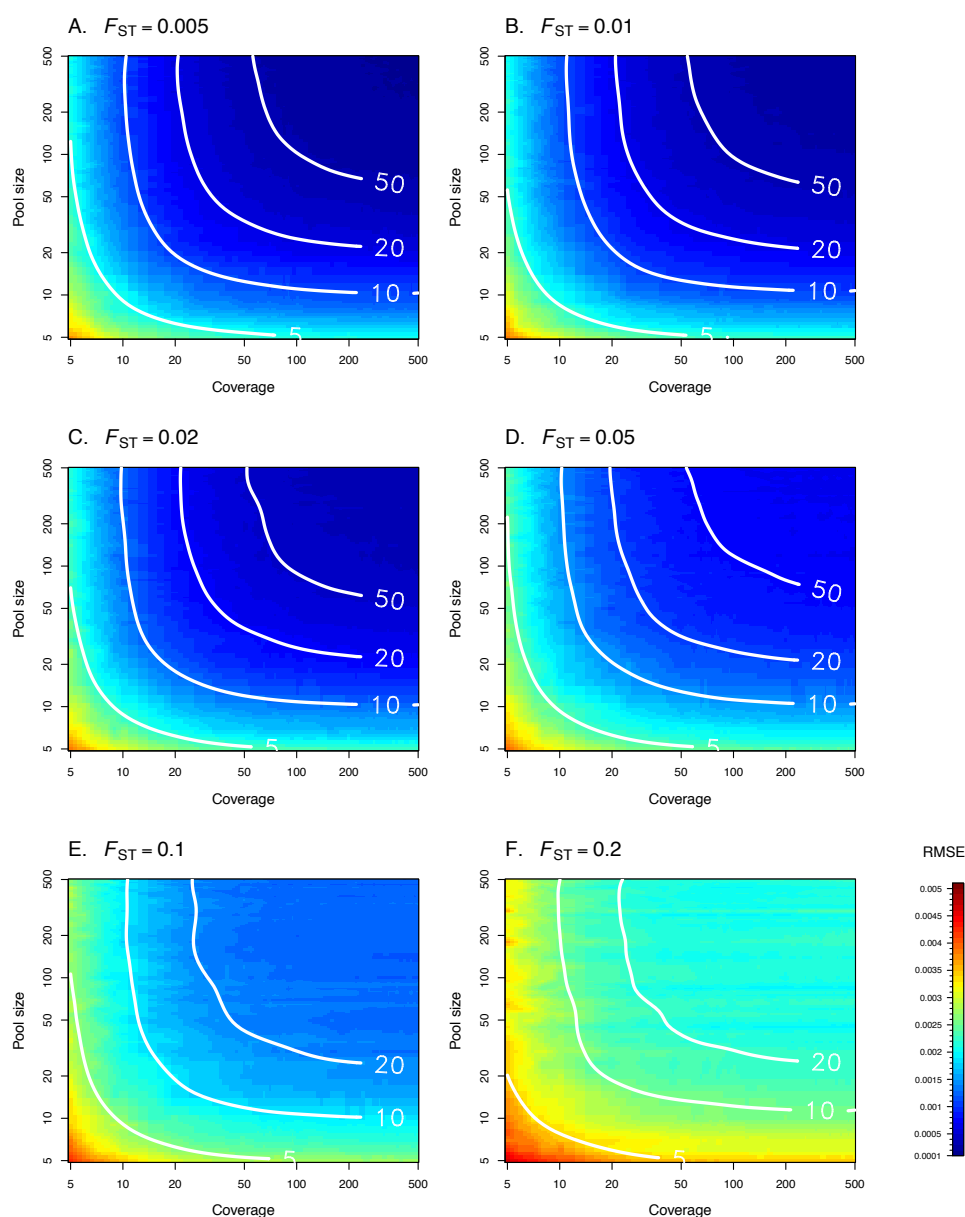
Multilocus  $\hat{F}_{ST}^{pool}$  estimates were computed for various conditions of expected  $F_{ST}$ , pool size ( $n$ ) and coverage (Cov.) in an island model with  $n_d = 8$  subpopulations (pools). The mean (RMSE) is over 50 independent simulated datasets, each made of 5,000 loci. For comparison, we computed multilocus WC<sub>84</sub> estimates 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  $\hat{F}_{ST}^{pool}$  estimates from Pool-seq data. We simulated 5,000 SNPs using `ms` in an island model with  $n_d = 8$  demes. We used two migration rates corresponding to  $F_{ST} = 0.05$  (A) and  $F_{ST} = 0.20$  (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  $F_{ST}$ .

