



HAL
open science

f-statistics estimation and admixture graph construction with Pool-Seq or allele count data using the R package poolfstat

Mathieu Gautier, Renaud Vitalis, Laurence Flori, Arnaud Estoup

► To cite this version:

Mathieu Gautier, Renaud Vitalis, Laurence Flori, Arnaud Estoup. f-statistics estimation and admixture graph construction with Pool-Seq or allele count data using the R package poolfstat. *Molecular Ecology Resources*, 2022, 22 (4), pp.1394-1416. 10.1111/1755-0998.13557 . hal-03481066

HAL Id: hal-03481066

<https://hal.inrae.fr/hal-03481066>

Submitted on 14 Jan 2022

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.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0
International License

f-statistics estimation and admixture graph construction with Pool-Seq or allele count data using the R package *poolfstat*

Mathieu Gautier^{1,*}, Renaud Vitalis¹, Laurence Flori² and Arnaud Estoup¹

¹CBGP, INRAE, CIRAD, IRD, Montpellier SupAgro, Univ Montpellier, Montpellier, France

²SelMet, INRAE, CIRAD, Montpellier SupAgro, Montpellier, France

*Corresponding author:

Mathieu Gautier

INRAE-CBGP (Centre de Biologie pour la Gestion des Populations), 755 avenue du campus Agropolis,

CS30016, 34988 Montferrier sur lez cedex

Tel: +33499623331; Fax: +33499623345

E-mail: mathieu.gautier@inrae.fr

Keywords: *f*-statistics; Admixture Graph; Pool-Seq; Demographic Inference; *Drosophila suzukii*

Running title: Inferring demographic history with the R package poolfstats v2.0

1 **Abstract**

2 By capturing various patterns of the structuring of genetic variation across populations, f -statistics have proved
3 highly effective for the inference of demographic history. Such statistics are defined as covariance of SNP allele
4 frequency differences among sets of populations without requiring haplotype information and are hence particu-
5 larly relevant for the analysis of pooled sequencing (Pool-Seq) data. We here propose a reinterpretation of the F
6 (and D) parameters in terms of probability of gene identity and derive from this unified definition unbiased estima-
7 tors for both Pool-Seq and standard allele count data obtained from individual genotypes. We implemented these
8 estimators in a new version of the R package `poolfstat`, which now includes a wide range of inference methods:
9 (i) three-population test of admixture; (ii) four-population test of treeness; (iii) F_4 -ratio estimation of admixture
10 rates; and (iv) fitting, visualization and (semi-automatic) construction of admixture graphs. A comprehensive eval-
11 uation of the methods implemented in `poolfstat` on both simulated Pool-Seq (with various sequencing coverages
12 and error rates) and allele count data confirmed the accuracy of these approaches, even for the most cost-effective
13 Pool-Seq design involving low sequencing coverages. We further analyzed a real Pool-Seq data made of 14 pop-
14 ulations of the invasive species *Drosophila suzukii* which allowed refining both the demographic history of native
15 populations and the invasion routes followed by this emblematic pest. Our new package `poolfstat` provides the
16 community with a user-friendly and efficient all-in-one tool to unravel complex population genetic histories from
17 large-size Pool-Seq or allele count SNP data.

18 **1 Introduction**

19 In their seminal paper, Reich *et al* (2009) introduced a new population genetics framework to decipher the his-
20 tory of Indian human populations. This inference approach relied on a set of so-called f -statistics that are aimed
21 at capturing various patterns of the structuring of genetic diversity across-population based on Single Nucleotide
22 Polymorphism (SNP) assayed on a genome-wide scale (see also Patterson *et al*, 2012). The parameters underly-
23 ing these statistics and denoted F following Patterson *et al* (2012) are defined as covariances in allele frequency
24 difference among sets of two (F_2), three (F_3) or four (F_4) populations and were demonstrated to be highly infor-
25 mative about populations demographic history when modeled as admixture graphs, i.e., population trees possibly
26 including admixture events (Patterson *et al*, 2012). Hence, formal tests of admixture, called *three-population* test,
27 between a target population and two source population surrogates can be derived from estimates of F_3 . Con-
28 versely, via the so-called *four-population* test, estimating F_4 among quadruplets of populations allows to test for
29 their treeness, i.e., if their joint history can be modeled as a simple (unrooted) bifurcating tree. Under certain
30 restrictive assumptions about the underlying phylogeny, accurate estimates of the relative contributions of the an-
31 cestral sources of an admixed population may be obtained from ratios of F_4 involving some of its related sampled
32 populations. A normalized version of the F_4 parameter, called Patterson's D , was also developed by Green *et al*
33 (2010) and has become quite popular to characterize introgression in phylogenies of closely related species (Du-
34 rand *et al*, 2011). Finally, f -statistics can directly be used to fit admixture graphs (i.e., estimate branch lengths
35 and/or admixture proportions) and to rigorously assess their support (Patterson *et al*, 2012; Lipson *et al*, 2013;
36 Lipson, 2020).

37 A critical advantage of F and D parameters is that they only depend on population allele frequencies and
38 their estimation does not require haplotype information. The non-independence of neighboring SNPs (Linkage
39 Disequilibrium or LD) can be accurately accounted for with block-jackknife statistical techniques (Patterson *et al*,
40 2012; Kunsch, 1989; Reich *et al*, 2009; Busing *et al*, 1999) when computing standard errors of the estimated
41 f -statistics which are noticeably required for the derivation of formal tests of admixture or treeness and also
42 to assess the residuals of fitted admixture graphs. These characteristics make the f -statistics based inference
43 framework particularly attractive for the analysis of Pool-Seq data that result from the massive sequencing of
44 pools of individual DNA and have become quite popular, most particularly in non-model species (Schlötterer *et al*,
45 2014). Indeed, although LD information is generally lost in Pool-Seq experiments (but see Long *et al*, 2011; or
46 Feder *et al*, 2012), they lead to accurate and cost-effective assessment of allele frequencies across populations on a

47 whole genome basis (Gautier *et al*, 2013; Schlötterer *et al*, 2014). If the derivation of unbiased estimates of allele
48 frequencies from Pool-Seq data is straightforward, estimation of more elaborated population genetics parameters
49 characterizing the structuring of genetic diversity within or across populations is more challenging (Gautier *et al*,
50 2013; Ferretti *et al*, 2013; Hivert *et al*, 2018). As the individual origin of the sequencing reads is not identifiable
51 within pools, it is not possible to assess whether reads are identical because they are sequenced copies of the
52 same individual chromosome or because they are copies of different chromosomes carrying the same allele. The
53 resulting additional level of variation thus needs to be accounted for in the estimation which, in contrast to the
54 nucleotide diversities (heterozygosities) or the well-known F_{ST} differentiation measure (Ferretti *et al*, 2013; Hivert
55 *et al*, 2018), has to our knowledge not been investigated for the estimation of F and D parameters (but see Leblois
56 *et al*, 2018; Collin *et al*, 2021).

57 In the present paper, we first propose a (re)interpretation of the different F and D parameters in terms of
58 probability of identity in state (IIS or AIS for Alike-In-State) of pairs of genes sampled either within the same
59 population (Q_1) or between two different populations (Q_2), extending results we introduced in some earlier studies
60 (Hivert *et al*, 2018; Leblois *et al*, 2018; Collin *et al*, 2021). This unified definition simplified the derivation
61 of the unbiased estimators for both allele-count and Pool-Seq read count data, that we implemented in a new
62 version of our R package `poolfstat` (Hivert *et al*, 2018) together with methods that rely on the estimated f -
63 statistics for historical and demographic inference. These methods include i) three-population test of admixture; ii)
64 four-population test of treeness; iii) F_4 -ratio estimation of admixture proportion; and iv) fitting, visualization and
65 (semi-automatic) construction of admixture graphs. For completion, we briefly present the underlying methods
66 as implemented in the package. We then carried out a comprehensive evaluation of the whole package on both
67 simulated allele count and Pool-Seq read count data, considering for the latter various sequencing coverages and the
68 presence or not of sequencing errors. Finally, we illustrate the power and limitations of `poolfstat` by analyzing
69 real Pool-Seq data available from a previous study (Olazcuaga *et al*, 2020) for 14 populations of the invasive species
70 *Drosophila suzukii*. This example illustrates how f -statistics based inference and admixture graph construction
71 may confirm previous inferences and provide new insights into both the history of populations from the native area
72 and the invasion routes followed by an emblematic invasive species. We provide as Supplementary Materials, a first
73 vignette (Supplementary Vignette V1) designed as a detailed hands-on manual to outline the main functionalities
74 of `poolfstat` and a second vignette (Supplementary Vignette V2) detailing the analysis of the *D. suzukii* data to
75 make it fully reproducible.

2 Material and Methods

2.1 Definition, estimation and f -statistics based inference methods

2.1.1 A unified definition of F_2 , F_3 and F_4 parameter and their scaled version F_{ST} , F_3^* and D in terms of Q_1 and Q_2 probabilities

Let p_A, p_B, p_C and p_D the allele frequency of an arbitrarily chosen allele at a random SNP segregating in populations A, B, C and D respectively. The parameters F_2, F_3 and F_4 were originally defined in terms of covariance in allele frequencies difference among different sets of populations as follows (Reich *et al*, 2009; Patterson *et al*, 2012):

$$\begin{aligned}
 \bullet \quad F_2(A; B) &\equiv \mathbb{E}[(p_A - p_B)^2] \\
 \bullet \quad F_3(A; B, C) &\equiv \mathbb{E}[(p_A - p_B)(p_A - p_C)] = \frac{1}{2}(F_2(A; B) + F_2(A; C) - F_2(B; C)) \\
 \bullet \quad F_4(A, B; C, D) &\equiv \mathbb{E}[(p_A - p_B)(p_C - p_D)] = \frac{1}{2}(F_2(A; D) + F_2(B; C) - F_2(A; C) - F_2(B; D))
 \end{aligned} \tag{1}$$

In total, with n populations, there are $\binom{n}{2} = \frac{1}{2}n(n-1)$ possible F_2 ; $3\binom{n}{3} = \frac{1}{2}n(n-1)(n-2)$ possible F_3 ; and $3\binom{n}{4} = \frac{1}{8}n(n-1)(n-2)(n-3)$ possible F_4 . Note that these values exclude the alternative equivalent configurations that result from the permutation of populations within pairs (since $F_2(A; B) = F_2(B; A)$; $F_3(A; B, C) = F_3(A; C, B)$ and $F_4(A, B; C, D) = F_4(B, A; D, C) = -F_4(B, A; C, D) \dots$). Due to the linear dependency of all these parameters (eq. 1), the $\frac{1}{8}n(n-1)(n^2 - n + 2)$ F parameters actually span a vector space of dimension $\frac{1}{2}n(n-1)$ the basis of which may be specified by the set of all the $\binom{n}{2}$ possible F_2 or, given a reference population i (randomly chosen among the n ones) the set of all the $n-1$ F_2 of the form $F_2(i; j)$ (with $j \neq i$) and all the $\binom{n-1}{2}$ F_3 of the form $F_3(i; j, k)$ (with $j \neq i$; $k \neq i$ and $j \neq k$) (Patterson *et al*, 2012; Lipson, 2020). As mentioned by Patterson *et al* (2012), it is important to notice that these definitions are invariant in the choice of the reference SNP allele since:

$$F_2(A; B) \equiv \mathbb{E}[(p_A - p_B)^2] = \mathbb{E}[(1 - p_A) - (1 - p_B)]^2$$

It directly follows from this property that:

$$\begin{aligned}
 F_2(A; B) &= \frac{1}{2} \left(\mathbb{E} \left[(p_A - p_B)^2 \right] + \mathbb{E} \left[((1 - p_A) - (1 - p_B))^2 \right] \right) \\
 &= \frac{1}{2} \left(\mathbb{E} \left[p_A^2 \right] + \mathbb{E} \left[(1 - p_A)^2 \right] \right) + \frac{1}{2} \left(\mathbb{E} \left[p_B^2 \right] + \mathbb{E} \left[(1 - p_B)^2 \right] \right) - (\mathbb{E} [p_A p_B] + \mathbb{E} [(1 - p_A)(1 - p_B)]) \quad (2) \\
 &= \frac{Q_1^A + Q_1^B}{2} - Q_2^{A,B}
 \end{aligned}$$

where Q_1^A (resp. Q_1^B) is actually the probability of sampling two genes (or alleles) identical in state (IIS) within population A (resp. B) and $Q_2^{A,B}$ is the probability of sampling two IIS genes from A and B . It directly follows from equations 1 and 2 that:

$$F_3(A; B, C) = \frac{1}{2} \left(Q_1^A + Q_2^{B,C} - Q_2^{A,B} - Q_2^{B,C} \right)$$

and,

$$F_4(A, B; C, D) = \frac{1}{2} \left(Q_2^{A,C} + Q_2^{B,D} - Q_2^{A,C} - Q_2^{B,C} \right)$$

The Q_1 and Q_2 probabilities, and hence the F_2 , F_3 and F_4 parameters depend on both demographic parameters (i.e., population sizes, divergence times and other historical events) and marker polymorphism (i.e., their mutation rates and ascertainment process). For instance, under a simple pure-drift model with no mutation, if p_r denotes the allele frequency of the ancestral population R of two isolated populations A and B then $1 - Q_2^{A,B} = \mathbb{E} [2p_A p_B | p_r] = 2p_r(1 - p_r)$ which is the heterozygosity in R . Similarly, $1 - Q_1^A = 2p_r(1 - p_r)e^{-\tau_A}$ (resp., $1 - Q_1^B = 2p_r(1 - p_r)e^{-\tau_B}$) where τ_A (resp. τ_B) is the divergence time separating R and A (resp. B) on a diffusion timescale (i.e., in drift units of $\frac{1}{2N_e}$ where N_e is the effective population along the branch). As a consequence, the resulting estimates of F_2 , F_3 and F_4 strongly depend on the underlying set of genetic markers and may not be compared across different datasets, even from the same populations. Various scaling procedure may actually helps in reducing this dependence. Scaling the F_2 with respect to the across population heterozygosity $1 - Q_2$ leads to the standard definition of pairwise-population F_{ST} in terms of IIS probabilities (Rousset, 2007; Hivert *et al*, 2018) which is also concordant with its original definition as the numerator of F_{ST} (Reich *et al*, 2009; Peter, 2016):

$$F_{ST}(A; B) \equiv \frac{Q_1 - Q_2^{A,B}}{1 - Q_2^{A,B}} = \frac{F_2(A; B)}{1 - Q_2^{A,B}}$$

111 where $Q_1 = \frac{1}{2}(Q_1^A + Q_1^B)$ is the overall probability of sample two IIS genes within the same population (i.e.,
 112 averaged over populations A and B). Similarly, the scaled versions of the F_3 and F_4 statistics named F_3^* and D
 113 respectively (Patterson *et al*, 2012; Green *et al*, 2010; Durand *et al*, 2011), can be expressed as: $F_3^*(A; B, C) \equiv$
 114 $\frac{F_3(A; B, C)}{1 - Q_1^A}$ and $D(A, B; C, D) \equiv \frac{F_4(A, B; C, D)}{(1 - Q_2^{A,B})(1 - Q_2^{C,D})}$. To sum up, expressions of the F and D parameters as a function of
 115 Q_1 and Q_2 probability are finally defined as follows:

$$\begin{aligned}
 F_2(A; B) &\equiv \frac{Q_1^A + Q_1^B}{2} - Q_2^{A,B} & \text{and } F_{ST}(A; B) &\equiv \frac{F_2(A; B)}{1 - Q_2^{A,B}} = \frac{Q_1^A + Q_1^B - 2Q_2^{A,B}}{2(1 - Q_2^{A,B})} \\
 F_3(A; B, C) &\equiv \frac{Q_1^A + Q_2^{B,C} - Q_2^{A,B} - Q_2^{A,C}}{2} & \text{and } F_3^*(A; B, C) &\equiv \frac{F_3(A; B, C)}{1 - Q_1^A} = \frac{Q_1^A + Q_2^{B,C} - Q_2^{A,B} - Q_2^{A,C}}{2(1 - Q_1^A)} \\
 F_4(A, B; C, D) &\equiv \frac{Q_2^{A,C} + Q_2^{B,D} - Q_2^{A,D} - Q_2^{B,C}}{2} & \text{and } D(A, B; C, D) &\equiv \frac{F_4(A, B; C, D)}{(1 - Q_2^{A,B})(1 - Q_2^{C,D})} = \frac{Q_2^{A,C} + Q_2^{B,D} - Q_2^{A,D} - Q_2^{B,C}}{2(1 - Q_2^{A,B})(1 - Q_2^{C,D})}
 \end{aligned} \tag{3}$$

117 2.1.2 Unbiased parameter estimators from Pool-Seq read count and standard allele count data

118 Let y_{ij} be the allele count for an arbitrarily chosen reference allele and n_{ij} the total number of sampled alleles (e.g.,
 119 twice the number of genotyped individuals for a diploid species) at SNP i in population j . For Pool-Seq read count
 120 data, the y_{ij} 's are not observed and for a given pool j , it is assumed that $n_{ij} = n_j$ (the haploid sample size) for each
 121 and every SNP. We thus similarly defined r_{ij} as the read counts for the reference allele and c_{ij} the overall coverage
 122 observed at SNP i in population j .

