MOLECULAR ECOLOGY RESOURCES

Supplemental Information for:

Biased assessment of ongoing admixture using STRUCTURE in the absence of reference samples

Sara Ravagni, Ines Sanchez-Donoso & Carles Vilà

Table of Contents:

Supplementary text

Simulation of ancestral populations

A high number of simulations for markers with different number of alleles were carried out in Easypop v. 2.0.1 (Balloux, 2001) in order to obtain pairs of populations with the desired levels of genetic differentiation. The parameters used in the simulations to reach the desired level of differentiation are indicated in Table S1. A total of 100 pairs of populations were simulated with each set of parameters. Although each simulation was different, the F_{ST} values observed in all cases were very close to the intended values and no outliers were detected (Figure S1). Consequently, all simulated datasets were appropriate to represent the targeted F_{ST} values.

Similarly, since the Easypop simulations were run for thousands of generations, there was a risk that the final mean number of alleles per locus could be very much lower than the initial value due to drift. However, we simulated large populations and the average number of alleles remained very close to the initial number in all cases (Figure S2). Only in the case of simulations for F_{ST}=0.2 for markers with a maximum of 10 alleles, the final average number of alleles was clearly lower (8.04). However, this average remains very much higher than 5 alleles and the simulations remain relevant to represent sets of markers with high polymorphism.

We are aware that our simulations may imply a high level of homoplasy due to the combination of a high mutation rate (specially for microsatellite-like markers, for which it was set at 10^{-3}), elevated effective population size, and long simulations. Therefore, we carried out tests with different running parameters to test the potential effect of homoplasy on downstream analysis. On one hand, we simulated populations with smaller effective size (N=400) and higher migration rate (0.001) and we run these simulations during a very high number of generations (50000) to reach drift-migration-mutation equilibrium (as in Evanno, Regnaut, & Goudet, 2005;

Puechmaille, 2016). On the other hand, to eliminate completely the effect of homoplasy we assumed a mutation rate of 0 and a lower number of generations (600). In both cases, for a given value of population differentiation we obtained practically identical results, the same as for the analyses shown in the main text of the manuscript, which shows that homoplasy did not have an effect on the analyses.

REFERENCES

Balloux, F. (2001). EASYPOP (Version 1.7): A Computer Program for Population Genetics Simulations. *Journal of Heredity*, *92*(3), 301–302. https://doi.org/10.1093/jhered/92.3.301

Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Molecular Ecology*. https://doi.org/10.1111/j.1365-294X.2005.02553.x

Puechmaille, S. J. (2016). The program STRUCTURE does not reliably recover the correct population structure when sampling is uneven: Subsampling and new estimators alleviate the problem. *Molecular Ecology Resources*, *16*(3), 608–627. https://doi.org/10.1111/1755-0998.12512

Supplementary table

Table S1. Input parameters used in Easypop to simulate pairs of populations with the desired level of differentiation (F_{ST}) for the different types of markers (with different number of alleles). Mutation model: 1, KAM (each allele has the same probability to mutate to the other allelic state); 2, Mixed model of SSM (single step mutation model) with a 0.3 proportion of KAM mutation events.

F_{ST}	Number of alleles	Number of generations	Individuals per population	Mutation rate	Mutation model
0.005	2	1000	5000	10^{-8}	1
0.1	2	1300	3000	10^{-8}	1
0.2	2	1300	1400	10^{-8}	1
0.005	5	1500	3000	10^{-3}	2
0.1	5	3000	1500	10^{-3}	2
0.2	5	4000	700	10^{-3}	2
0.005	10	1500	3000	10^{-3}	2
0.1	10	3000	1500	10^{-3}	2
0.2	10	4000	700	10^{-3}	2

Supplementary figures

Figure S1. F_{ST} obtained in the population simulations. Input parameters in Table S1 were chosen to target the expected F_{ST} values. Those parameters were used to obtain 100 pairs of populations for each combination of expected F_{ST} and number of alleles. The small variance in all cases indicates that all the simulated populations indeed represented the intended level of differentiation and were suitable for subsequent analyses.