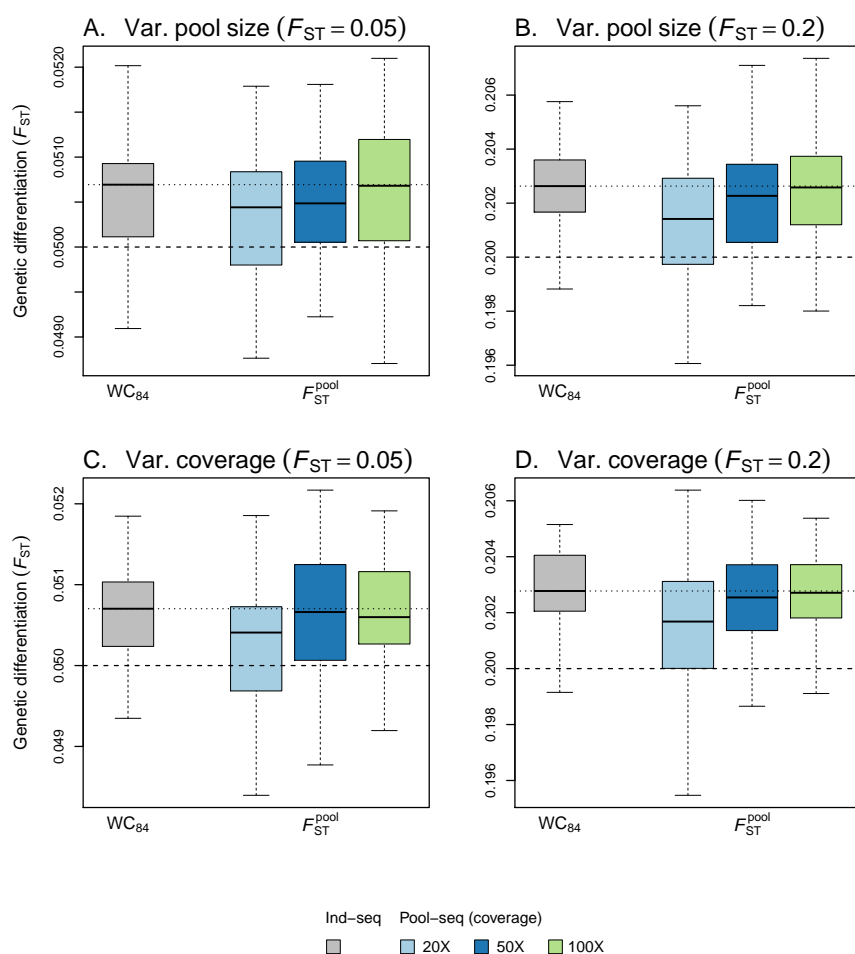


**Figure 2** Precision and accuracy of pairwise estimators of  $F_{ST}$ . We considered two estimators based on allele count data inferred from individual genotypes (Ind-seq):  $WC_{84}$  and  $NC_{83}$ . For pooled data, we computed the two estimators implemented in the software package POPOOLATION2, that we refer to as  $PP2_d$  and  $PP2_a$ , as well as the  $FRP_{13}$  estimator and our estimator  $\hat{F}_{ST}^{pool}$ . Each boxplot represents the distribution of multilocus  $F_{ST}$  estimates across all pairwise comparisons in an island model with  $n_d = 8$  demes, and across 50 independent replicates of the `ms` simulations. We used two migration rates, corresponding to  $F_{ST} = 0.05$  (A–B) and  $F_{ST} = 0.20$  (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.