123 If allele count data are directly observed, unbiased estimators of the IIS probability within population j ($Q_{1,i}^j$)
 124 and between a pair of populations j and k ($Q_{2,i}^{j,k}$) for a given SNP i are:

$$\begin{aligned}
 \widehat{Q}_{1,i}^j &= \frac{y_{ij}(y_{ij} - 1) + (n_{ij} - y_{ij})(n_{ij} - y_{ij} - 1)}{n_{ij}(n_{ij} - 1)} = 1 - 2 \frac{y_{ij}(n_{ij} - y_{ij})}{n_{ij}(n_{ij} - 1)} \\
 \text{and } \widehat{Q}_{2,i}^{j,k} &= \frac{y_{ij}y_{ik} + (n_{ij} - y_{ij})(n_{ik} - y_{ik})}{n_{ij}n_{ik}}
 \end{aligned} \tag{4}$$

126 For Pool-Seq read count, unbiased estimators of $Q_{1,i}$ and $Q_{2,i}$ are similarly defined as (Hivert *et al*, 2018, eqns
 127 A37 and A40):

$$\begin{aligned}
 \widehat{Q}_{1,i}^j &= 1 - \frac{n_j}{n_j - 1} \left(1 - \frac{r_{ij}(r_{ij} - 1) + (c_{ij} - r_{ij})(c_{ij} - r_{ij} - 1)}{c_{ij}(c_{ij} - 1)} \right) = 1 - 2 \frac{n_j}{n_j - 1} \frac{r_{ij}(c_{ij} - r_{ij})}{c_{ij}(c_{ij} - 1)} \\
 \text{and } \widehat{Q}_{2,i}^{j,k} &= \frac{r_{ij}r_{ik} + (c_{ij} - r_{ij})(c_{ik} - r_{ik})}{c_{ij}c_{ik}}
 \end{aligned} \tag{5}$$

129 Genome-wide estimates of all the parameters defined in eq. 3 above are then simply obtained from these
 130 unbiased estimators of IIS probabilities over all the I SNPs as:

$$\begin{aligned}
 f_2(A; B) = \widehat{F}_2(A; B) &= \frac{1}{2I} \sum_{i=1}^I (\widehat{Q}_{1,i}^A + \widehat{Q}_{1,i}^B - 2\widehat{Q}_{2,i}^{A,B}) & \text{and } \widehat{F}_{ST}(A; B) &= \frac{\sum_{i=1}^I (\widehat{Q}_{1,i}^A + \widehat{Q}_{1,i}^B - 2\widehat{Q}_{2,i}^{A,B})}{2 \sum_{i=1}^I (1 - \widehat{Q}_{2,i}^{A,B})} \\
 f_3(A; B, C) = \widehat{F}_3(A; B, C) &= \frac{1}{2I} \sum_{i=1}^I (\widehat{Q}_{1,i}^A + \widehat{Q}_{2,i}^{B,C} - \widehat{Q}_{2,i}^{A,B} - \widehat{Q}_{2,i}^{A,C}) & \text{and } f_3^*(A; B, C) = \widehat{F}_3^*(A; B, C) &= \frac{\sum_{i=1}^I (\widehat{Q}_{1,i}^A + \widehat{Q}_{2,i}^{B,C} - \widehat{Q}_{2,i}^{A,B} - \widehat{Q}_{2,i}^{A,C})}{2 \left(I - \sum_{i=1}^I \widehat{Q}_{1,i}^A \right)} \\
 f_4(A, B; C, D) = \widehat{F}_4(A, B; C, D) &= \frac{1}{2I} \sum_{i=1}^I (\widehat{Q}_{2,i}^{A,C} + \widehat{Q}_{2,i}^{B,D} - \widehat{Q}_{2,i}^{A,D} - \widehat{Q}_{2,i}^{B,C}) & \text{and } D(A, B; C, D) &= \frac{\sum_{i=1}^I (\widehat{Q}_{2,i}^{A,C} + \widehat{Q}_{2,i}^{B,D} - \widehat{Q}_{2,i}^{A,D} - \widehat{Q}_{2,i}^{B,C})}{2 \sum_{i=1}^I ((1 - \widehat{Q}_{2,i}^{A,B})(1 - \widehat{Q}_{2,i}^{C,D}))}
 \end{aligned}
 \tag{6}$$

131

132 Similarly, the within-population heterozygosity \widehat{h}_j for each population is simply estimated as:

$$\widehat{h}_j = 1 - \frac{1}{I} \sum_{i=1}^I \widehat{Q}_{1,i}^j \tag{7}$$

133 Importantly, for the three scaled parameters F_{ST} , F_3^* and D , multi-locus estimators consist of ratios of the
 134 numerator and denominator averages and not average of ratios (see e.g., Rousset, 2007; Patterson *et al*, 2012;
 135 Bhatia *et al*, 2013; Weir & Goudet, 2017; Hivert *et al*, 2018). Hence, for pairwise F_{ST} , the above estimator is
 136 similar to the one described in Rousset (2007) for allele count data and identical to the alternative PID estimator
 137 described in Hivert *et al* (2018) for Pool-Seq read count data (so-called ‘‘Identity’’ method of the `computeFST`
 138 function from the `poolfstat` package).

139 2.1.3 Block-Jackknife estimation of standard errors

140 Following Reich *et al* (2009), standard-errors of genome-wide estimates of the different statistics are computed
 141 using block-jackknife (Kunsch, 1989; Busing *et al*, 1999) which consists of dividing the genome into contiguous
 142 chunks of a predefined number of SNPs and then removing each block in turn to quantify the variability of the
 143 estimator. For a given parameter F , if n_b blocks are available and \widehat{F}_b is the estimated statistics when removing all
 144 SNPs belonging to block b , the standard error $\widehat{\sigma}_F$ of the genome-wide estimator \widehat{F} is computed as:

$$\widehat{\sigma}_F = \sqrt{\frac{n_b - 1}{n_b} \sum_{b=1}^{n_b} (\widehat{F}_b - \widehat{\mu}_F)^2}$$

145 where $\widehat{\mu}_F = \frac{1}{n_b} \sum_{b=1}^{n_b} \widehat{F}_b$, which may be slightly different than the estimator obtained with all the I markers since
 146 the latter may include SNPs that are not eligible for block-jackknife sampling (e.g., those at the chromosome
 147 or scaffolds boundaries). Finally, block-jackknife sampling may also be used to obtain estimates of the error
 148 covariance between two estimates \widehat{F}^u and \widehat{F}^v as:

$$\widehat{\text{Cov}}(\widehat{F}^u, \widehat{F}^v) = \frac{n_b - 1}{n_b} \sum_{b=1}^{n_b} (\widehat{F}_b^u - \widehat{\mu}_{F^u})(\widehat{F}_b^v - \widehat{\mu}_{F^v})$$

149 For convenience, we here chose to specify the same number of SNPs for each block instead of a block size in ge-
 150 netic distance (Patterson *et al*, 2012; Reich *et al*, 2009). We therefore do not recourse to a weighted block-jackknife
 151 (Busing *et al*, 1999). In practice, this has little impact providing the distribution of markers is homogeneous along
 152 the genome and the amount of missing data is negligible.

153 2.1.4 Admixture Graph fitting

154 The approach implemented in the new version of `poolfstat` to fit admixture graphs from f -statistics is directly
 155 inspired from the one proposed by Patterson *et al* (2012) and implemented in the `qpGraph` software (see also
 156 Lipson, 2020). Briefly, let $\widehat{\mathbf{f}}$ the vector (of length $\frac{n_l(n_l-1)}{2}$ where n_l is the number of graph leaves) of the estimated
 157 f_2 and f_3 statistics forming the basis of all the f -statistics (see above). Similarly, let $\mathbf{g}(\mathbf{e}; \mathbf{a}) = \mathbf{X}(\mathbf{a}) \times \mathbf{e}$ the vector
 158 of their expected values given the graph edge lengths vector \mathbf{e} and an incidence matrix $\mathbf{X}(\mathbf{a})$, which summarize
 159 the structure of the graph given the vector \mathbf{a} of proportions of all admixture events (for a tree-topology, $\mathbf{X}(\mathbf{a})$ only
 160 consists of 0 or 1). In `poolfstat`, $\mathbf{X}(\mathbf{a})$ is derived using simple operations from another n_l by n_e matrix (where n_e
 161 is the number of graph edges) that specifies the weights of each edge along all the paths connecting the graph leaves
 162 to the root. It should be noticed that an admixture event is modeled as an instantaneous mixing of two populations
 163 S_1 and S_2 into a population S directly ancestral to a child population A . An admixture event may thus be specified
 164 by i) one admixture rate α quantifying the relative S_1 and S_2 ancestry proportions (α and $1-\alpha$) in population S ; and
 165 ii) three edge lengths $e_{S \leftrightarrow A}$ for the branch connecting S and A and $e_{S_1 \leftrightarrow \mathcal{G}}$ and $e_{S_2 \leftrightarrow \mathcal{G}}$ for the branches connecting the
 166 two source populations to the rest of the graph \mathcal{G} . Yet, these three edge lengths are not identifiable and can only be
 167 estimated jointly in a single compound parameter (Pickrell & Pritchard, 2012; Patterson *et al*, 2012; Lipson, 2020):
 168 $\zeta = \alpha^2 \times e_{S_1 \leftrightarrow \mathcal{G}} + (1 - \alpha)^2 \times e_{S_2 \leftrightarrow \mathcal{G}} + e_{S \leftrightarrow A}$. Following Lipson (2020), this identifiability issue is solved by setting
 169 $e_{S_1 \leftrightarrow \mathcal{G}} = e_{S_2 \leftrightarrow \mathcal{G}} = 0$ (i.e., nullifying the edges connecting the two source populations to the graph). Although it has
 170 no impact on the interpretation of the graph, this may overestimate the length of $e_{S \leftrightarrow A}$ (i.e, the divergence between

171 the admixed population A and its direct ancestor S). Proceeding this way differs from the choice made by Pickrell
 172 & Pritchard (2012) in the `Treemix` package which consists, following our notations, of setting $e_{S \leftrightarrow A} = e_{S_2 \leftrightarrow \mathcal{G}} = 0$
 173 if $\widehat{\alpha} > 0.5$ and $e_{S \leftrightarrow A} = e_{S_1 \leftrightarrow \mathcal{G}} = 0$ otherwise.

174 We finally define \mathbf{Q} as the $\frac{n_l(n_l-1)}{2}$ by $\frac{n_l(n_l-1)}{2}$ covariance matrix of the basis f -statistics estimated by block-
 175 jackknife. Graph fitting consists of finding the graph parameter values ($\widehat{\mathbf{e}}$ and $\widehat{\mathbf{a}}$) that minimize a cost (score of the
 176 model) defined as:

$$S(\mathbf{e}; \mathbf{a}) = (\widehat{\mathbf{f}} - \mathbf{g}(\mathbf{e}; \mathbf{a}))' \mathbf{Q}^{-1} (\widehat{\mathbf{f}} - \mathbf{g}(\mathbf{e}; \mathbf{a})) = (\widehat{\mathbf{\Gamma}}\mathbf{f} - \mathbf{\Gamma}\mathbf{X}(\mathbf{a})\mathbf{e})' (\widehat{\mathbf{\Gamma}}\mathbf{f} - \mathbf{\Gamma}\mathbf{X}(\mathbf{a})\mathbf{e}) \quad (8)$$

177 where $\mathbf{\Gamma}$ results from the Cholesky decomposition of \mathbf{Q}^{-1} (i.e., $\mathbf{Q}^{-1} = \mathbf{\Gamma}'\mathbf{\Gamma}$). Given admixture rates \mathbf{a} , $S(\mathbf{e}; \mathbf{a})$
 178 is quadratic in the edge lengths \mathbf{e} (Patterson *et al*, 2012) leading us to rely on the Lawson-Hanson non-negative
 179 linear least squares algorithm implemented in the R package `nnls` (Lawson & Hanson, 1995)) to estimate the
 180 vector $\widehat{\mathbf{e}}$ that minimizes $S(\mathbf{e}; \mathbf{a})$ (subject to the constraint of positive edge lengths). Full minimization of $S(\mathbf{e}; \mathbf{a})$
 181 is thus reduced to the identification of the admixture rates \mathbf{a} which is performed using the L-BFGS-B algorithm
 182 implemented in the `optim` function of the R package `stats` (Nocedal & Wright, 1999).

183 2.1.5 Confidence Intervals and model fit assessment

184 Assume $\widehat{\mathbf{f}} \sim N(\mathbf{g}(\widehat{\mathbf{e}}; \widehat{\mathbf{a}}), \mathbf{Q})$, i.e., the vector of the basis f -statistics follows a multivariate normal distribution
 185 centered on the vector $\mathbf{g}(\widehat{\mathbf{e}}; \widehat{\mathbf{a}})$ specified by the fitted admixture graph parameters and the estimated error covariance
 186 matrix \mathbf{Q} . The optimized score $S(\widehat{\mathbf{e}}; \widehat{\mathbf{a}})$ then verifies $S(\widehat{\mathbf{e}}; \widehat{\mathbf{a}}) = -2\log(L) - K$ where L is the likelihood of the
 187 fitted graph and $K = n \log(2\pi) + \log(|\mathbf{Q}|)$. This makes it straightforward to compute a *BIC* (Bayesian Information
 188 Criterion) for the fitted graph from the optimized score as:

$$\text{BIC} = S(\widehat{\mathbf{e}}; \widehat{\mathbf{a}}) + n_{\text{par}} \log\left(\frac{1}{2}n_l(n_l - 1)\right) - \frac{1}{2}n_l(n_l - 1) \log(2\pi) - \log(|\mathbf{Q}|)$$

189 *BIC* may then be useful to compare different fitted admixture graph topologies. When comparing two graphs \mathcal{G}_1
 190 and \mathcal{G}_2 with BIC equal to BIC_1 and BIC_2 respectively, we have $\Delta_{12} = \text{BIC}_2 - \text{BIC}_1 \approx 2 \log(\text{BF}_{12})$ where BF_{12} is
 191 the Bayes Factor associated to the comparison of the graphs \mathcal{G}_1 and \mathcal{G}_2 (Kass & Raftery, 1995, eq. 9). We may
 192 further rely on the modified Jeffreys' rule proposed by Kass & Raftery (1995) to assess to which extent the data
 193 support either the \mathcal{G}_1 or \mathcal{G}_2 graphs, with $\Delta_{12} > 6$ (respectively $\Delta_{12} > 10$) providing “strong” (respectively “very
 194 strong”) evidence in favor of \mathcal{G}_1 (Supplementary Vignette V1).

195 Moreover, the likelihood interpretation of the optimized score $S(\hat{\mathbf{e}}; \hat{\mathbf{a}}) = -2\log(L) - K$ allows constructing
 196 confidence intervals (CI) for the fitted parameters of a given graph (i.e., elements of the \mathbf{e} and \mathbf{a} vectors) using
 197 the following uni-dimensional procedure. For a given parameter ν (either a edge length or an admixture rate), the
 198 difference $S_\nu(x) - S(\hat{\mathbf{e}}; \hat{\mathbf{a}})$ (where $S_\nu(x)$ is the score when $\nu = x$ and all the other parameters are set to their best fitted
 199 values) can be interpreted as a likelihood-ratio test statistics following a χ^2 distribution with one degree of freedom.
 200 Lower and upper boundaries ν_{\min} and ν_{\max} of the 95% CI (such $S_\nu(x) - S(\hat{\mathbf{e}}; \hat{\mathbf{a}}) < 3.84$ for all $\nu_{\min} < x < \nu_{\max}$) may
 201 then simply be computed using a bisection method, as implemented in `poolfstat`.

202 Finally, a straightforward (but highly informative and recommended) approach to assess the fit of an admixture
 203 graph is to evaluate to which extent the f -statistics derived from the fitted admixture graph parameters ($g(\hat{\mathbf{e}}; \hat{\mathbf{a}})$)
 204 depart from the estimated ones (Patterson *et al*, 2012; Lipson, 2020). This can be summarized via a Z-score of
 205 residuals computed as $Z = \frac{f - \hat{G}}{\sigma_F^2}$ where \hat{G} is a given fitted f -statistics; f is its corresponding estimated values;
 206 and σ_F^2 the block-jackknife standard error. The presence of outlying Z-scores for at least one f -statistics (e.g.,
 207 $|Z| > 1.96$ at a 95% significance threshold) may suggest poor model fit while also providing insights into the
 208 leaves or graph edges that are the most problematic (Lipson, 2020).

209 2.1.6 Scaling of branch lengths in drift units

210 Admixture graph fitting results in estimated edge lengths on the same scale as F_2 which limits their interpretation,
 211 because they depend both on the overall level of SNP polymorphism and on their distance to the root (Patterson
 212 *et al*, 2012). Lipson *et al* (2013) proposed an empirical approach to rescale edge lengths on a diffusion timescale
 213 using estimates of overall marker heterozygosities within (i.e., $1 - Q_1$) or across (i.e., $1 - Q_2$) populations. The
 214 argument echoes the aforementioned interpretation of pairwise F_{ST} as a scaled F_2 . If p_C and p_P are the reference
 215 allele frequencies in a child node C and its direct parent node P and their divergence time (on a diffusion timescale)
 216 is $\tau_{C,P} = \frac{t}{N_e}$ (where t is the branch length in generations), then conditional on p_C , $F_2(C; P) = (1 - e^{-\tau_{C,P}}) p_C(1 - p_C)$
 217 and $Q_2^{C,P} = 1 - 2p_C(1 - p_C)$ leading to $F_{ST}^{C,P} = \frac{F_2(C;P)}{1 - Q_2^{C,P}} = \frac{1}{2} (1 - e^{-\tau_{C,P}})$ (i.e., $F_{ST}^{C,P} \approx \frac{t}{2N_e}$ when $\tau_{C,P} \ll 1$). Hence,
 218 the estimated graph edges length $\widehat{F}_2(C; P) = \hat{e}_{P \leftrightarrow C}$ are scaled in units of drift by a factor equal to $= \frac{2}{\hat{h}_P}$ where \hat{h}_P
 219 is the estimated heterozygosity (i.e., $1 - \widehat{Q}_1^P$) in the (parent) node P . Rearranging equation 2 and using $Q_2^{C,P} = Q_1^P$
 220 (conditional on p_P) shows that $h_P = F_2(C; P) + h_C$, where $h_C = (1 - Q_1^C)$ is the heterozygosity of the child node
 221 C . Hence, all the node heterozygosities can be inferred iteratively from the leaves to the root along the admixture
 222 graph using the leave heterozygosities (directly estimated from the data) and the fitted edge lengths (Lipson *et al*,
 223 2013).