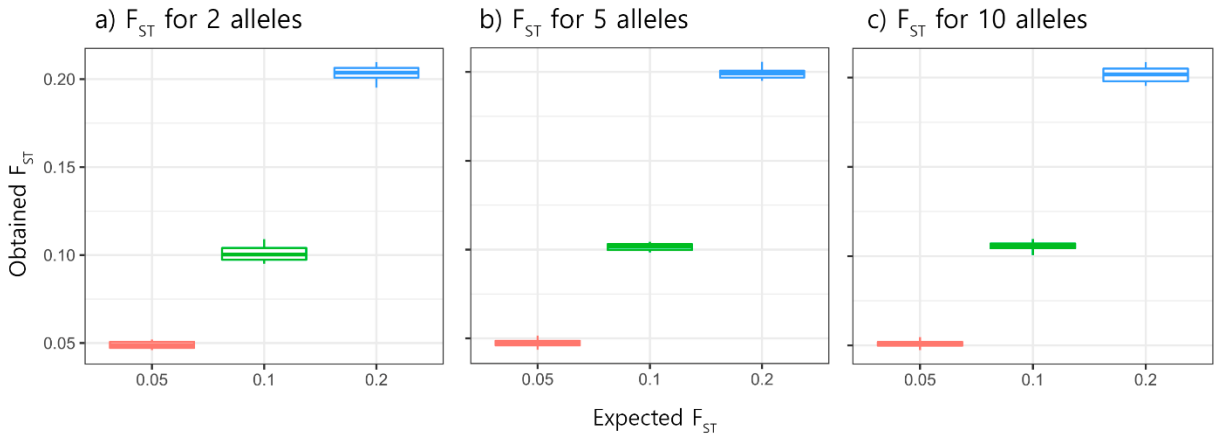

Figure S2. Average number of alleles per locus in the simulated populations. This was evaluated to confirm that the number of alleles in the simulated Easypop populations were close to the maximum number of alleles set in the simulations and had not declined excessively due to drift. Except for markers with 10 alleles and F_{ST}= 0.2, the effective number of alleles remained very close to the initial number set in the Easypop simulations.