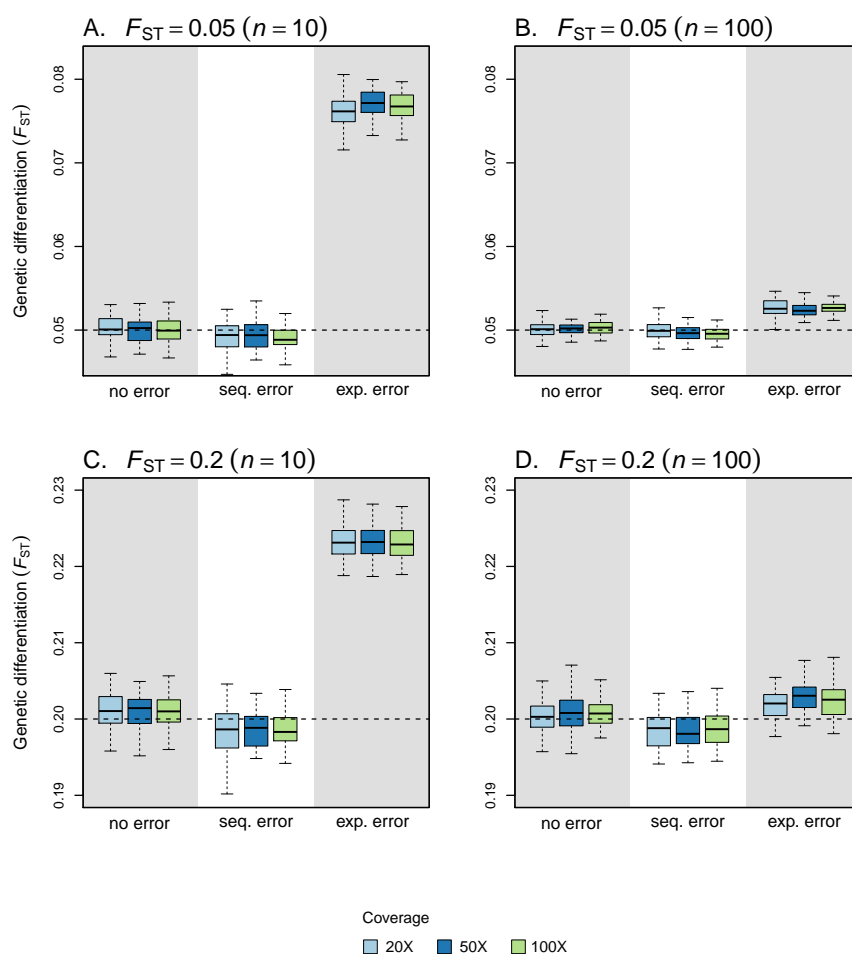


**Figure 3** Precision and accuracy of our estimator  $\hat{F}_{ST}^{pool}$  as a function of pool size and coverage, for simulated  $F_{ST}$  values ranging from 0.005 to 0.2 (A–F). Each density plot, which represents the root mean squared error (RMSE) of the estimator  $\hat{F}_{ST}^{pool}$ , was obtained using simple linear interpolation from a set of  $44 \times 44$  pairs of pool size and coverage values. For each pool size and coverage, 500 replicates of 5,000 markers were simulated from an island model with  $n_d = 8$  demes. Plain white isolines represent the RMSE of the  $WC_{84}$  estimator computed from Ind-seq data, for various sample sizes ( $n = 5, 10, 20, 50$ ). 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).

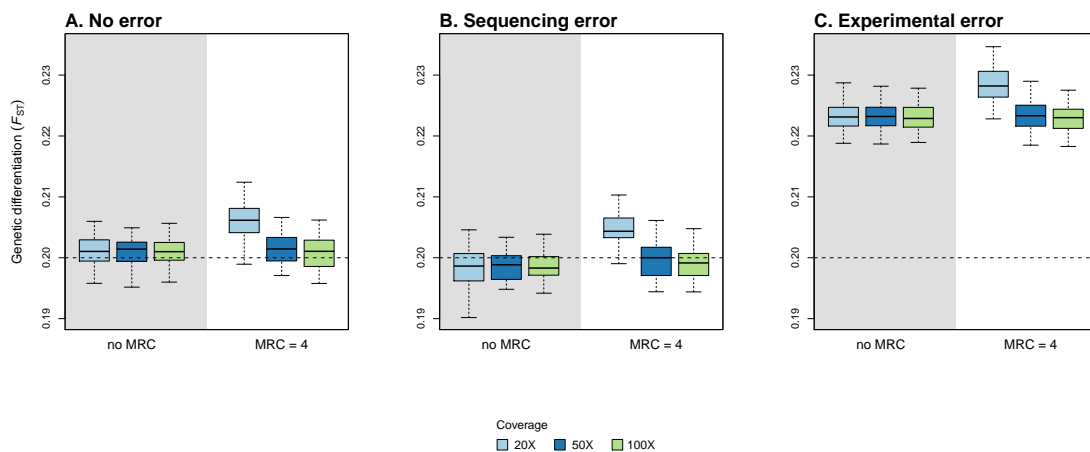




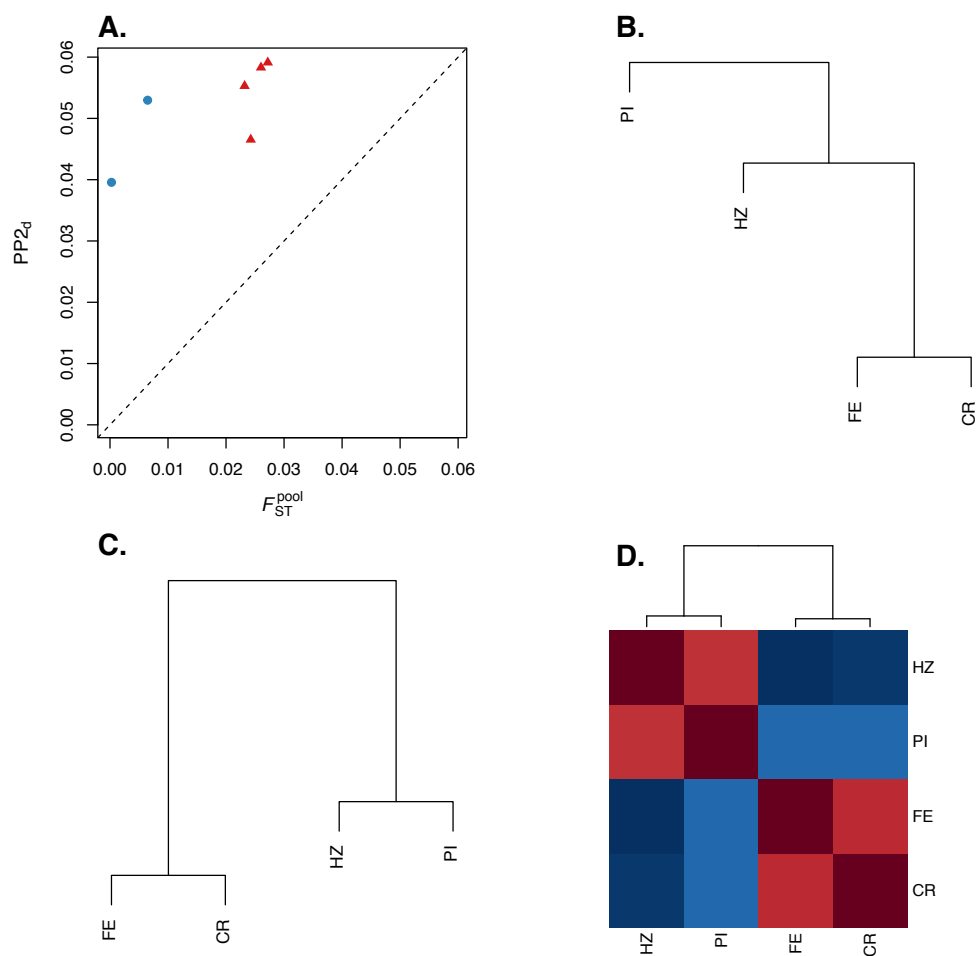
**Figure 4** Precision and accuracy of  $F_{ST}$  estimates with varying pool size or varying coverage. Our estimator  $\hat{F}_{ST}^{pool}$  was calculated from Pool-seq data over all demes and loci and compared to the estimator  $WC_{84}$ , computed from individual genotypes (Ind-seq). Each boxplot represents the distribution of multilocus  $F_{ST}$  estimates across 50 independent replicates of the `ms` simulations. We used two migration rates, corresponding to  $F_{ST} = 0.05$  (A and C) and  $F_{ST} = 0.20$  (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 ( $\delta_i$ ) was varying across demes and loci, with  $\delta_i \sim \text{Pois}(\Delta)$  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  $\hat{F}_{ST}^{\text{pool}}$  was computed from Pool-seq data over all demes and loci without error, with sequencing error (occurring at rate  $\mu_e = 0.001$ ), 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  $F_{ST} = 0.05$  (A–B) or  $F_{ST} = 0.20$  (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  $\hat{F}_{ST}^{\text{pool}}$  was computed from Pool-seq data over all demes and loci 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  $F_{ST} = 0.20$ , 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** Reanalysis of the prickly sculpin (*Cottus asper*) Pool-seq data. In (A) we compare the pairwise  $F_{ST}$  estimates  $PP2_d$ , and  $\hat{F}_{ST}^{pool}$  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 UPGMA hierarchical cluster analyses based on  $PP2_d$  (B) and  $\hat{F}_{ST}^{pool}$  (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.