224 **2.1.7 Admixture Graph construction**

225 Comprehensive exploration of the space of possible admixture graphs rapidly becomes impossible even for a
226 moderate number of populations. We implemented in `poolfst` different heuristics to facilitate admixture graph
227 construction based on a supervised approach (see Supplementary Vignette V1 for details). First the `add.leaf`
228 function allows exploring all the possible connections of a new population to an existing admixture graph. If n_e is
229 the number of edges of the admixture graphs, $n_e + 1$ possible graphs connecting the new leaf with a non-admixed
230 edges (i.e., including a new rooting with the candidate leaf as an outgroup) and $\frac{1}{2}n_e(n_e - 1) - 1$ connecting the
231 new leaf with a two-way admixture event are then tested. Note that an admixture between the two root edges is
232 excluded from the exploration since it results in a singular model. More generally, the different possible graphs are
233 always checked for singularity by empirically verifying that the rank of the model incidence matrix $\mathbf{X}(\mathbf{a})$ is equal
234 to the number of edges to fit. The different fitted graph can then be ranked according to their *BIC*, the graph with
235 the lowest *BIC* having the strongest support.

236 The `graph.builder` function allows a larger exploration of the graph space by successively adding several
237 leaves in a given order to an existing admixture graph. At each step of the process, a heap stores the best resulting
238 graph together with some intermediary sub-optimal graphs based on their *BIC*. After initializing the heap with
239 some graph (or a list of graphs), the `add.leaf` function is called to evaluate, for each candidate leaf in turn, all
240 its possible connections (with non-admixed or admixed edges) to all the graphs stored in the heap. Among the
241 obtained graphs, the one with the lowest *BIC* together with those with a *BIC* within a given Δ_{BIC} ($\Delta_{BIC} = 6$ by
242 default) are included in a newly generated heap. If the resulting heap contains more than a predefined number of
243 graphs n_g^{\max} ($n_g^{\max} = 25$ by default), only the n_g^{\max} graphs with the lowest *BIC* are finally kept in the heap of graphs
244 to be used for the addition of the next leaf. Although helpful, such heuristic should be used cautiously and we
245 recommend to only try adding a small number of populations (i.e., ≤ 5) to an existing graph. One also needs to
246 evaluate different orders of population inclusion (Supplementary Vignettes V1 and V2).

247 It is also critical to start these supervised procedures with graphs that are representative of the whole history of
248 the populations under study and not too unbalanced with respect to the candidate leaves. In particular, starting with
249 a small tree of closely related populations which are distantly related to the candidate leaves must be avoided. When
250 prior knowledge about the history of the investigated population is limited (which is usually the case), Lipson *et al*
251 (2013) proposed to start admixture graph construction with a scaffold tree of populations displaying no evidence
252 of admixture. As in the absence of admixture, F_2 statistics are expected to be additive along the paths of the
253 (binary) population tree, its unrooted topology and branch lengths may simply be inferred with a neighbor-joining

254 algorithm. In `poolfstat`, we implemented two functions that allow i) identifying candidate sets of unadmixed
255 populations among all the genotyped ones (`find.tree.popset`); and ii) building rooted neighbor-joining tree
256 (`rooted.njtree.builder`). Briefly, `find.tree.popset` implements a procedure consisting of i) discarding all
257 the populations showing at least one significant three-population test (i.e., displaying a negative F_3 Z-score lower
258 than -1.65 by default) among all the possible ones; and ii) identifying via a greedy algorithm the largest sets of
259 populations for which all the possible quadruplets pass the four-population test of treeness (i.e., with an absolute
260 F_4 Z-score lower than 1.96 by default). The `rooted.njtree.builder` function builds a scaffold tree from a
261 candidate set of (presumably) unadmixed populations using the `nj` function from the `ape` package (Paradis *et al*,
262 2004) and then compare the consistency of population heterozygosities between the partitions of the tree to root it
263 (Lipson *et al*, 2013). Note that this latter procedure may be sensitive to long-branch attraction and should thus be
264 used carefully when including highly divergent populations.

265 **2.2 Overview of the new poolfstat package**

266 Tables 1 and 2 describe the main objects and functions implemented in our new version (v2.0.0) of the R package
267 `poolfstat` publicly available from the CRAN repository ([https://cran.r-project.org/web/packages/
268 poolfstat/index.html](https://cran.r-project.org/web/packages/poolfstat/index.html)). In-depth analyses of two Pool-Seq and allele count simulated datasets (see below)
269 are described for illustration purposes in the package vignette provided as Supplementary Vignette V1. Detailed
270 documentation page of the different objects and functions can also be directly accessed from an R terminal with
271 `poolfstat` loaded using the `help` function (or the `?` operator).

272 The package includes several functions to parse allele count (e.g., `genotremix2countdata`) or Pool-Seq
273 (e.g., `vcf2pooldata`) input data stored in various formats commonly used in population genomics studies (Ta-
274 ble 2). These functions allow to clearly distinguish these two different types of data by producing objects of either
275 the so-called `countdata` (for allele count) or `pooldata` (for Pool-Seq data) classes (Table 1). This step is critical
276 to further rely on the appropriate unbiased estimators for the F and D parameters. Some functions allow to per-
277 form subsequent manipulation of the input data, for instance to only consider some of the populations or to remove
278 SNPs according to various criteria (Table 2).

279 The three functions `computeFST`, `compute.pairwiseFST` and `compute.fstats` implement the unbiased
280 estimators of the different f -, D - and within-population heterozygosities (based on allele IIS probabilities within
281 and between pairs of populations) together with block-jackknife estimation of their standard errors. Importantly,
282 these three functions automatically detect the appropriate estimators given the type of data (either allele or Pool-

S4 object	Description
<code>countdata</code>	Standard allele count data (i.e., obtained from individual genotyping or sequencing data)
<code>pooldata</code> [†]	Pool-Seq read count data
<code>pairwisefst</code>	Store pairwise F_{ST} estimates. This object is generated by the <code>compute.pairwiseFST</code> function. Estimates can be conveniently visualized with the <code>heatmap</code> or <code>plot</code> functions, the latter interfacing the <code>plot.fstats</code> function of <code>poolfstat</code> .
<code>fstats</code>	Store F_2 , pairwise F_{ST} , F_3 , F_3^* , F_4 and D estimation results. This object is generated by the <code>compute.fstats</code> function. Estimates can be conveniently visualized with the <code>heatmap</code> or <code>plot</code> functions, the latter interfacing the <code>plot.fstats</code> function of <code>poolfstat</code> .
<code>graph.params</code>	Represent a population tree or an admixture graph and its parameter. This object is generated by the <code>generate.graph.params</code> function. The graph can be visualized with the <code>plot</code> function that interfaces the <code>grViz</code> function from the <code>DiagrammeR</code> package (Iannone, 2020).
<code>fitted.graph</code>	Represent a population tree or an admixture graph and its underlying fitted parameters as obtained from the <code>fit.graph</code> or other fitting functions. The graph can be visualized with the <code>plot</code> function that interfaces the <code>grViz</code> function from the <code>DiagrammeR</code> package (Iannone, 2020).

Table 1. Description of the main S4 objects of the `poolfstat` package. [†]Object already existing in the first `poolfstat` version.

283 Seq read counts) according to the input object class (either `countdata` or `pooldata`). For the estimation of F_{ST} ,
 284 the `computeFST` and `compute.pairwiseFST` also implement (by default) estimators based on an Analysis of
 285 Variance framework that correspond to those developed by Weir (1996) for allele count data and by Hivert *et al*
 286 (2018) for Pool-Seq data.

287 The `fit.graph` function implements the approach described above to estimate the parameters (i.e., edge
 288 lengths and admixture rates) of an admixture graph that is stored in a `graph.params` object (Table 1). Such ob-
 289 jects can be generated with the `generate.graph.params` function (Table 2) to include the target basis f -statistics
 290 and the error covariance matrix (denoted above $\widehat{\mathbf{f}}$ and \mathbf{Q} , respectively) estimated with `compute.fstats` (stored
 291 in an `fstats` object) and to specify the topology and the parameters of the admixture graph. Note that the
 292 `graph.params2symbolic.fstats` function allows exploring in details the properties of an admixture graph
 293 specified by a `graph.params` object by deriving a symbolic representation of all the F_2 , F_3 , F_4 and the model
 294 equations (see above) by internally relying on the `Ryacas` package for symbolic computation (Andersen & Højsgaard,
 295 2019). The `fit.graph` function then produced an object of class `fitted.graph` that includes the estimated edge
 296 lengths (in F_2 and also optionally in drift units) and admixture proportions together with (optionally) their 95% CI.
 297 For model fit assessment purposes, `fitted.graph` objects also include the BIC and Z-score of the residuals of the
 298 fitted basis f -statistics. Such a comparison can (and should) be generalized to all the f_2 , f_3 and f_4 statistics (not just
 299 the ones forming the basis) using `compare.fitted.fstats` jointly applied to a `fitted.graph` and a `fstats` ob-

Function	Type	Detail
<code>genotreeMix2countdata</code>	data import	Generate a <code>countdata</code> object (Table 1) from allele count data stored in <code>TreeMix</code> format (Pickrell & Pritchard, 2012)
<code>genobypass2countdata</code>	data import	Generate a <code>countdata</code> object (Table 1) from allele count data stored in <code>BayPass</code> format (Gautier, 2015)
<code>vcf2pooldata*</code>	data import	Generate a <code>pooldata</code> object (Table 1) from Pool-Seq read count data stored in a <code>vcf</code> file generated by commonly used calling software as <code>VarScan</code> (Koboldt <i>et al.</i> , 2012), <code>bcftools</code> (Li <i>et al.</i> , 2009), <code>GATK</code> (McKenna <i>et al.</i> , 2010) or <code>FreeBayes</code> (Garrison & Marth, 2012). Parsing of <code>vcf</code> files uses C++ routines inspired by the <code>vcfR</code> package (Knaus & Grünwald, 2017).
<code>popsync2pooldata†</code>	data import	Generate a <code>pooldata</code> object (Table 1) from Pool-Seq read count data stored in the <code>sync</code> <code>Popoolation</code> format (Kofler <i>et al.</i> , 2011)
<code>genobypass2pooldata†</code>	data import	Generate a <code>pooldata</code> object (Table 1) from Pool-Seq read count data stored in <code>BayPass</code> format (Gautier, 2015)
<code>countdata.subset</code>	data handling	Subsets a <code>countdata</code> object (Table 1) according to various criteria (e.g., population or SNP indexes, marker polymorphism, call rate)
<code>pooldata.subset†</code>	data handling	Subsets a <code>countdata</code> object (Table 1) according to various criteria (e.g., population or SNP indexes, marker polymorphism, pool coverage)
<code>pooldata2genobypass†</code>	data handling	Export Pool-Seq read count data stored in a <code>pooldata</code> object in <code>BayPass</code> format (Gautier, 2015).
<code>computeFST*</code>	estimation	Estimate genome-wide F_{ST} over all the populations.
<code>compute.pairwiseFST*</code>	estimation	Estimate pairwise-population population F_{ST} . The function generates a <code>pairwiseFST</code> object (Table 1).
<code>compute.fstats</code>	estimation	Estimate f_2 , pairwise F_{ST} , f_3 , f_4^* and f_4 , D -statistics and within-population heterozygosities together with their standard errors via block-jackknife. The function generates an <code>fstats</code> object (Table 1).
<code>compute.f4ratio</code>	estimation	Estimate admixture rates and their standard errors using ratios of f_4 -statistics from an <code>fstats</code> object (Table 1).
<code>generate.graph.params</code>	graph fitting	Generate a <code>graph.params</code> object synthesising the structure and properties of an admixture graph (Table 1).
<code>fit.graph</code>	graph fitting	Estimate parameters (edge lengths and admixture rates) of an admixture graph. The function generates a <code>fitted.graph</code> object (Table 1).
<code>compare.fitted.fstats</code>	graph fitting	Compare all the fitted f_2 , f_3 and f_4 derived from a fitted graph and stored in a <code>fitted.graph</code> object (Table 1) to the estimated ones (stored in a <code>fstats</code> object). This function allows graph fitting assessment.
<code>find.tree.popset</code>	graph building	Find sets of populations that may be used as scaffold tree based on the estimated f_3 and f_4 stored in a <code>fstats</code> object (Table 1).
<code>rooted.njtree.builder</code>	graph building	Construct and root a <code>Neighbor-Joining</code> tree of presumably unadmixed scaffold populations. This function generates a <code>fitted.graph</code> object (Table 1).
<code>add.leaf</code>	graph building	Evaluate all possible connections of a new leaf (i.e., genotyped population) to an existing graph (stored in either a <code>graph.params</code> or <code>fitted.graph</code> object) with both non-admixed and admixed edges. This function generates a list of <code>fitted.graph</code> objects including other information (e.g., BIC of all the possible graphs, index of the best fitted graph)
<code>graph.builder</code>	graph building	Implement a graph builder heuristic by successively adding leaves (i.e., genotyped populations) to an existing graph or tree (stored in either a <code>graph.params</code> or <code>fitted.graph</code> object) by successively calling the <code>add.leaf</code> functions and keeping sub-optimal graphs in a heap at each step. This function generates a list of <code>fitted.graph</code> objects including other information (e.g., BIC of all the possible graphs, index of the best fitted graph)
<code>plot.fstats</code>	visualization	Plot f -statistics with their Confidence Intervals (can be called directly using <code>plot</code>)
<code>graph.params2symbolic.fstats</code>	Other utilities	Derive a symbolic representation of F_2 , F_3 and F_4 (and the graph model system of equation) as a function of admixture graph parameters specified in a <code>graph.params</code> object (Table 1) using functionalities of the <code>Ryacas</code> package (Andersen & Højsgaard, 2019).

Table 2. Description of the main `poolfstat` functions. † and * objects existing or significantly improved since the first `poolfstat` version, respectively.

jects. Notice that we developed for comparison purposes a function named `graph.params2qpGraphFiles` to export admixture graph specification and their underlying estimated basis f -statistics (both stored in a `graph.params` object) into `qpGraph` format (Patterson *et al*, 2012), allowing independent fitting based on the same estimated statistics to be carried out with this later program.

The `poolfstat` package includes several functions to assist construction of admixture graphs. As mentioned in the previous section, the `find.tree.popset` and `rooted.njtree.builder` functions allow to identify and build rooted tree(s) of scaffold of (presumably) unadmixed populations that may be used as starting graph(s). Besides, the `add.leaf` and `graph.builder` functions implement the above described heuristic to extent an existing graph (or tree) by adding one or several leaves (i.e., genotyped populations). These functions generate a list of `fitted.graph` objects together with other information that may be helpful for graph comparison (e.g., *BIC* of all the graphs or index of the best fitted graph).

Finally, as detailed and exemplified in the Supplementary Vignette V1, fitted graphs (stored in `fitted.graph` objects) and non-fitted graphs (stored in `graph.params` objects) can be directly and conveniently plotted with the `plot` function which internally interfaces the `grViz` function from the `DiagrammeR` package (Iannone, 2020).

2.3 Data analyses

2.3.1 Simulation study

Genetic data for a total of 150 diploid individuals belonging to six different populations (n=25 individuals per populations) related by the demographic scenario depicted in Figure 1 were simulated using the `msprime` coalescent simulator (Kelleher *et al*, 2016) with the following command:

```
mspms 300 20 -t 4000 -I 6 50 50 50 50 50 50 0 -r 4000 1000000000 -p 8 -es 0.0125 6 0.25  
-ej 0.0125 6 2 -ej 0.0125 7 3 -ej 0.025 2 1 -ej 0.05 3 1 -ej 0.075 5 4 -ej 0.1 4 1
```

Each genome thus consisted of 20 independent chromosomes of $L = 100$ Mb assuming a scaled chromosome-wide recombination rate of $\rho = 4LN_e r = 4,000$ as expected for instance in a population of constant diploid effective size of $N_e = 10^3$ when the per-base and per-generation recombination rate is $r = 10^{-8}$ (i.e., one cM per Mb). The scaled chromosome-wide mutation rate was set to $\theta = 4LN_e \mu = 4,000$ which is also the expected nucleotide diversity in a population with $N_e = 10^3$ at mutation-drift equilibrium when the per-base mutation rate is $\mu = 10^{-8}$. A total of 250 independent genotyping datasets were simulated and each was subsequently processed to generate 32 different types of datasets corresponding to:

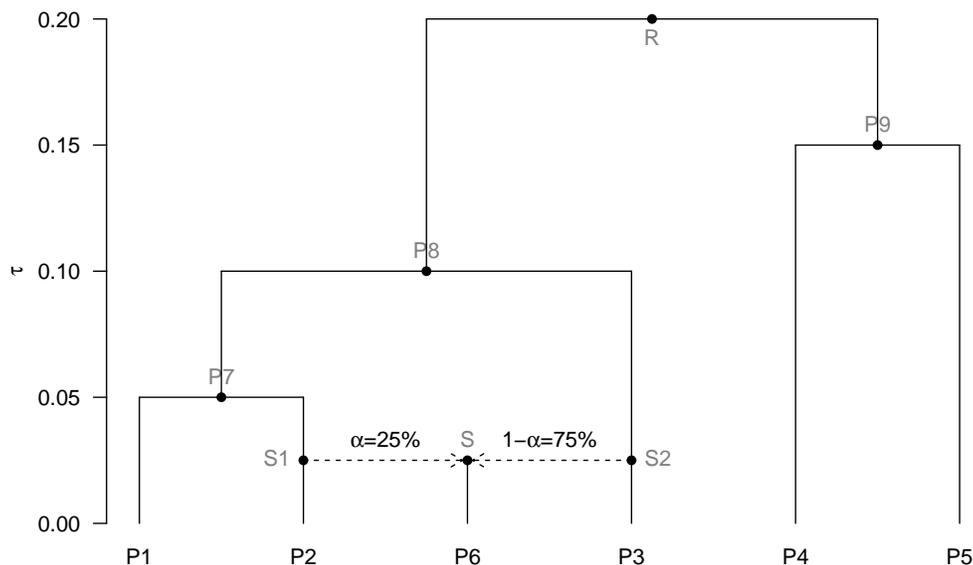


Figure 1. Simulated scenario relating six sampled populations. The population $P6$ derived from a population S which is admixed between two ancestral sources ($S1$ and $S2$) directly related to populations $P2$ and $P3$ and contributing to $\alpha = 25\%$ and $1 - \alpha = 75\%$ of its genome, respectively. The branch lengths are in a diffusion timescale i.e., with $\tau = \frac{t}{2N_e}$ under a pure-drift model of divergence (where t is the number of non-overlapping generations and N_e the average diploid effective population sizes along the branch). The names of the internal node populations (not sampled) are represented in grey.