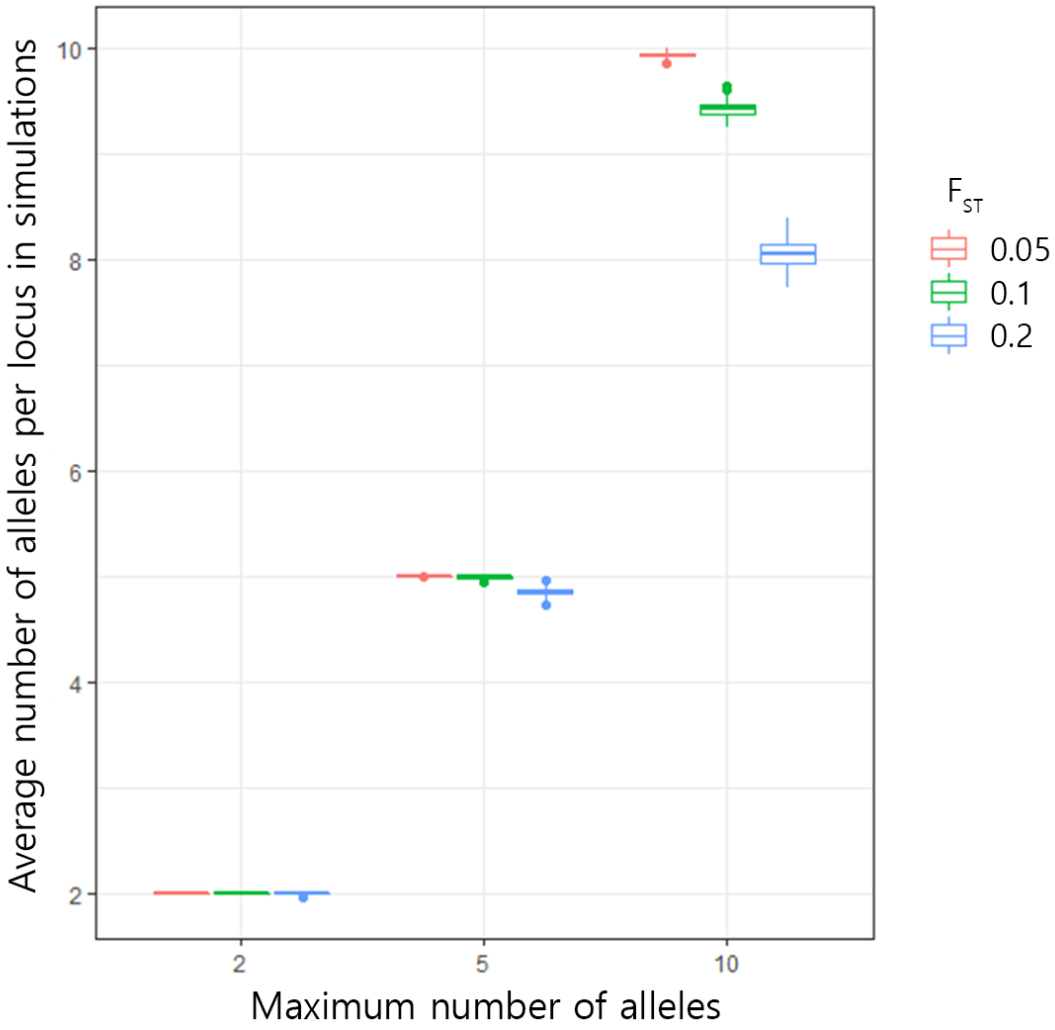

Figure S3. Individual proportion of the genome belonging to population A estimated with STRUCTURE ($q_{STRUCTURE}$) compared to the real value (q_{real}) calculated during the simulations, for F_{ST} = 0.05. (a) Hybridization rate of 1%, (**b)** Hybridization rate of 5%. Compared to Figure 2 (with F_{ST} = 0.1**)**, precision and accuracy in the estimates decrease, showing that a lower genetic differentiation between the ancestral populations affects negatively the estimates of ancestry.