- 328 • Two standard allele count datasets (namely $AC_{m \geq 1\%}$ and $AC_{m \geq 5\%}$) obtained by simple counting of the simu-
329 lated individual (haploid) genotypes for each population (i.e., assuming Hardy-Weinberg equilibrium within
330 population) and removing SNPs with a Minor Allele Frequency (MAF) computed over all the individuals
331 lower than 1% (for $AC_{m > 1\%}$ datasets) or 5% (for $AC_{m > 5\%}$ datasets)
- 332 • Thirty Pool-Seq datasets (coded as $PS_{m > m_t, \%}^{\lambda, \epsilon = \epsilon}$) for i) five different average sequencing coverages λ (equal to
333 30, 50, 75, 100 or 200 reads; a 30X Pool-Seq coverage representing a lower limit for population genomics
334 studies); ii) two different MAF thresholds m_t of 1% and 5% (MAF being estimated on the read counts over
335 all the pools); and iii) three different sequencing error rates ϵ of 0 (no error), 1‰ and 2.5‰ the two latter
336 being representative of Illumina sequencers (Glenn, 2011).

337 Pool-Seq datasets were simulated from the $AC_{m \geq 1\%}$ allele count datasets following a procedure similar to that
338 described in Hivert *et al* (2018). Briefly, the vector $r_{ij} = \{r_{ijk}\}$ of read counts at SNP position i in population j
339 for the nucleotide k (where by convention $k = 1$ and $k = 2$ for the derived and ancestral alleles respectively) was
340 sampled from a Multinomial distribution parameterized as:

$$r_{ij} \sim \text{Multin} \left(\left\{ \frac{y_{ij}}{n_j} \left(1 - \frac{\epsilon}{3}\right) + \left(1 - \frac{y_{ij}}{n_j}\right) \frac{\epsilon}{3}; \left(1 - \frac{y_{ij}}{n_j}\right) \left(1 - \frac{\epsilon}{3}\right) + \frac{y_{ij}}{n_j} \frac{\epsilon}{3}; \frac{\epsilon}{3}; \frac{\epsilon}{3} \right\}; c_{ij} \right)$$

341 where y_{ij} is the derived allele count for SNP i in population j (from the corresponding $AC_{m>1\%}$ dataset); n_j is
 342 the haploid sample size of population j (here $n_j = 50$ for all j); and $c_{ij} = \sum_{k=1}^{k=4} r_{ijk}$ is the overall read coverage. To
 343 introduce variation in read coverages across pools and SNPs, each c_{ij} was sampled from a Poisson distribution with
 344 a parameter λ (the target Pool-Seq mean coverage). When $\epsilon = 0$, only reads for the derived ($k = 1$) or ancestral
 345 ($k = 2$) alleles can be generated and the above Multinomial sampling actually reduces to a Binomial sampling
 346 following $r_{ij1} \sim \text{Bin} \left(\frac{y_{ij}}{n_j}; c_{ij} \right)$ (and $r_{ij2} = c_{ij} - r_{ij1}$). However, when $\epsilon > 0$, sequencing errors might lead to non-null
 347 read counts for the two other alleles leading to tri- or tetra- allelic SNPs. Moreover, sequencing errors may also
 348 introduce spurious additional variation by generating false SNPs at monomorphic sites. To account for the latter,
 349 read count vectors $r_{i'j}$ for all the $2 \times 10^9 - I$ monomorphic positions i' (where I is the number of SNPs observed in
 350 the considered $AC_{m>1\%}$ dataset) were sampled as $r_{i'j} \sim \text{Multin} \left(\left\{ 1 - \frac{\epsilon}{3}; \frac{\epsilon}{3}; \frac{\epsilon}{3}; \frac{\epsilon}{3} \right\}; c_{i'j} \right)$ with coverages $c_{i'j}$ sampled
 351 from a Poisson distribution (as c_{ij} for polymorphic positions). Yet, as usually done with empirical datasets, we
 352 applied a minimum read count filtering step consisting of disregarding all the alleles with less than 2 observed
 353 reads (over all the populations). Only bi-allelic SNPs passing the overall MAF threshold m_f were finally retained
 354 in the final $PS\lambda_{m \geq m_f, \%}^{\epsilon = \epsilon}$ datasets.

355 Analyses of the simulated data were carried out with `poolfstat` (Supplementary Vignette V1). Briefly, each
 356 `msms` simulated dataset was converted into an $AC_{m \geq 1\%}$ dataset in `TreeMix` format (Pickrell & Pritchard, 2012)
 357 further imported into a `countdata` object with `genotreeemix2countdata` (Tables 1 and 2) and used to generate
 358 each corresponding $AC_{m \geq 5\%}$ dataset using `countdata.subset`. To improve computational efficiency, the different
 359 $PS\lambda_{m > m_f, \%}^{\epsilon = \epsilon}$ Pool-Seq datasets were generated from the $AC_{m \geq 1\%}$ `countdata` objects in the form of `pooldata` object
 360 using custom functions (not included in the package) coded in C++ and integrated within R using `Rcpp` (Eddelbuet-
 361 tel, 2013). In addition, to evaluate the impact of the (bad) practice consisting of analyzing Pool-Seq data as if they
 362 were allele count data (i.e., overlooking the sampling of reads from individual genes of the pool), we also created
 363 “fake” `countdata` objects from the different `pooldata` objects. We then used default options (unless otherwise
 364 stated) of i) `computeFST` to estimate genome-wide F_{ST} over all the populations; ii) `compute.fstats` to estimate
 365 all the f - and D - statistics; iii) `compute.f4ratio` to estimate admixture proportions; and iv) `fit.graph` to esti-
 366 mate the admixture graph parameters (Table 2). As the number of SNPs was variable across the different simulated
 367 datasets, we adjusted the number of successive SNPs defining a block for block-jackknife estimation of standard

368 errors by dividing the total number of available SNPs by 500. This thus resulted on average in 490 blocks of 4.1
369 Mb over the genome for all the analyzed datasets (the simulated genomes consisting of 20 chromosomes). Note
370 that the parameter estimates were always taken as the block-jackknife mean values rather than estimates over all
371 SNPs (i.e., including those in the chromosome ends). In practice, the differences between the two are insignificant
372 (e.g., Supplementary Vignette V1).

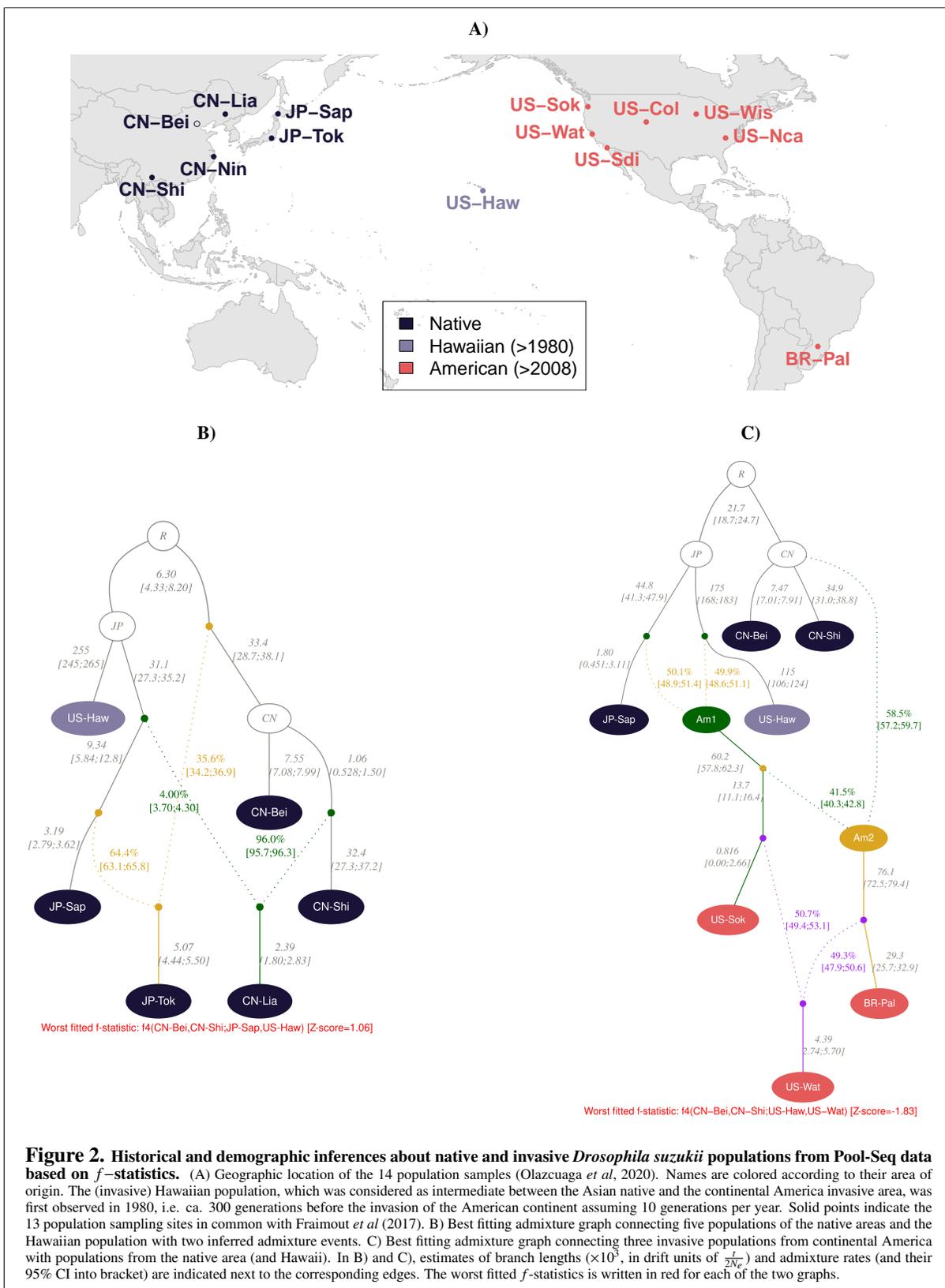
373 For validation purposes, we also analyzed the 250 $AC_{m \geq 1\%}$ datasets with programs from the `AdmixTools` suite
374 (Patterson *et al*, 2012) after conversion to the appropriate input format using custom `awk` scripts. More specifically,
375 we ran `qpfstats` (v. 200) to estimate the 15 basis f -statistics, i.e., taking $P1$ as the reference population, the five
376 f_2 of the form (P1,Px) and the ten f_3 of the form (P1;Px,Py) (where $x = 2, \dots, 6$ and $y = 3, \dots, 6$ with $y \neq x$)
377 and their corresponding error covariance matrix. Default options were considered except for the disabling of the
378 scaling of estimated values (using option `-l 1`) to facilitate their comparison with `poolfstat` estimates. We also
379 ran with default options `qp3Pop` (v. 650) to estimate f_3^* for all the 60 possible triplet configurations and `qpDstat`
380 (v. 970) to estimate the D -statistics for all the 45 possible quadruplet configurations together with their associated
381 Z -scores. By default, these three programs define blocks of 5 cM to implement the (weighted) block-jackknife
382 procedure. As we here converted the simulated SNP positions from Mb to cM assuming one cM per Mb (see
383 above), the sizes of the 400 blocks was thus about 20% than for `poolfstat` analyses.

384 **2.3.2 Analysis of a real *Drosophila suzukii* Pool-Seq data**

385 The spotted wing drosophila, *Drosophila suzukii*, represents an attractive model to study biological invasion and
386 hence recent historical and demographic history. Native to South East Asia, this pest species was first observed
387 outside its native range in Hawaii in 1980, and later rapidly invaded America and Europe simultaneously between
388 2008 and 2013 (Fraimout *et al*, 2017). Using DNA sequences and microsatellite markers, Adrion *et al* (2014)
389 and Fraimout *et al* (2017) deciphered the routes taken by *D. suzukii* during its worldwide invasion. Both studies
390 showed that America and European populations globally represent separate invasion routes with different native
391 source populations. Olazcuaga *et al* (2020) recently generated Pool-Seq genomic data from 22 worldwide pop-
392 ulation samples to detect genetic variants associated with the historical status (i.e. invasive versus native) of the
393 sampled populations. We here focused our illustration on 14 Pool-Seq data from this study (with 50 to 100 diploid
394 individuals per pool) for populations representative of the Asian native area (six populations), Hawaii (one popu-
395 lation) and the invaded continental America (seven populations), where the species was first observed in 2008 on
396 the Western coast of the USA (around Watsonville, CA; Figure 2A). Beside native populations, we have restricted

397 our analysis to the American continent because the invasion of this area is characterized by multiple admixture
398 events between different source populations (Fraitout *et al*, 2017), which makes it an appealing situation to eval-
399 uate the power and the limitation of `poolfstat` analyzes. Moreover, 13 of our 14 population samples consist of
400 individuals originating from the same sites (albeit sometimes collected at different dates for some pools; Table 2
401 in Supplementary vignette V2) as those genotyped at 25 microsatellite markers and analyzed with an Approximate
402 Bayesian Computation Random Forest (ABC-RF) approach to infer the routes of invasion on a worldwide scale by
403 Fraitout *et al* (2017).

404 To allow for complete reproduction (and exploration) of our analyses, all the command lines used to analyze the
405 *D. sukukii* Pool-Seq dataset are described in the Supplementary vignette V2. Briefly, we combined the 14 (bam)
406 files, obtained by Olazcuaga *et al* (2020) after aligning the 14 Pool-Seq data onto the latest near-chromosome
407 scale *D. sukukii* assembly (Paris *et al*, 2020), into an `mpileup` file using `SAMtools` 1.9 with options `-q 20 -Q20`
408 (Li *et al*, 2009). Variant calling was then performed using `VarScan mpileup2snp` v2.3.4 Koboldt *et al* (2012)
409 run with options `--min-coverage 10 --min-avg-qual 25 --min-var-freq 0.005 --p-value 0.5` (i.e.,
410 with very loose criteria). After discarding positions mapping to non-autosomal contigs (Paris *et al*, 2020), the
411 resulting `vcf` file was parsed with the `vcf2pooldata` function of `poolfstat` with default options except for i) the
412 overall MAF threshold (computed from read counts) that was set to 5%; and ii) the minimal read coverage for each
413 pool that was set to 50. The resulting `pooldata` object was further filtered with `pooldata.subset` to discard i)
414 all positions with a coverage higher than the 99th coverage percentile within at least one pool; and ii) discard all
415 SNPs with $MAF < 5\%$ over all the populations from the native area to favor ancestral SNPs. The final dataset then
416 consisted of read counts for 1,588,569 bi-allelic SNPs with a median read coverage varying from 64 (US-Sok) to
417 95 (CN-Bei and US-Haw) among the 14 pools (Table 2 in the Supplementary Vignette V2). We defined blocks of
418 10,000 consecutive SNPs for block-jackknife estimation of standard errors leadint to a total of 145 blocks of 698
419 kb on average (varying from 414 kb to 2.03 Mb). Hence, most analyses actually relied on 1,450,000 SNPs that
420 mapped to the 15 largest contigs of the assembly (totaling 116 Mb). In other words, SNPs mapping to the smallest
421 (and less reliable) contigs were discarded in addition to the few ones mapping to the end of the 15 retained contigs.



3 Results

3.1 Evaluation of poolfstat on simulated data

Historical and demographic inference based on f - and D - statistics has already been extensively evaluated in previous studies (e.g., Patterson *et al*, 2012; Lipson *et al*, 2013; Peter, 2016). Therefore, the purpose of our simulation study was essentially threefold: i) to validate the estimators implemented in poolfstat by comparing, for allele count data, with those obtained with the reference AdmixTools suite (Patterson *et al*, 2012); ii) to evaluate the performance of the estimators for Pool-Seq data as a function of read coverage and sequencing errors; and iii) to provide example datasets with known ground truth for illustration purposes.

3.1.1 Description of the simulated datasets

We simulated 250 genetic datasets for six populations (named $P1$ to $P6$) each consisting of 25 diploid individuals and that were historically related by the admixture graph represented in Figure 1 (Material and Methods). Each of these datasets was further used as template to generate 2 allele count datasets (applying 1% or 5% threshold on the overall MAF for $AC_{m>1\%}$ and $AC_{m>5\%}$ datasets respectively) and to simulate 30 Pool-Seq datasets with five different mean read coverages ($\lambda \in \{30; 50; 75; 100; 200\}$); three sequencing error rates ($\epsilon \in \{0; 10^{-3}; 2.5 \times 10^{-3}\}$) and two MAF (computed over all read counts) thresholds (referred to as $PS\lambda_{m>1\%}^{\epsilon=\epsilon}$ and $PS\lambda_{m>5\%}^{\epsilon=\epsilon}$ for 1% and 5% MAF thresholds, respectively). This thus lead to a total of 8,000 simulated datasets. The average number of available SNPs and false SNPs (for $PS\lambda_{m>1\%}^{\epsilon=1\%}$ and $PS\lambda_{m>5\%}^{\epsilon=2.5\%}$ datasets) is given in Table S1 for each of the 32 different types of datasets and represented as a function of the mean coverages λ and MAF thresholds in Figure S1.

Overall, 471,919 SNPs and 240,369 SNPs were available on average for allele count datasets at the 1% ($AC_{m>1\%}$) and 5% ($AC_{m>5\%}$) MAF thresholds respectively consistent with the L-shaped distribution of allele frequencies (Figure S2A). As expected from binomial sampling (Figure S2B), for Pool-Seq datasets generated with no sequencing error, the number of SNPs remained always lower than the $AC_{m>1\%}$ datasets at the 1% MAF threshold although increasing with coverages from 13.8% for $PS30_{m>1\%}^{\epsilon=0}$ to 2.01% for $PS200_{m>1\%}^{\epsilon=0}$ datasets (see Table S1 legend for details). Conversely, at the 5% MAF threshold, the number of SNPs was slightly higher than the $AC_{m>5\%}$ datasets (from 2.58% for $PS30_{m>5\%}^{\epsilon=0}$ to 1.51% for $PS200_{m>5\%}^{\epsilon=0}$) which is related to i) the shape of the allele frequency spectrum (stochastic variation in read sampling leading to include more SNPs with $0.01 < MAF < 0.05$ than exclude SNPs with $MAF > 0.05$ from the simulated genotyping data because they are more numerous); and ii) variation in the simulated read coverages that explains the decreasing trend with λ .

450 With sequencing errors, our filtering steps proved efficient to remove false SNPs except at the 1% MAF thresh-
451 old when $\epsilon = 2.5\%$ or when $\epsilon = 1\%$ at the lowest coverage ($\lambda = 30$ and $\lambda = 50$). These configurations displayed
452 substantial to very high proportions of false SNPs (up to 93.8% for $PS50_{m>1\%}^{\epsilon=2.5\%}$) although decreasing with cover-
453 age (Figure S1B). A 5% MAF threshold always resulted in the complete removal of all the false SNPs for all the
454 investigated scenarios (Table S1). Note that for the highest coverages, sequencing errors lead to a relative reduction
455 of the number of SNPs due to the generation of spurious tri- or tetra- allelic SNPs from the simulated bi-allelic
456 SNPs (compare e.g., $PS200_{m>5\%}^{\epsilon=2.5\%}$ and $PS100_{m>5\%}^{\epsilon=2.5\%}$ on Figure S1A).

457 3.1.2 Comparison of poolfstat and Admixtools estimates for allele count data

458 We first analyzed the 250 simulated $AC_{m>1\%}$ datasets to estimate with both poolfstat and Admixtools programs
459 i) the 15 basis f -statistics (taking $P1$ as the reference population) consisting of five f_2 and ten f_3 (Figure 3A) and
460 their corresponding error covariance matrix (Figure 3B); ii) the 60 f_3^* (Figure 3C) and their associated Z-scores
461 (Figure 3D); and iii) the 45 D -statistics (Figure 3E) and their associated Z-scores (Figure 3F). The estimates
462 were all found in almost perfect agreement between the two implementations with Mean Absolute Differences
463 (MAD) negligible when compared to the range of variation of the underlying values. For f - and D - statistics,
464 slight differences were mostly due to the plotted poolfstat estimates corresponded to block-jackknife means
465 (i.e., excluding SNPs outside blocks as those from chromosome ends). Using poolfstat estimates based on all
466 the SNPs indeed resulted in almost null MAD (MAD' in Figure 3A, C and E), up to rounding errors due to lower
467 decimal precision in the printed output of the Admixtools programs. Note that the differences in block-jackknife
468 implementation among the two programs (Material and Methods) had very minor impact on the estimation of error
469 variance and covariance of the estimates (Figures 3B). Accordingly, the MAD computed on Z-scores remained
470 very small (although inflated for higher values) and Z-score based decision for the underlying three-population
471 admixture (Figures 3D) or four-population treeness tests (Figures 3F) were highly consistent (with a proportion
472 $\beta = 97.7\%$ and $\beta = 98.0\%$ respectively of Z-scores significant with the two programs among the ones significant
473 with at least one program).

474 3.1.3 Performance of f_3 and f_3^* based tests of admixture and f_4 and D - based tests of treeness for allele 475 count and Pool-Seq data

476 We ran the compute.fstats function on all the simulated allele count and Pool-Seq datasets to estimate all
477 f - and D - statistics. To further evaluate the impact of (improperly) treating read counts as allele counts when

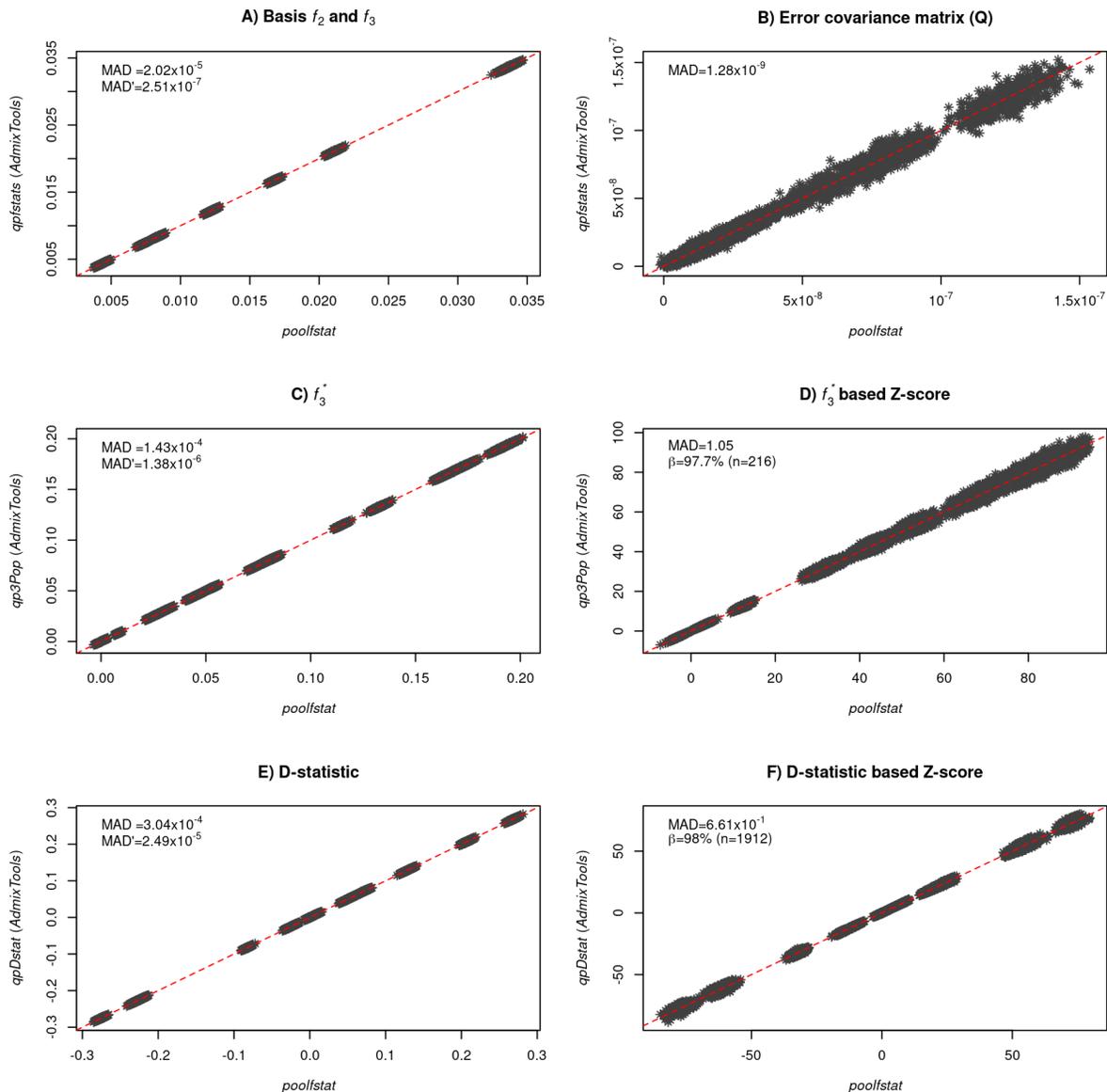


Figure 3. Comparison of poolfstat and AdmixTools estimates across 250 simulated allele count datasets ($AC_{m \geq 1\%}$). A) All estimates of the 15 basis f -statistics taking P_1 as the reference population and corresponding to 5 f_2 of the form (P_1, P_x) and the 10 f_3 of the form $(P_1; P_x, P_y)$ (with $x = 2, \dots, 6$; $y = 3, \dots, 6$ and $y > x$). B) All Block-jackknife estimates of the covariance matrix Q of the 15 basis f -statistics (15 error variances and 105 error covariances). C) All estimates of the 60 f_3^* (scaled f_3) and their associated Z-scores (D). E) All estimates of the 45 D -statistics (scaled f_4) and their associated Z-scores (F). For each comparison, the Mean Absolute Difference (MAD) between the parameter estimates of the two programs are given on the upper left corner of the plots. In A), C) and E), poolfstat estimates correspond to block-jackknife means (i.e., they only include SNPs eligible for block-jackknife). The given MAD' value is the MAD between AdmixTools and poolfstat estimates that include all SNPs (see documentation for the `compute.fstats` function). In D), a consistency score β is also given and was computed as the proportion of Z-scores < -1.65 (i.e., significant three-population test of admixture at a 5% threshold) with both programs among the $n = 216$ ones significant in at least one of the two programs. Similarly, in F), the given consistency score β is computed as the proportion of absolute Z-scores < 1.96 (i.e., passing the four-population treeness at a 5% threshold) with both programs among the $n = 1,912$ ones with an absolute Z-scores < 1.96 in at least one of the two programs)

478 analyzing Pool-Seq data we also analyzed the simulated Pool-Seq datasets (focusing only on PS $\lambda_{m>m_t\%}^{\epsilon=0}$ datasets,
479 i.e., simulated without sequencing error) as if they were allele count data. Overall, 42 different configurations were
480 thus investigated each originating from the 250 allele count datasets simulated under the demographic scenario
481 represented in Figure 1, leading to a total of $42 \times 250 = 10,500$ analyses.

482 Tables 3 and S2 provide the estimated power (True Positive Rate, TPR) and False Positive Rate (FPR) of the
483 f_3 - and f_3^* -based test of admixture for each configurations. As P6 was the only admixed population, each TPR was
484 estimated as the proportion of f_3 (respectively f_3^*) with an associated Z-score < -1.65 (95% significance thresh-
485 old) for the (P6;P2,P3) population triplet (i.e., among 250 estimates). Conversely, the FPR was computed as the
486 proportion of f_3 (respectively f_3^*) with an associated Z-score < -1.65 among all the 50 population triplets that do
487 not involve P6 as a target (i.e., among $12,250 = 250 \times 50$ estimates). Consistent with Patterson *et al* (2012), the per-
488 formance of f_3 - and f_3^* -based test of admixture were virtually the same for all the configurations. When the same
489 MAF threshold was applied, the performance of Pool-Seq data generated with no sequencing error were very close
490 to that obtained with allele count data although the power tended to slightly decrease with decreasing sequencing
491 coverage. Interestingly, increasing the MAF threshold from 1% to 5% increased the power by more than 10% and
492 in all cases, no false positive signal of admixture was detected. Surprisingly, sequencing errors in Pool-Seq data
493 also tended to increase the power from a negligible amount (less or close to 1%) at 5% MAF threshold to a quite
494 substantial amount at 1% MAF threshold (decreasing with coverage and increasing with sequencing error rate).
495 At the extreme, a power of 100% was even observed when $\lambda \leq 50$ and $\epsilon \geq 1\%$. This trend was actually directly
496 related to the proportion of false SNPs introduced by sequencing error (Figure S1B) that resulted in a downward
497 bias of f_3 and f_3^* estimates, although the underlying tests remained robust as all the estimated FPR were null except
498 for PS50 $\epsilon=2.5\%$ $_{m>1\%}$ data (FPR=6.47%) which displayed the highest proportion of false SNPs ($> 90\%$, Figure S1B).
499 However, this observed apparent robustness of the three-population tests to false SNPs should be interpreted cau-
500 tiously since it may rather result from the moderate to high expected f_3 and f_3^* values in our simulated scenario
501 for the population triplets that do not involve P6 as a target. Overall, applying a 5% MAF threshold on Pool-Seq
502 data (even with $\epsilon=2.5\%$) to remove false SNPs (see above) allowed recovering the performances similar to that
503 obtained when analyzing datasets with no sequencing error. Finally, it is worth stressing that analyzing Pool-Seq
504 data as allele counts, whatever the coverage or MAF threshold considered, lead to no power in detecting admixture
505 event with f_3 or f_3^* based tests due to a strong upward estimation bias.

506 Tables 4 and S3 similarly provide the estimated power (TPR) and FPR of the f_4 - and D -based tests of treeness
507 for the 42 configurations investigated in the simulation study. Given the simulated scenario, eight of the 45 differ-

MAF threshold	seq. error ϵ	Pool-Seq (read counts) data					allele count data
		$\lambda = 30$	$\lambda = 50$	$\lambda = 75$	$\lambda = 100$	$\lambda = 200$	
>1%	0	82.0 (0.00)	84.4 (0.00)	86.0 (0.00)	86.0 (0.00)	85.2 (0.00)	85.6 (0.00)
	1‰	100 (0.00)	100 (0.00)	86.8 (0.00)	87.2 (0.00)	86.4 (0.00)	
	2.5‰	100 (0.00)	100 (6.47)	99.6 (0.00)	92.8 (0.00)	88.4 (0.00)	
	0	<i>0.00 (0.00)*</i>	<i>0.00 (0.00)*</i>	<i>0.00 (0.00)*</i>	<i>0.00 (0.00)*</i>	<i>0.00 (0.00)*</i>	
>5%	0	93.6 (0.00)	95.2 (0.00)	96.4 (0.00)	96.0 (0.00)	96.0 (0.00)	96.8 (0.00)
	1‰	94.0 (0.00)	96.8 (0.00)	96.4 (0.00)	97.2 (0.00)	96.8 (0.00)	
	2.5‰	94.0 (0.00)	96.0 (0.00)	96.0 (0.00)	97.2 (0.00)	96.8 (0.00)	
	0	<i>0.00 (0.00)*</i>	<i>0.00 (0.00)*</i>	<i>0.00 (0.00)*</i>	<i>0.00 (0.00)*</i>	<i>0.00 (0.00)*</i>	

Table 3. Comparison of the performance of f_3 -based tests of admixture for different types of data simulated under the Figure 1 scenario processing poolfstat analyses. For each MAF threshold (MAF > 1% or MAF > 5%), the table gives True and False (in parenthesis) Positive Rates (in %) for 21 different types of analyses relying on i) allele count data; ii) 15 different Pool-Seq read count data (five mean coverages λ and three sequencing error rates ϵ); and iii) Pool-Seq read count data simulated with $\epsilon = 0$ treated as allele counts (corresponding results of this bad practice are highlighted in italics and *). Each TPR was computed from the analysis of 250 independent datasets (generated from the data simulated under Figure 1 demographic scenario) as the proportion of f_3 with an associated Z-score < -1.65 (95% significance threshold) for the (P6;P2,P3) population triplet (n=250 estimates). The FPR was similarly computed as the proportion of f_3 with an associated Z-score < -1.65 among all the 50 population triplets that do not involve P6 as target population (n=250×50=12,250 estimates).

ent population quadruplets (namely (P1,P2;P3,P4); (P1,P2;P3,P5); (P1,P2;P4,P5); (P1,P3;P4,P5); (P1,P6;P4,P5);
(P2,P3;P4,P5); (P2,P6;P4,P5); and (P3,P6;P4,P5)) have a null expected F_4 (and D) value. Note that this may
easily be shown with the symbolic calculus derivation implemented in `graph.params2symbolic.fstats` (Ta-
ble 2). For each configuration, the TPR of the treeness test was then estimated as the proportion of f_4 (respectively
 D) with an associated absolute Z-score < 1.96 (95% significance threshold) for these eight population quadru-
plets ((P1,P2;P3,P4); (P1,P2;P3,P5); (P1,P2;P4,P5); (P1,P3;P4,P5); (P1,P6;P4,P5); (P2,P3;P4,P5); (P2,P6;P4,P5);
(P3,P6;P4,P5)) over all the 250 different underlying analyses (i.e., among 2,000=250×8 estimates). Conversely,
the FPR was estimated as the proportion of f_4 (respectively D) with an associated absolute Z-score < 1.96 among
all the 37 remaining population quadruplets (i.e., among 9,250=250×37 estimates). The power for both F_4 - and
 D -based tests were remarkably consistent across all the different configurations. In addition, the tests were all
found almost perfectly calibrated since the estimated power were close to 95%, the probability of rejecting the null
hypothesis at the chosen 95% significance threshold for Z-scores. Likewise, all FPR remained low ($\leq 0.15\%$),
although increasing with MAF thresholds (more than twice higher for a given type of data when increasing the
MAF threshold from 1% to 5%). Overall, sequencing errors and coverage had no impact on the performance of
the f_4 - and D -based test of treeness. As expected, analyzing read counts as allele count data did not affect the
performance of these tests (see Discussion).