Figure S3a

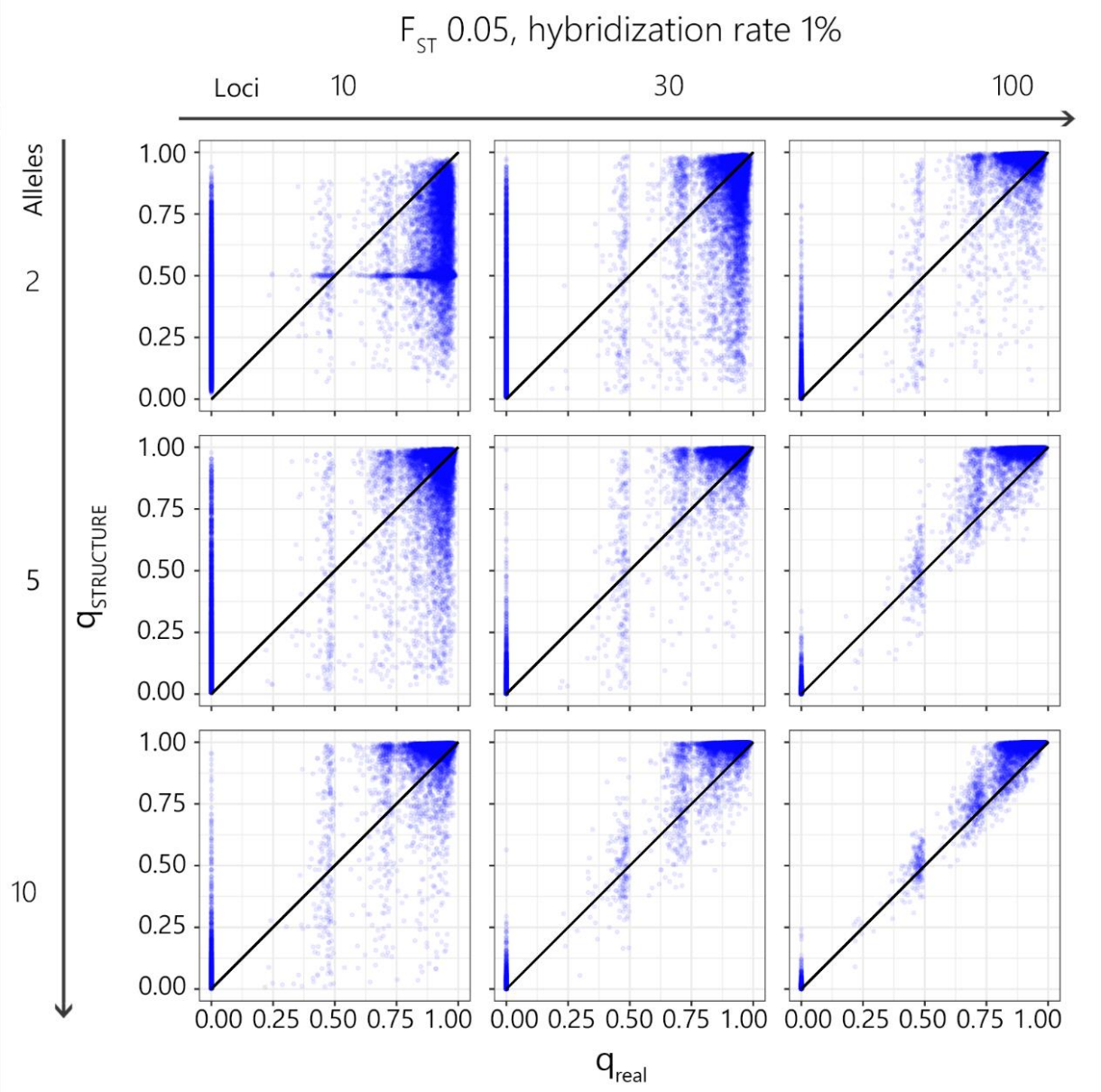

Figure S3b

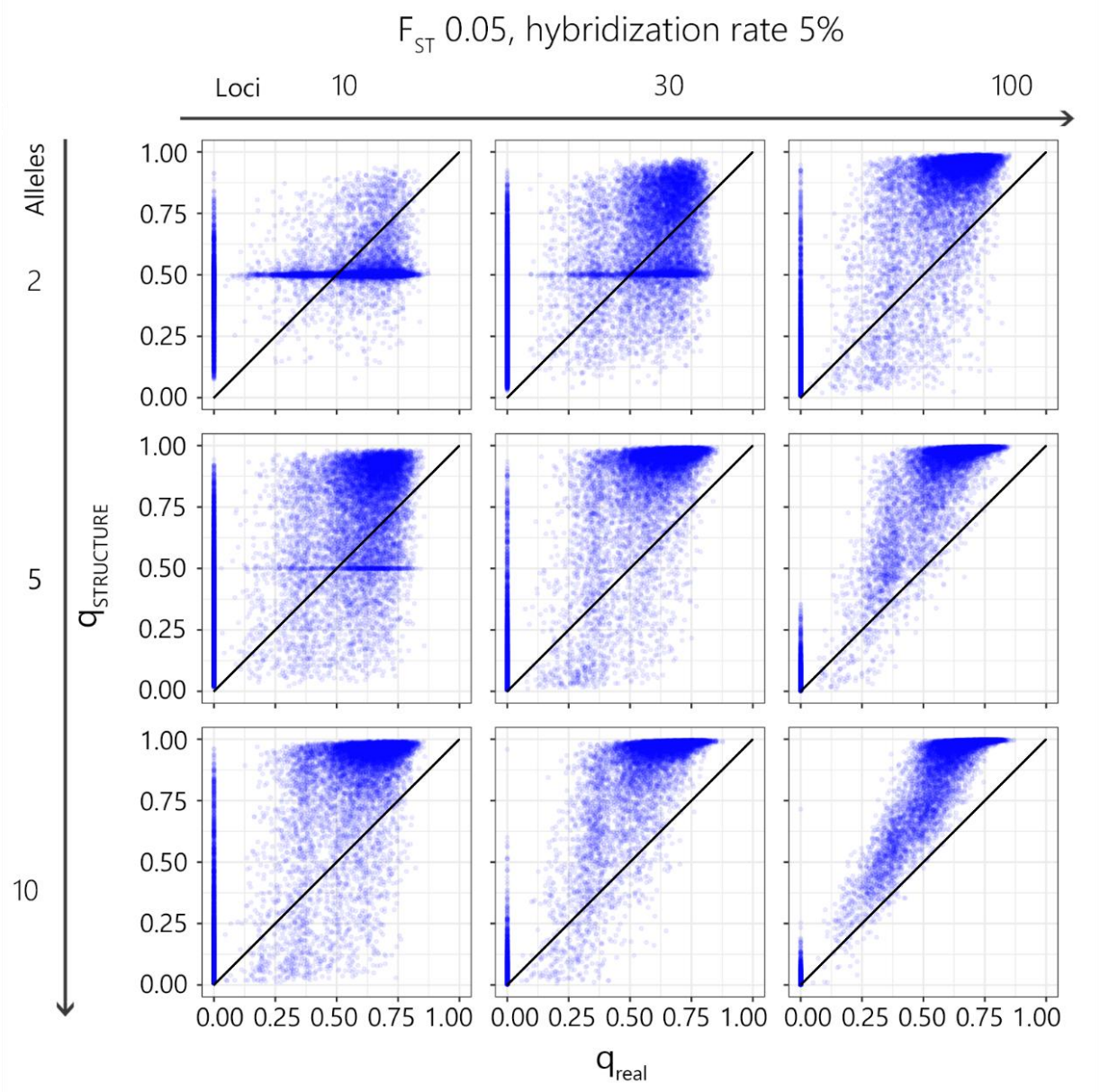
F_ST 0.05, hybridization rate 5%
Loci
10
30
100
Alleles
2
5
10
q_STRUCTURE
q_real
0.00
0.25
0.50
0.75
1.00

Figure S4. Individual proportion of the genome belonging to population A estimated with STRUCTURE ($q_{STRUCTURE}$) compared to the real value (q_{real}) calculated during the simulations, for F_{ST} = 0.2. (a) Hybridization rate of 1%, (**b)** Hybridization rate of 5%. Compared to Figures 2 and S3 (F_{ST} = 0.1 and 0.05, respectively**)**, $q_{STRUCTURE}$ shows an improvement in accuracy and precision with increasing genetic differentiation between the ancestral populations.