MAF threshold	seq. error ϵ	Pool-Seq (read counts) data					allele count data
		$\lambda = 30$	$\lambda = 50$	$\lambda = 75$	$\lambda = 100$	$\lambda = 200$	
>1%	0	94.0 (0.05)	94.4 (0.06)	94.1 (0.04)	94.5 (0.05)	94.3 (0.02)	94.2 (0.02)
	1‰	94.3 (0.04)	94.2 (0.03)	94.3 (0.03)	94.3 (0.03)	94.1 (0.05)	
	2.5‰	94.8 (0.06)	94.5 (0.05)	94.8 (0.03)	94.5 (0.06)	94.3 (0.04)	
	0	<i>94.0 (0.05)*</i>	<i>94.4 (0.06)*</i>	<i>94.1 (0.04)*</i>	<i>94.5 (0.05)*</i>	<i>94.3 (0.02)*</i>	
>5%	0	94.5 (0.14)	94.3 (0.11)	94.1 (0.14)	94.9 (0.09)	94.3 (0.08)	94.3 (0.11)
	1‰	94.5 (0.09)	94.5 (0.11)	94.5 (0.13)	94.2 (0.09)	94.1 (0.15)	
	2.5‰	95.2 (0.13)	93.8 (0.11)	94.2 (0.12)	94.5 (0.11)	94.3 (0.13)	
	0	<i>94.5 (0.14)*</i>	<i>94.3 (0.11)*</i>	<i>94.1 (0.14)*</i>	<i>94.9 (0.09)*</i>	<i>94.3 (0.08)*</i>	

Table 4. Comparison of the performance of f_4 -based test of treeness for different types of data simulated under the Figure 1 scenario processing poolfstat analyses. For each MAF threshold (MAF > 1% or MAF > 5%), the table gives True and False (in parenthesis) Positive Rates (in %) for 21 different types of analyses relying on i) allele count data; ii) 15 different Pool-Seq read count data (five mean coverages λ and three sequencing error rates ϵ); and iii) Pool-Seq read count data simulated with $\epsilon = 0$ treated as allele counts (corresponding results of this bad practice are highlighted in italics and *). Each TPR was computed from the analysis of 250 independent datasets (generated from the data simulated under Figure 1 demographic scenario) as the proportion of f_4 with an associated absolute Z-score < 1.96 (95% significance threshold) among all the eight population quadruplets ((P1,P2;P3,P4); (P1,P2;P3,P5); (P1,P2;P4,P5); (P1,P3;P4,P5); (P1,P6;P4,P5); (P2,P3;P4,P5); (P2,P6;P4,P5); (P3,P6;P4,P5)) with a null expected F_4 ($n=250 \times 8=2,000$ estimates). The FPR was similarly computed as the proportion of f_4 with an associated absolute Z-score < 1.96 among all the 37 remaining population quadruplets ($n=250 \times 37=9,250$ estimates).

3.1.4 Precision of the F_4 -ratio based estimation of the admixture rate α

Given the simulated scenario, two different ratios of f_4 estimates could be used to estimate the admixture proportion $\alpha = 0.25$ (Figure 1), namely $\hat{\alpha}_1 = \frac{f_4(P1,P4;P3,P6)}{f_4(P1,P4;P2,P3)}$ and $\hat{\alpha}_2 = \frac{f_4(P1,P5;P3,P6)}{f_4(P1,P5;P2,P3)}$ (Patterson *et al.*, 2012). The `graph.params2symbolic.fstats` function (Table 2) may also prove useful to identify appropriate quadruplets (Supplementary Vignette V2). We used the `compute.f4ratio` function to obtain these two estimates from all the simulated datasets together with their 95% CI (defined as $\hat{\alpha} \pm 1.96\hat{\sigma}_\alpha$ where $\hat{\sigma}_\alpha$ is the block-jackknife standard-error estimate). Tables 5 and S4 provide the mean of the estimated $\hat{\alpha}_1$ and $\hat{\alpha}_2$ respectively over the 250 analyzed datasets for each of the 42 investigated configurations. As expected from the above evaluation of treeness tests, estimates of α were highly consistent among all the investigated configurations and similar for the two considered f_4 -ratio with a mean value varying between 0.245 and 0.248. Yet, a slight downward bias (always < 2%) could be noticed but the estimated 95% CIs were almost always optimal (or close to) since they contained the true simulated value ($\alpha = 0.25$) from 90.0% to 95.2% of the time (Tables 5 and S4).

3.1.5 Evaluation of graph fitting

We further estimated for all the simulated datasets branch lengths in drift units and admixture proportion α with their 95% CIs by fitting the simulated graph with `fit.graph`. As for the f_4 -ratio based estimation, estimates of

MAF threshold	seq. error ϵ	Pool-Seq (read counts) data					allele count data
		$\lambda = 30$	$\lambda = 50$	$\lambda = 75$	$\lambda = 100$	$\lambda = 200$	
>1%	0	0.247 (92.4)	0.247 (92.8)	0.247 (94.4)	0.247 (93.6)	0.247 (92.0)	0.247 (92.8)
	1‰	0.248 (91.6)	0.247 (91.6)	0.247 (92.4)	0.247 (93.2)	0.247 (92.0)	
	2.5‰	0.247 (91.6)	0.246 (94.0)	0.248 (93.2)	0.247 (91.2)	0.248 (92.0)	
	0	<i>0.247 (92.4)</i>	<i>0.247 (92.8)*</i>	<i>0.247 (94.4)*</i>	<i>0.247 (93.6)*</i>	<i>0.247 (92.0)*</i>	
>5%	0	0.247 (92.4)	0.248 (93.2)	0.247 (93.6)	0.247 (93.2)	0.247 (92.4)	0.247 (92.4)
	1‰	0.248 (93.2)	0.247 (91.6)	0.247 (91.6)	0.247 (92.8)	0.247 (91.2)	
	2.5‰	0.247 (91.6)	0.246 (95.2)	0.248 (93.2)	0.247 (90.0)	0.248 (93.2)	
	0	<i>0.247 (92.4)*</i>	<i>0.248 (93.2)*</i>	<i>0.247 (93.6)*</i>	<i>0.247 (93.2)*</i>	<i>0.247 (92.4)*</i>	

Table 5. Comparison of F_4 -ratio based estimation of the simulated admixture proportion α in Figure 1 scenario for different types of data processing poolfstat analyses. For each MAF threshold (MAF > 1% or MAF > 5%), the table gives the mean of the estimated $\hat{\alpha} = \frac{f_3(P1,P4,P3,P6)}{f_4(P1,P4,P2,P3)}$ (across 250 independent simulated datasets) for 21 different types of analyses relying on i) allele count data; ii) 15 different Pool-Seq read count data (five mean coverages λ and three sequencing error rates ϵ); and iii) Pool-Seq read count data simulated with $\epsilon = 0$ treated as allele counts (corresponding results of this bad practice are highlighted in italics and *). The proportion (in %) of the 250 estimated 95% CIs that contain the true simulated value ($\alpha = 0.25$) is given in parenthesis.

539 α were virtually unbiased and consistent across all the 42 different investigated configurations (Figure S3). Nev-
 540 ertheless, the 95% CIs were always too narrow since they contained the actual value ($\alpha = 0.25$) from only 40.8%
 541 to 74.4% of the time (Table S5) as expected from the χ^2 approximation of the LRT underlying the computation of
 542 these CIs. Figures 4 and S4 plot the distributions of the estimated lengths for the ten branches of the simulated
 543 admixture graph branches (over the 250 estimates per configuration) when applying 5% and 1% MAF threshold
 544 respectively. The corresponding mean estimates and proportions of 95% CI's including the true value are provided
 545 in Tables S6 to S15. Note that the branches $P8 \leftrightarrow R$ and $P9 \leftrightarrow R$ that are connected to the root R (Figure 1) can
 546 only be estimated jointly (as $\tau_{P8 \leftrightarrow P9} = \tau_{P8 \leftrightarrow R} + \tau_{P9 \leftrightarrow R}$, R being arbitrarily set in its middle).

547 At the 5% MAF threshold, very similar performance were obtained for the allele count and the different Pool-
 548 Seq datasets whatever the simulated read coverage or sequencing error rates (Figures 4A, 4B, and 4C). Hence,
 549 mostly unbiased branch lengths were estimated for the four leaves (terminal branches) $\tau_{P1 \leftrightarrow P7}$, $\tau_{P2 \leftrightarrow S1}$, $\tau_{P6 \leftrightarrow S}$
 550 and $\tau_{P3 \leftrightarrow S2}$. As previously observed with α , the estimated 95% CI's remained too narrow particularly for $\tau_{P2 \leftrightarrow S1}$
 551 for which less than 50% of the CI's contained the true value (Table S7) compared to more than 80% for $\tau_{P1 \leftrightarrow P7}$
 552 (Table S6). As expected from the drift-scaling approximation, the estimated branch lengths tended to be
 553 slightly downwardly biased (ca. 2%) for the two other leaves ($\tau_{P4 \leftrightarrow P9}$ and $\tau_{P5 \leftrightarrow P9}$) but the estimated 95% CI
 554 displayed similar characteristics since from 48.0% to 87.6% contained the true values, the proportion increasing in
 555 Pool-Seq datasets when coverage and sequencing error decreased (Tables S10 and S11). Conversely, the internal
 556 branch lengths tended to be upwardly biased from a slight (ca. 2%) for $\tau_{S1 \leftrightarrow P7}$, $\tau_{P7 \leftrightarrow P8}$ and $\tau_{S2 \leftrightarrow P8}$ (Tables S12

557 to S14), to a moderate amount (ca. 20%) for the root including branch $\tau_{P8 \leftrightarrow P9}$, the true value being then always
558 outside the estimated 95% CI's (Table S15). Yet, when analyzing data with a lower MAF threshold of 1%, this
559 bias almost completely vanished (Figure S4 and Table S15).

560 For the other branches, the estimates had similar characteristics (yet with a slightly decreased performance
561 for the $\tau_{P4 \leftrightarrow P9}$ and $\tau_{P5 \leftrightarrow P9}$ leaves) for allele count data or Pool-Seq data simulated without sequencing error (Fig-
562 ure S4A). In agreement with previous observations, at the 1% MAF threshold, sequencing errors lead to strong
563 downward bias at the lowest simulated coverages, i.e., when the percentage of false SNPs became non-negligible
564 (Figures S4B and S4C). Finally, whatever the chosen MAF threshold, improperly analyzing read counts as allele
565 count data always lead to a substantial upward bias of the lengths of all the leaves (Figure 4D). Notice however,
566 that this had no or limited impact on the estimation of internal branch lengths.

567 3.1.6 Evaluation of graph construction

568 To provide insights into the reliability of graph construction, we evaluated the performance of the `add.leaf`
569 function in positioning the admixed population P6 on the underlying $((P1,P2),P3),(P4,P5)$ tree (Figure 1) for
570 the different types of simulated datasets. Table S16 gives the proportion of correctly inferred admixture graphs
571 (i.e., corresponding to the simulated scenario) with a $\Delta_{BIC} > 6$ support with all other tested graphs over the
572 250 analyzed datasets for each of the 42 investigated configurations. As the reference tree with rooted topology
573 $((P1,P2),P3),(P4,P5)$ consists of eight branches, P6 may be connected with either i) nine non-admixed edges
574 (connection to either one of the eight branches or as an outgroup) or; ii) $\binom{8}{2} - 1 = 27$ admixed edges from two-way
575 admixture events. Except for the $PS50_{m>1\%}^{\epsilon=2.5\%}$ dataset (the one with the highest percentage of false SNPs), the
576 correct admixture graph was always retrieved with a fairly high support ($\Delta_{BIC} > 15$).

577 3.2 Analysis of real *Drosophila suzukii* Pool-Seq data

578 We here sketched the main findings from the analyses using `poolfst` of a subset of the Pool-Seq data previously
579 generated by Olazcuaga *et al* (2020) focusing on 14 population samples of the invasive species *D. suzukii*. For more
580 details, we encourage readers to consult the Supplementary Vignette V2.

581 3.2.1 Structuring of genetic diversity across the 14 populations

582 Overall, the estimated global F_{ST} across the 14 populations was 7.03% (95% CI; [6.90%; 7.32%]). Estimates of
583 all the pairwise-population F_{ST} confirmed that populations tended to cluster according to their geographic area of

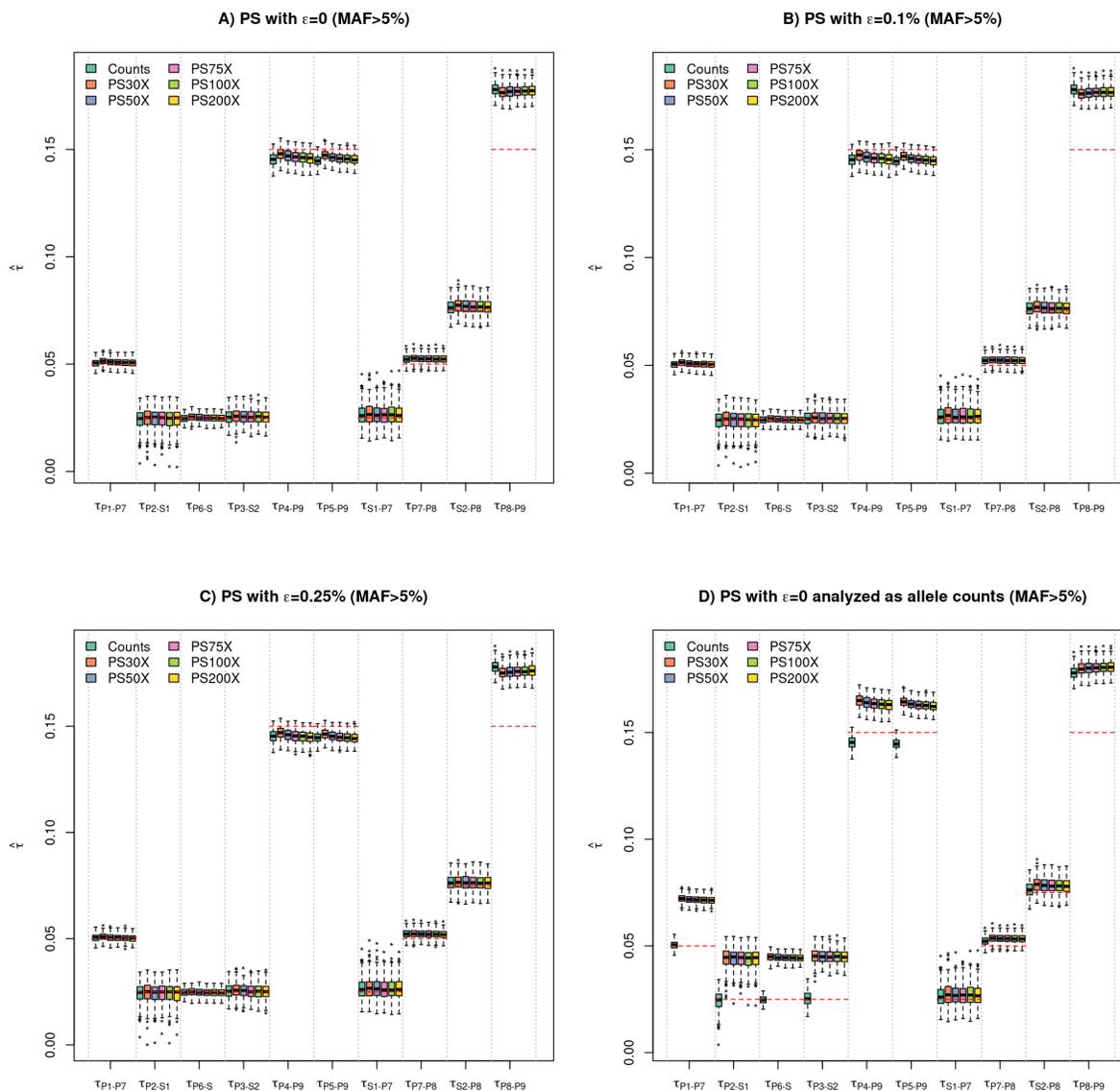


Figure 4. Distribution of the estimated drift-scaled lengths for all the branches in Figure 1 simulated scenario using admixture graph fitting (as implemented in the `fit.graph` function of `poolfstat`) for different types of data with a 5% threshold on the overall SNP MAF. Each box plot summarize the distribution of the 250 estimated lengths of each of the ten branches obtained from the analysis of either allele count dataset (“Counts”) or one of the five different simulated Pool-Seq read count datasets (“PS. λ X”) with different mean coverages ($\lambda = 30; 50; 75; 100; \text{and } 200$) as generated from the genotyping data simulated under the scenario depicted in Figure 1. Pool-Seq read count data were generated with no sequencing errors ($\epsilon = 0$) in A) and D) and with a sequencing error rate of $\epsilon = 1\%$ and $\epsilon = 2.5\%$ in panel B) and C) respectively (Table S1). In D), the read count data were analyzed as allele counts which corresponds to a bad practice. Note that the two branches coming from the root are combined since the position of the root is not identifiable by the model (i.e., $\tau_{P8 \leftrightarrow P9} = \tau_{P8 \leftrightarrow R} + \tau_{P9 \leftrightarrow R}$). Note that the box plots obtained from the analysis of count data are replicated in each panel for comparison purposes. For each branch, a red dotted line indicates the underlying simulated value. For Pool-Seq data, the overall MAF was estimated from read counts.

584 origin (i.e., Asia, America and Hawaii; Figure 2A), with some geographically close populations showing low level
 585 of differentiation. For instance, in the American invasive area the US-Nca, US-Col and US-Nca populations all
 586 displayed pairwise F_{ST} significantly lower than 1%. Likewise, in the native area, the three populations CN-Bei,
 587 CN-Nin and CN-Lia originating from North-Western China were all found very closely related (all pairwise F_{ST}
 588 being lower or very close to 1%). Conversely, the Hawaiian sample (US-Haw) was found the most highly differ-
 589 entiated with all the other populations, all pairwise F_{ST} including US-Haw ranging from 11.7% (with US-Sok) to
 590 17.0% (with US-Col) suggesting strong drift in this population as confirmed by its lowest estimated heterozygosity
 591 (Supplementary Vignette V2).

592 3.2.2 f_3 -based tests of admixture suggest pervasive admixture in the invaded area

593 Out of the 14 sampled populations, two (CN-Lia and JP-Tok) from the Asian native area and four (US-Col, US-
 594 Nca, US-Wat and US-Wis) from the continental American invasive areas showed at least one significantly negative
 595 f_3 at the 95% significance threshold (i.e., Z-score < -1.65). Table 6 summarizes for each of these 6 populations
 596 the number of significantly negative f_3 together with the triplet with the lowest Z-score giving insights into the pair
 597 of populations that branch the closest to the two original sources (assuming a two-way admixture event). Except
 598 for CN-Lia, all the detected signals were significant at a far more stringent threshold (e.g., Z-score < -2.33 at 99%
 599 significance threshold). The f_3 and f_3^* statistics gave almost exactly the same results (Supplementary Vignette
 600 V2).

Population	Origin	nb. of signif. tests ($f_3 Z < -1.65$)	triplet with the lowest f_3 Z-score
CN-Lia	Native	1	CN-Lia;CN-Shi,JP-Sap (Z=-1.66)
JP-Tok	Native	11	JP-Tok;CN-Nin,JP-Sap (Z=-7.11)
US-Col	Invasive (AM)	2	US-Col;BR-Pal,US-Wis (Z=-3.31)
US-Nca	Invasive (AM)	6	US-Nca;JP-Sap,US-Col (Z=-3.89)
US-Wat	Invasive (AM)	13	US-Wat;US-Sdi,US-Sok (Z=-23.6)
US-Wis	Invasive (AM)	4	US-Wis;JP-Sap,US-Col (Z=-5.02)

Table 6. Results of the f_3 -based tests of admixture on populations from the *D. sukikii* invasive species. For all the population displaying at least one significant signal of admixture at the 95% significance threshold ($f_3 Z < -1.65$), the table gives the number of significant tests (out of the $C_2^3 = 78$ performed per population) and the triplet displaying the lowest Z-score (i.e., most significant test).

601 In the native area, JP-Tok showed clear evidence of admixture with 11 significant tests that all involved JP-Sap
 602 (from Northern Japan) as a source proxy. The three lowest f_3 values were obtained with three Chinese populations
 603 (CN-Nin, CN-Bei and CN-Shi in increasing order of f_3). Assuming an admixture-graph like history, this suggests

604 that the two populations branching the closest to the two sources of JP-Tok were JP-Sap and CN-Nin. The remain-
605 ing Chinese population, CN-Lia showed some weak evidence for admixture with only one test barely significant
606 at the 95% threshold for the triplet involving CN-Shi and JP-Sap as source proxies (Table 6).

607 Out of the seven invasive populations from continental America, the four populations US-Col, US-Wis, US-Nca
608 and US-Wat showed strong evidence of admixture. Although it has up to now been considered as the closest to the
609 first invading population of Continental America (based on historical records), the Western American US-Wat pop-
610 ulation displayed the strongest signals with 11 (strongly) significant tests. Interestingly, the three signals supported
611 by the lowest (and hence more significant) Z-score all involved pairs of source population proxies originating from
612 the continental American invasive area namely, in order of increasing Z-score (i.e., decreasing evidence), the (US-
613 Sdi,US-Sok); (BR-Pal,US-Sok) and (US-Col,US-Sok) pairs. As the underlying f_3 CI's did not overlap with those
614 of the other triplet configurations, these three pairs of populations may be considered as the closest (among the
615 sampled populations) to the original US-Wat source populations. It is worth noting that the Western American
616 US-Sok population was involved in nine of the 11 significant negative f_3 statistic with US-Wat as a target. The
617 three others populations, US-Col, US-Wis and US-Nca only had a moderate number of significant tests (compared
618 to others). Such tests always involved at least one of the two other populations and overlapping f_3 CI's. This
619 suggests complex patterns of recurrent admixture event among US-Col, US-Wis and US-Nca, a feature consistent
620 with their low level of differentiation and close geographic origins.

621 **3.2.3 Exploring invasion scenarios with admixture graph construction and fitting**

622 To provide further insights into the relationships of the surveyed populations and the probable scenarios of invasion
623 of *D. sukuzii* in the American area, we relied on admixture graph construction. Our purpose was not to build a
624 comprehensive admixture graph for the 14 populations, which may be elusive given the close relationships of
625 some populations and the pervasiveness of recent and presumably recurrent admixture events among the different
626 populations, but rather to identify key regional event that occurred at early time of the invasion history of the
627 species. From our extensive analyses (Supplementary Vignette V2), we were in fine able to build and estimate
628 the parameters of two admixture graphs represented in Figure 2B and C. The first admixture graph described the
629 somewhat complex and so far non-investigated relationships among the populations of the native area (including
630 the early invasive population established in Hawaii since 1980) with a very good fit since the Z-score of the
631 residuals for the worst fitted f -statistics was 1.06 (Figure 2B). In agreement with previous findings (and geographic
632 proximity), the Hawaiian population was found more closely related to the Japanese population JP-Sap than to the

633 other Chinese populations but it experienced a strong differentiation from their common ancestor (named JP in
634 Figure 2B) with an estimated branch length of 0.255 drift units ($\frac{t}{2N_e}$). Yet, it was not possible from our data to
635 definitively conclude that US-Haw originates from a Japanese population since we have no element to claim that
636 the (ancestral) node population JP was located in Japan. To that end additional sampling of Japanese populations
637 would be required. The inferred graph also confirmed above f_3 -based test results of an admixed origins of JP-Tok
638 between a population closely related to JP-Sap (the main contributor) and a second source likely of Chinese origin
639 although the same caution as for JP are needed regarding the geographic origins of this internal node populations.
640 Similarly, CN-Lia was found admixed with a contributing source of Chinese ancestry related to CN-Shi largely
641 predominant (estimated contribution $\widehat{\alpha}_C = 96.0\%$; 95% CI, [95.7;96.3]), and a second (minor) contributing source
642 of presumably Japanese origin (related to JP-Sap). This may explain why the corresponding f_3 -based test was
643 only barely significant (Table 6). Interestingly, the graph topology also allowed estimating the Chinese ancestry of
644 CN-Lia based on F_4 -ratio resulting in consistent but larger 95% CI ($\widehat{\alpha}_C = 95.6$; 95% CI, [94.4;96.8]) as expected
645 from above simulation study. CN-Nin, the remaining population from the native area, could not be positioned with
646 reasonable accuracy onto the admixture graph of Figure 2B, the resulting worst fitted f -statistics associated to the
647 best fitting graph having a Z-score=3.43. However, both its genetic proximity with CN-Lia and the best fitting
648 admixture graph resulting from its positioning onto the scaffold tree including US-Haw, JP-Sap, CN-Bei and CN-
649 Shi suggested a small amount of Japanese introgression (see Supplementary Vignette V2 for more details).

650 The second admixture graph represented in Figure 2C allowed providing insights into the history of intro-
651 duction of *D. sukukii* into the American continent. It related the three continental American population, US-Sok,
652 US-Wat and BR-Pal to a scaffold including the four unadmixed populations US-Haw, JP-Sap, CN-Shi and CN-Bei
653 with a good fit (the worst fitted f -statistics had a Z-score=-1.83). The underlying scenario suggested that continen-
654 tal American populations originated from at least two major and successive admixture events. The first admixture
655 event lead to the internal node population named Am1 and occurred in balanced proportions between two sources,
656 a Japanese one closely related to JP-Sap and a Hawaiian one relatively distantly related to US-Haw (according to
657 the estimated branch lengths). The US-Sok population was the sampled continental American population the clos-
658 est to Am1 and may thus be assumed the most closely related to the first invading population (in agreement with
659 f_3 -based test results). Yet US-Sok remained separated by about 0.0816 drift units from Am1 which may explain
660 why no significantly negative f_3 were found for triplets with US-Sok as a target.

661 The second major admixture events occurred between the internal node population named Am2 and a Chinese
662 population closely related to the common ancestor of CN-Bei and CN-Shi, with CN-Shi contributing slightly

663 more (58.5% against 41.5% for the other Am1 related ancestor). Interestingly, the closest Am2 representatives
664 among the sampled populations were BR-Pal and US-Sdi (also in agreement with f_3 -based tests) suggesting a
665 more Southern geographical origin for Am2. We found that some additional ancestry from a ghost population or
666 recurrent admixture events (e.g., related to Hawaiian populations) may also have contributed to US-Sdi, but this
667 lead to a poor fit (worst fitted f -statistics Z -score=-5.87 for the best fitting graph resulting from the positioning
668 of US-Sdi onto the graph, see Supplementary Vignette V2). Therefore, US-Sdi is not represented in Figure 2C.
669 Although geographically distant, the Brazilian population BR-Pal thus appeared as the best proxy for Am2 thereby
670 suggesting a rapid spread of *D. sukukii* in South America from this population without any subsequent admixture
671 events. Additional (and preferably ancient) sample from South-American populations would help refining this
672 scenario. Finally, according to the inferred graph, US-Wat was found to originate from a recent admixture between
673 a population very closely related to US-Sok (and thus Am1) and a population deriving from Am2 with similar
674 contributions of both.

675 In agreement with f_3 -based admixture tests that suggested complex admixture histories among the closely
676 related US-Col, US-Wis and US-Nca populations, no satisfactory admixture graph could be found when trying to
677 position each of these onto the Figure 2C graph. Nevertheless, their resulting best fitted graphs all suggested a
678 high contribution of the Am2 admixed source, a second contributing source being related to Japanese populations
679 (Supplementary Vignette V2).

680 **4 Discussion**

681 **4.1 A new version of poolfstat for f -statistics estimation and associated inference from** 682 **both Pool-Seq and allele count data**

683 The R package `poolfstat` was originally developed by Hivert *et al* (2018) to implement an unbiased estimator
684 of F_{ST} for PoolSeq data and provide utilities to facilitate manipulation of such data. We here proposed a sub-
685 stantially improved version that implements unbiased estimators of F_2 , F_3 and F_4 parameters together with their
686 scaled versions (i.e., pairwise F_{ST} , F_3^* and D respectively). Although we primarily focused on the analysis of
687 Pool-Seq data, we extended the package to analyze standard allele count (as obtained from individual genotyping
688 or sequencing data) and to implement unbiased estimators equivalent to those available in the `AdmixTools` suite
689 (Patterson *et al*, 2012) allowing us in turn to validate our estimation procedure. Recently, the `admixr` package was

690 developed to interface most of the `AdmixTools` programs with R for the estimation of f -statistics (only from allele
691 count data), with the noticeable exception of the admixture graph fitting program `qpGraph` (Petr *et al*, 2019). We
692 implemented in `poolfstat` our own functions for fitting, building, visualizing and quality assessment of admix-
693 ture graphs based on the estimated f -statistics. The underlying procedures shared strong similarities with those
694 implemented in `qpGraph` (Patterson *et al*, 2012) resulting on the same fitting on some examples (e.g., Supple-
695 mentary Vignette V1) or also `MixMapper` (Lipson *et al*, 2013, 2014) programs. As recognized by the developers
696 themselves, the latter program specifically developed for admixture inference from allele count data was written
697 in C++ and MATLAB making it ‘cumbersome to use’ for users, as ourselves, with no MATLAB license. More-
698 over, to facilitate local exploration of the admixture graphs space, we also implemented in `poolfstat` efficient
699 semi-automated building utilities (`add.leaf` and `graph.builder` functions). It should be noticed that although
700 it does not include functions for the estimation of f -statistics, the `admixturegraph` R package (Leppälä *et al*,
701 2017) also provides several alternative valuable utilities for the fitting (based on a slightly different approach), the
702 manipulation, and the visualization of admixture graphs together with utilities for the plotting of the statistics with
703 their confidence intervals or the symbolic derivation f -statistics (as `poolfstat`). Overall, our effort of developing
704 with `poolfstat` a self-contained, efficient and user-friendly R package capable of performing the entire workflow
705 for f -statistics based demographic inference from both standard allele count and Pool-Seq read count data will
706 hopefully make such a powerful framework accessible to a wider range of researchers and biological models.

707 **4.2 A unified definition of the F parameters in terms of probability of gene identities**

708 To derive our unbiased estimators, we proposed to recapitulate and unify the different definitions of the F and D
709 parameters in terms of probability of gene identity within population (Q_1) or between pairs of populations (Q_2) as
710 summarized in equation 1. This formulation offers a complementary perspective to the original description of these
711 parameters in terms of covariance of allele frequencies (Patterson *et al*, 2012). In practice, a little algebra shows
712 that the unbiased estimators derived from these two alternative formulations are strictly equivalent (i.e., when com-
713 paring eq. 6 for allele count data with Appendix A in Patterson *et al*, 2012). Formally, the Q_1 and Q_2 probabilities
714 can be viewed as expected identity (in state) of genes across independent replicates of the (stochastic) evolutionary
715 process (Rousset, 2007) that may themselves be expressed as a function of other demo-genetic population parame-
716 ters. Hence, the obtained expressions for F_2 , F_3 and F_4 in terms of Q_1 and Q_2 probabilities can be directly related
717 to those by Peter (2016) in terms of coalescent times which allowed him to provide an in-depth exploration of their
718 theoretical properties under a wide range of demographic models other than admixture graphs (see e.g., Figure 7 in

719 Peter, 2016). More precisely, under an infinite-site mutation model with constant per-generation mutation rate μ ,
720 the probability that two genes are identical in state is $Q = \sum_{t=1}^{\infty} C_t(1-\mu)^{2t} = 1 - 2\mu\mathbb{E}[T] + O(\mu^2)$, where C_t is the prob-
721 ability that the two genes coalesced t generations in the past and $\mathbb{E}[T] \equiv \sum_{t=1}^{\infty} tC_t$ is the expected coalescence time
722 of two genes (see Slatkin, 1991; Rousset, 2007, pp.58-59). Using $Q_1^{(1)} = 1 - 2\mu\mathbb{E}[T_{11}]$ and $Q_1^{(2)} = 1 - 2\mu\mathbb{E}[T_{22}]$ as
723 the IIS probabilities within populations 1 and 2 respectively and $Q_2 = 1 - 2\mu\mathbb{E}[T_{12}]$ as the IIS probability between
724 1 and 2 allows recovering equations 16 (after fixing a typo into it), 20c and 24 by Peter (2016) for F_2 , F_3 and
725 F_4 respectively. Likewise, the estimators derived from (unbiased) estimators of Q_1 and Q_2 are equivalent to those
726 expressed in terms of average pairwise differences between and within populations which are natural estimators
727 for $2\mu\mathbb{E}[T]$ terms as proposed by Peter (2016, eq. 17) for F_2 estimator based on allele count data (e.g., noting that
728 $\hat{\pi}_{11} = 1 - \hat{Q}_1^1$, $\hat{\pi}_{22} = 1 - \hat{Q}_1^2$ and $\hat{\pi}_{12} = 1 - \hat{Q}_2$ following his notations). For Pool-Seq data, replacing the latter
729 estimators of nucleotide diversities by the unbiased estimators described in Ferretti *et al* (2013, eqs. 3 and 10)
730 would also result in the same estimator for F_2 (and other parameters) as those defined in our equation 6.

731 In practice, estimators are obtained by averaging over (a high) number of SNPs which amounts assuming that
732 each represent an independent outcome of a common demographic processes that shaped the genome-wide patterns
733 of genetic diversity. This generally allows to provide accurate estimations and LD between markers (i.e., violation
734 of the marker independence assumption) can be accounted for with block-jackknife estimation of standard errors
735 (Patterson *et al*, 2012). Importantly, as originally noticed by Patterson *et al* (2012), expressions of F_2 , F_3 and F_4
736 in terms of coalescent times (Peter, 2016) show explicitly that they both depend on the demography (via $\mathbb{E}[T]$)
737 and the marker mutation rate (μ). In the scaled versions of F_2 and F_3 (F_{ST} and F_3^* respectively), the parameter
738 μ cancels out making them presumably more comparable across different datasets. It should however be noticed,
739 that for demographic inference purposes, scaling of the f -statistics is not needed. Indeed, the three-population test
740 of admixture is informed by the sign of f_3 which is not affected by the denominator of F_3^* . Similarly, the four-
741 population test evaluates departure of f_4 (i.e., the numerator of D) from a null value expected under the hypothesis
742 of treeness. Patterson *et al* (2012) also showed both analytically and using simulations that F_3 and F_4 estimates
743 remained mostly robust to various realistic SNP ascertainment scheme. It is finally worth stressing that admixture
744 graph inference only requires additivity of F_2 (Patterson *et al*, 2012), a feature not fulfilled by F_{ST} (or F_3^* and D).

745 4.3 Estimation of f -statistics and inference from Pool-Seq data

746 Our simulation study showed that accurate estimates of F and D parameters could be obtained from Pool-Seq data
747 from the unbiased estimators we developed, thereby extending our findings for the Pool-Seq F_{ST} estimator (Hivert

748 *et al*, 2018). With no sequencing error, this remained true even at a read coverage as low as 30X which was here
749 lower than our simulated haploid sample size of 50. Increasing the coverage only provided marginal gain. When
750 introducing sequencing errors, the performance of the estimators tended to decrease for the lowest investigated read
751 coverages (up to 50X) and MAF filtering threshold. This was however essentially due to the presence of spurious
752 SNPs that were not completely filtered out when considering too loose criteria. As a result, simply increasing the
753 threshold on the overall MAF (computed from read counts over all the pool samples) to 5% allowed to remove all
754 the spurious SNPs and recover accurate estimates of the parameters at the lowest read coverages. In agreement
755 with original observations made for allele count data (Patterson *et al*, 2012), all the f -statistics based analyses (i.e.,
756 three-population test of admixture, four-population test of treeness, F_4 -ratio estimation of admixture proportions or
757 admixture graph fitting) remained remarkably robust to a MAF-based ascertainment process. From our simulation
758 study, discarding lowly polymorphic SNPs was only found to increase the bias of the drift-scaled length estimates
759 of internal branch in admixture graph. In practice, cost-effective designs consisting of sequencing pools of 30 to
760 50 individuals at a 50-100X coverage and applying MAF threshold of 5% to filter the called SNPs are expected to
761 provide good performance for all the different f -statistics based inference methods we presented here.

762 For Pool-Seq data, all the above conclusions were nevertheless only valid for the analyses based on the unbiased
763 estimator that accounts for the additional level of variation introduced by the sampling of the DNA of pooled
764 individuals (non identifiable) at the sequencing step. We found that improperly analyzing Pool-Seq read counts as
765 standard allele counts had high detrimental consequence on the estimation of all the F parameters that involved
766 Q_1 probabilities (within population probability of identity) in their definition, i.e., F_2 , F_{ST} (as previously observed
767 by Hivert *et al*, 2018, see also Figure S5), F_3 and F_3^* leading to a complete loss of power of the associated three-
768 population test in our simulation. When processing admixture graph fitting, this also resulted in a strong upward
769 bias in the estimation of branch lengths, including the external ones that were accurately estimated when relying on
770 unbiased estimators. Loosely speaking, not accounting for the extra-variance introduced by the sampling of reads
771 has the same effect of adding a (substantial) amount of extra drift explaining the two aforementioned observations.
772 Although not investigated here (and of little interest since we should definitely rely on unbiased estimators), the
773 amount of extra variance may be inversely proportional to the pool haploid sample size (i.e., bias may decrease
774 with increasing pool sample size). Conversely, analyzing Pool-Seq read counts as standard allele counts did not
775 affect the performance of the f_4 - (and D) based test of treeness or the estimation of admixture proportion from F_4 -
776 ratio. This was expected from the properties of the underlying parameters that only depends on the Q_2 probabilities
777 across the different pairs of population involved in the quadruplet of interest (eq. 3) resulting in the same estimators

778 (see eqns. 4 and 5) for both allele count and Pool-Seq data. More generally, analyzing Pool-Seq read count data
779 with popular programs that were developed for standard allele count data such as those from the AdmixTools
780 (Patterson *et al*, 2012) or TreeMix (Pickrell & Pritchard, 2012) suites should definitely be avoided and, if not,
781 results should be carefully interpreted.

782 **4.4 Insights into the history of the invasive species *D. sukukii* from Pool-Seq data analysis**

783 To illustrate both the power and limitations of f -statistics based methods for historical and demographic inference
784 as implemented in `poolfst`, we analyzed Pool-Seq data available for 14 populations of the invasive species
785 *D. sukukii* (Olazcuaga *et al*, 2020). These population samples were representative of both the Asian native area
786 and the recently invaded American area. Most of them consisted of individuals originating from the same sites as
787 those genotyped in Frainout *et al* (2017) at microsatellite markers and analyzed under an ABC-RF framework.
788 The results remained consistent between the two studies, both of them pointing to complex invasion pathways
789 including multiple introductions leading to admixed origins of the continental American populations. The main
790 source contributions were from Hawaii, where *D. sukukii* was described about 30 years earlier and the native
791 area (China and Japan). However, some inferred scenarios appeared somewhat conflicting. First, for the Hawaiian
792 population that played a key role in American invasion route, both `poolfst` and Frainout *et al* (2017) suggested
793 a Japanese origin. However, we here found that the sample the closest to the source (internal node population JP in
794 Figure 2) was JP-Sap (sampled in Sapporo) while Frainout *et al* (2017) rather concluded it was JP-Tok (sampled
795 in Tokyo) which was not found to be directly contributing to US-Haw in `poolfst` analyses and was even found
796 to be admixed by native populations from Japan and China. In the ABC-RF treatments by Frainout *et al* (2017),
797 all populations from the native area were assumed to be non-admixed and no “ghost” (i.e., unsampled) populations
798 were included in the model whereas such populations are present in admixture graphs through internal nodes.
799 Moreover, the samples from Hawaii and Tokyo both differed in their exact location and collection date (2013 and
800 2016 for Hawaii, 2014 and 2016 for Tokyo) between the two studies, which may further explain the observed
801 discrepancies and more generally promotes the sequencing of additional samples in this area to better resolve the
802 origin of the Hawaiian population(s).

803 Interestingly, `poolfst` results challenged the initial view about the pioneering origin of the Californian pop-
804 ulation US-Wat in the invasion of continental America (as suggested by historical records) suggesting it rather
805 originates from an admixture between two already established but unsampled continental American populations,
806 one presumably Northern (related to AmI and here represented by US-Sok) and the other presumably more South-

ern (related to Am2 and here represented by BR-Pal from Brazil and US-Sdi from South-California). This discrepancy between Fraimout *et al* (2017) and `poolfstat` results points to three key issues. First, a too strong reliance on the reported date of first observation of the species in the invasive area when formalizing the scenario to be compared in ABC modeling may actually mislead inference and this especially since *D. sukukii* was first observed at very close dates in the US-Wat, US-Sdi and US-Sok sampled locations (i.e., 2008, 2009 and 2009, respectively). As a matter of fact, Fraimout *et al* (2017) only considered scenarios in which US-Wat was the first population introduced in continental America. Second, in ABC, scenarios are defined by hand justifying the use of dates of first observation to minimize their number (Estoup & Guillemaud, 2010). The functions implemented in `poolfstat` circumvent this constraint by facilitating a quick and automatic exploration of the admixture graph space to identify key historical events relating the populations of interest. Third, our finding reinforces the concern that the formalization of invasion scenarios including the possibility of unsampled populations is crucial. This possibility is by construction included in admixture graph construction but is also possible in ABC modeling (e.g. Guillemaud *et al*, 2010). Similarly, Fraimout *et al* (2017) argued for an admixed origin of the Brazilian BR-Pal population (first observed in 2013) between undefined North-Western and North-Eastern American sources, while we here found that this population was the best proxy for the ancestral “ghost” American population Am2 (Figure 2C) which may be viewed as one of the main contributor of all the sampled North-American populations (but US-Sok). Again, this results underline advantages of not relying on historical dates as for `poolfstat` analyses, and promotes the sequencing of additional samples in South and North-Western America areas to more thoroughly decipher the invasion routes followed by continental American populations.

If Pool-Seq data analyzed with `poolfstat` allowed to refine historical and demographic scenarios in both the native and invasive areas, the *D. sukukii* Pool-Seq data analysis also illustrated some inherent constraints imposed to the modeled demographic history when fitting admixture graph. In particular, more complex histories involving recurrent admixture events turned out to be difficult or even impossible to fit unless a number of source key samples are included, as observed here for the North-Eastern American populations. In real-life applications involving a large number of invasive populations characterized by numerous and recurrent introduction events, summarizing precisely and with a good fit the history of all surveyed populations with a comprehensive admixture graph may remain elusive. However, as previously underlined (Patterson *et al*, 2012; Lipson & Reich, 2017; Lipson, 2020), in addition to providing robust formal tests of admixture or treeness, a decisive advantage of f -based inference methods is to allow straightforward assessment of the fitted admixture graph by carefully inspecting and reporting Z-score of the residuals of the fitted statistics, an option not available in other related methods such as `TreeMix`

837 (Pickrell & Pritchard, 2012). Beyond modeling the history of populations as admixture graphs (via formal tests of
838 admixture of treeness or graph fitting), Peter (2016) provided valuable theoretical insights to interpret the estimated
839 f -statistics under alternative demographic models such as island, stepping-stone or serial founder models. This
840 suggests in turn that these statistics should be informative to estimate the parameters of demographic scenarios
841 more complex than admixture graphs (e.g., under an ABC framework as in Collin *et al*, 2021).

842 **Acknowledgements**

843 This work was supported by the French National Research Agency (ANR) for the projects SWING (ANR-16-
844 CE02-0015-01) and GANDHI (ANR-20-CE02-0018).

845 **Data Accessibility**

846 The vcf file generated for the *D. suzukii* Pool Seq data is publicly available from the Zenodo repository (<http://doi.org/10.5281/zenodo.4709080>). Further details are provided in Supplementary Vignettes V1 and V2.
847

848 **Supporting Information**

849 Additional supporting information is available online and consists of three pdf files: i) vignette V1 for the R
850 package `poolfst` providing a detailed tutorial using Pool-Seq and allele count data simulated under the scenario
851 depicted in Figure 1; ii) vignette V2 detailing the analysis of the real *Drosophila suzukii* Pool-Seq data using the
852 R package `poolfst`; and iii) a single pdf file with the sixteen supplementary Tables and the five supplementary
853 Figures.

854 **References**

- 855 Adrion JR, Kousathanas A, Pascual M, *et al* (2014) *Drosophila suzukii*: The genetic footprint of a recent, world-
856 wide invasion. *Molecular Biology and Evolution*, **31**, 3148–3163.
- 857 Andersen MM, Højsgaard S (2019) Ryacas: A computer algebra system in R. *Journal of Open Source Software*,
858 **4**.
- 859 Bhatia G, Patterson N, Sankararaman S, Price AL (2013) Estimating and interpreting Fst: the impact of rare
860 variants. *Genome Research*, **9**, 1514–21.
- 861 Busing FMTA, Meijer E, Leeden RVD (1999) Delete-m jackknife for unequal m. *Statistics and Computing*, **9**,
862 3–8.
- 863 Collin FD, Durif G, Raynal L, *et al* (2021) Extending approximate bayesian computation with supervised machine
864 learning to infer demographic history from genetic polymorphisms using diyabc random forest. *Molecular*
865 *Ecology Resources*, **accepted**.
- 866 Durand EY, Patterson N, Reich D, Slatkin M (2011) Testing for ancient admixture between closely related popu-
867 lations. *Molecular biology and evolution*, **28**, 2239–2252.
- 868 Eddelbuettel D (2013) *Seamless R and C++ Integration with Rcpp*. Springer, New York.
- 869 Estoup A, Guillemaud T (2010) Reconstructing routes of invasion using genetic data: why, how and so what?
870 *Molecular ecology*, **19**, 4113–4130.
- 871 Feder AF, Petrov DA, Bergland AO (2012) LDx: estimation of linkage disequilibrium from high-throughput pooled
872 resequencing data. *PLoS One*, **7**, e48588.
- 873 Ferretti L, Ramos-Onsins SE, Pérez-Enciso M (2013) Population genomics from pool sequencing. *Molecular*
874 *Ecology*, **22**, 5561–5576.
- 875 Fraimout A, Debat V, Fellous S, *et al* (2017) Deciphering the routes of invasion of *drosophila suzukii* by means of
876 abc random forest. *Molecular biology and evolution*, **34**, 980–996.
- 877 Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing. *arXiv*, p. 1207.3907.

- 878 Gautier M (2015) Genome-wide scan for adaptive divergence and association with population-specific covariates.
879 *Genetics*, **201**, 1555–1579.
- 880 Gautier M, Foucaud J, Gharbi K, *et al* (2013) Estimation of population allele frequencies from next-generation
881 sequencing data: pool-versus individual-based genotyping. *Molecular Ecology*, **22**, 3766–3779.
- 882 Glenn TC (2011) Field guide to next-generation dna sequencers. *Molecular Ecology Resources*, **11**, 759–769.
- 883 Green RE, Krause J, Briggs AW, *et al* (2010) A draft sequence of the neandertal genome. *Science (New York, N.Y.)*,
884 **328**, 710–722.
- 885 Guillemaud T, Beaumont MA, Ciosi M, Cornuet JM, Estoup A (2010) Inferring introduction routes of invasive
886 species using approximate bayesian computation on microsatellite data. *Heredity*, **104**, 88–99.
- 887 Hivert V, Leblois R, Petit EJ, Gautier M, Vitalis R (2018) Measuring genetic differentiation from pool-seq data.
888 *Genetics*, **210**, 315–330.
- 889 Iannone R (2020) *DiagrammeR: Graph/Network Visualization*. R package version 1.0.6.1.
- 890 Kass RE, Raftery AE (1995) Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795.
- 891 Kelleher J, Etheridge AM, McVean G (2016) Efficient coalescent simulation and genealogical analysis for large
892 sample sizes. *PLoS Computational Biology*, **12**, e1004842.
- 893 Knaus BJ, Grünwald NJ (2017) vcfr: a package to manipulate and visualize variant call format data in R. *Molecular
894 Ecology Resources*, **17**, 44–53.
- 895 Koboldt DC, Zhang Q, Larson DE, *et al* (2012) VarScan 2: somatic mutation and copy number alteration discovery
896 in cancer by exome sequencing. *Genome Research*, **22**, 568–576.
- 897 Kofler R, Orozco-terWengel P, De Maio N, *et al* (2011) Popoolation: a toolbox for population genetic analysis of
898 next generation sequencing data from pooled individuals. *PloS One*, **6**, e15925.
- 899 Kunsch HR (1989) The jackknife and the bootstrap for general stationary observations. *Annals of Statistics*, **17**,
900 1217–1241.
- 901 Lawson CL, Hanson RJ (1995) *Solving least squares problems*. No. 15 in Classics in applied mathematics. Society
902 for industrial and applied mathematics, 1995, 1st edn..

- 903 Leblois R, Gautier M, Rohfritsch A, *et al* (2018) Deciphering the demographic history of allochronic differentiation
904 in the pine processionary moth *thaumetopoea pityocampa*. *Molecular Ecology*, **27**, 264–278.
- 905 Leppälä K, Nielsen SV, Mailund T (2017) admixturegraph: an r package for admixture graph manipulation and
906 fitting. *Bioinformatics*, **33**, 1738–1740.
- 907 Li H, Handsaker B, Wysoker A, *et al* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*,
908 **25**, 2078–2079.
- 909 Lipson M (2020) Applying f-statistics and admixture graphs: Theory and examples. *Molecular Ecology Resources*,
910 **20**, 1658–1667.
- 911 Lipson M, Loh PR, Levin A, Reich D, Patterson N, Berger B (2013) Efficient moment-based inference of admixture
912 parameters and sources of gene flow. *Molecular Biology and Evolution*, **30**, 1788–1802.
- 913 Lipson M, Loh PR, Patterson N, *et al* (2014) Reconstructing austronesian population history in island southeast
914 asia. *Nature Communications*, **5**, 4689.
- 915 Lipson M, Reich D (2017) A working model of the deep relationships of diverse modern human genetic lineages
916 outside of africa. *Molecular Biology and Evolution*, **34**, 889–902.
- 917 Long Q, Jeffares DC, Zhang Q, *et al* (2011) PoolHap: inferring haplotype frequencies from pooled samples by
918 next generation sequencing. *PLoS One*, **6**, e15292.
- 919 McKenna A, Hanna M, Banks E, *et al* (2010) The genome analysis toolkit: a mapreduce framework for analyzing
920 next-generation dna sequencing data. *Genome Research*, **20**, 1297–1303.
- 921 Nocedal J, Wright SJ (1999) *Numerical optimization*. Springer series in operations research. Springer, New York,
922 NY [u.a.].
- 923 Olazcuaga L, Loiseau A, Parrinello H, *et al* (2020) A whole-genome scan for association with invasion success in
924 the fruit fly *drosophila suzukii* using contrasts of allele frequencies corrected for population structure. *Molecular
925 biology and evolution*, **37**, 2369–2385.
- 926 Paradis E, Claude J, Strimmer K (2004) Ape: Analyses of phylogenetics and evolution in r language. *Bioinformat-
927 ics*, **20**, 289–290.

- 928 Paris M, Boyer R, Jaenichen R, *et al* (2020) Near-chromosome level genome assembly of the fruit pest *Drosophila*
929 *suzukii* using long-read sequencing. *Scientific reports*, **10**, 11227.
- 930 Patterson N, Moorjani P, Luo Y, *et al* (2012) Ancient admixture in human history. *Genetics*, **192**, 1065–1093.
- 931 Peter BM (2016) Admixture, population structure, and f-statistics. *Genetics*, **202**, 1485–1501.
- 932 Petr M, Vernet B, Kelso J (2019) admixr-r package for reproducible analyses using admixtools. *Bioinformatics*,
933 **35**, 3194–3195.
- 934 Pickrell JK, Pritchard JK (2012) Inference of population splits and mixtures from genome-wide allele frequency
935 data. *PLoS Genetics*, **8**, e1002967.
- 936 Reich D, Thangaraj K, Patterson N, Price AL, Singh L (2009) Reconstructing Indian population history. *Nature*,
937 **461**, 489–494.
- 938 Rousset F (2007) Inferences from spatial population genetics. In *Handbook of Statistical Genetics* (edited by
939 DJ Balding, M Bishop, C Cannings), pp. 945–979. John Wiley and Sons, Ltd, Chichester, England, 3rd edn..
- 940 Schlötterer C, Tobler R, Kofler R, Nolte V (2014) Sequencing pools of individuals - mining genome-wide poly-
941 morphism data without big funding. *Nature Reviews Genetics*, **15**, 749–763.
- 942 Slatkin M (1991) Inbreeding coefficients and coalescence times. *Genetical research*, **58**, 167–175.
- 943 Weir BS (1996) *Genetic data analysis II : methods for discrete population genetic data*. Sinauer Associates,
944 Sunderland, Mass.
- 945 Weir BS, Goudet J (2017) A unified characterization of population structure and relatedness. *Genetics*, **206**, 2085–
946 2103.