Figure S4a

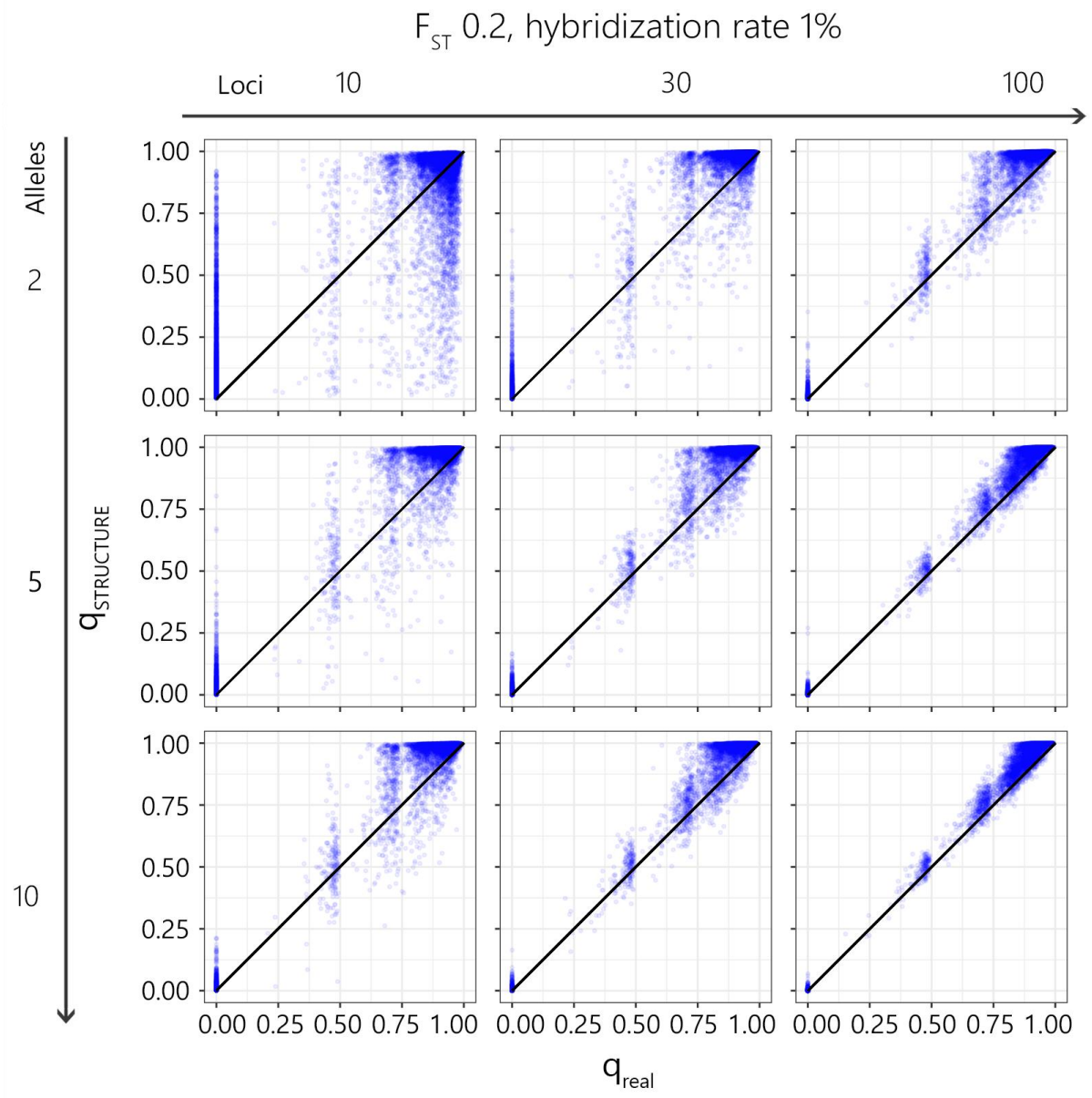

Figure S4b

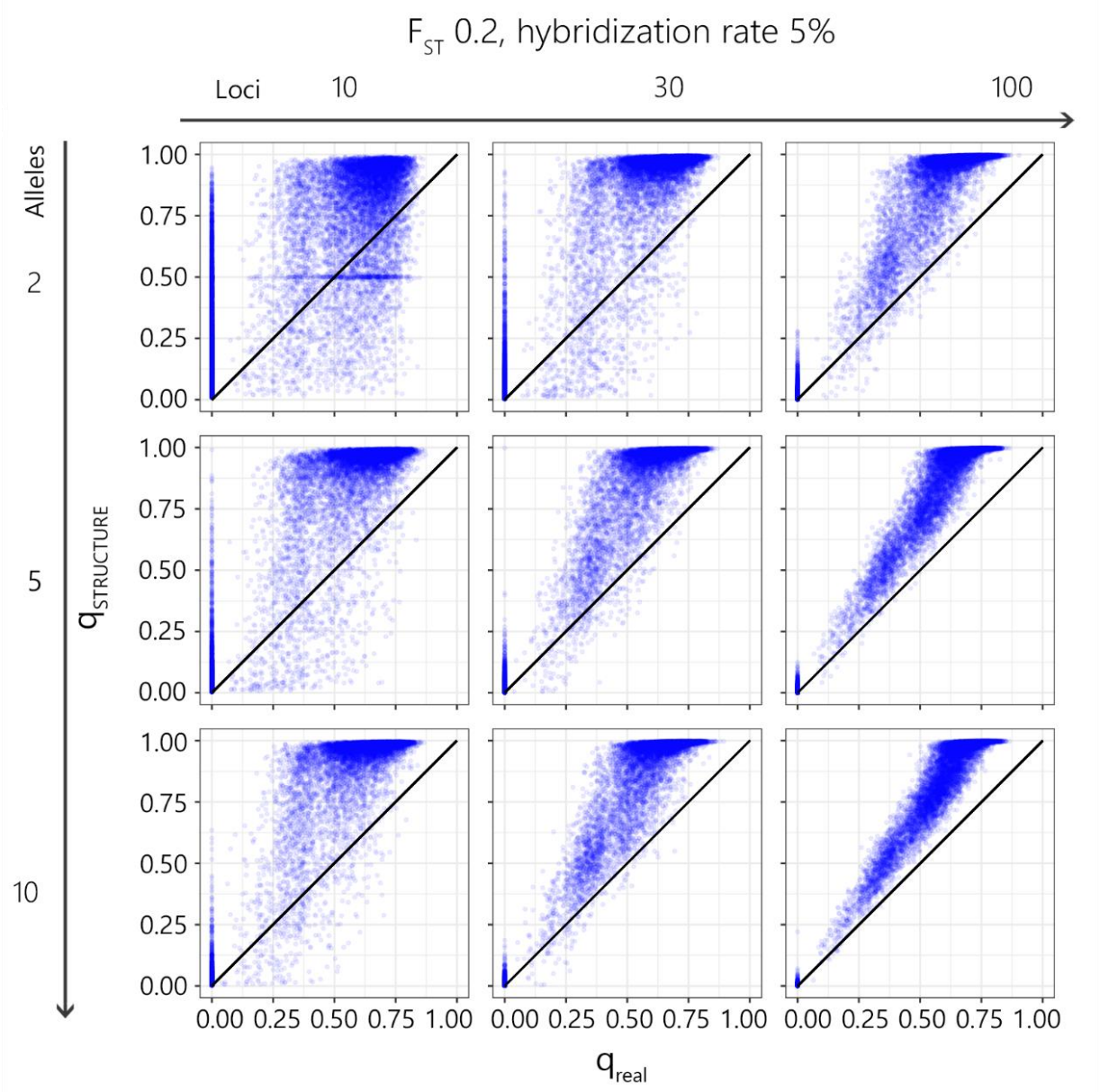
F_ST 0.2, hybridization rate 5%
Loci
10
30
100
Alleles
2
5
10
q_STRUCTURE
q_real
0.00
0.25
0.50
0.75
1.00

Figure S5. Linear regressions for $q_{STRUCTURE}$ in relation to q_{real} for individuals from population A. We only show the cases where regression lines could be adjusted (residuals followed normal distribution). In all cases the intercept was significantly higher than 0, showing that STRUCTURE tended to overestimate the proportion of genome from population A. **a,** intercept; **b,** slope. All coefficients were significantly different from 0 ($p < 10^{-16}$).

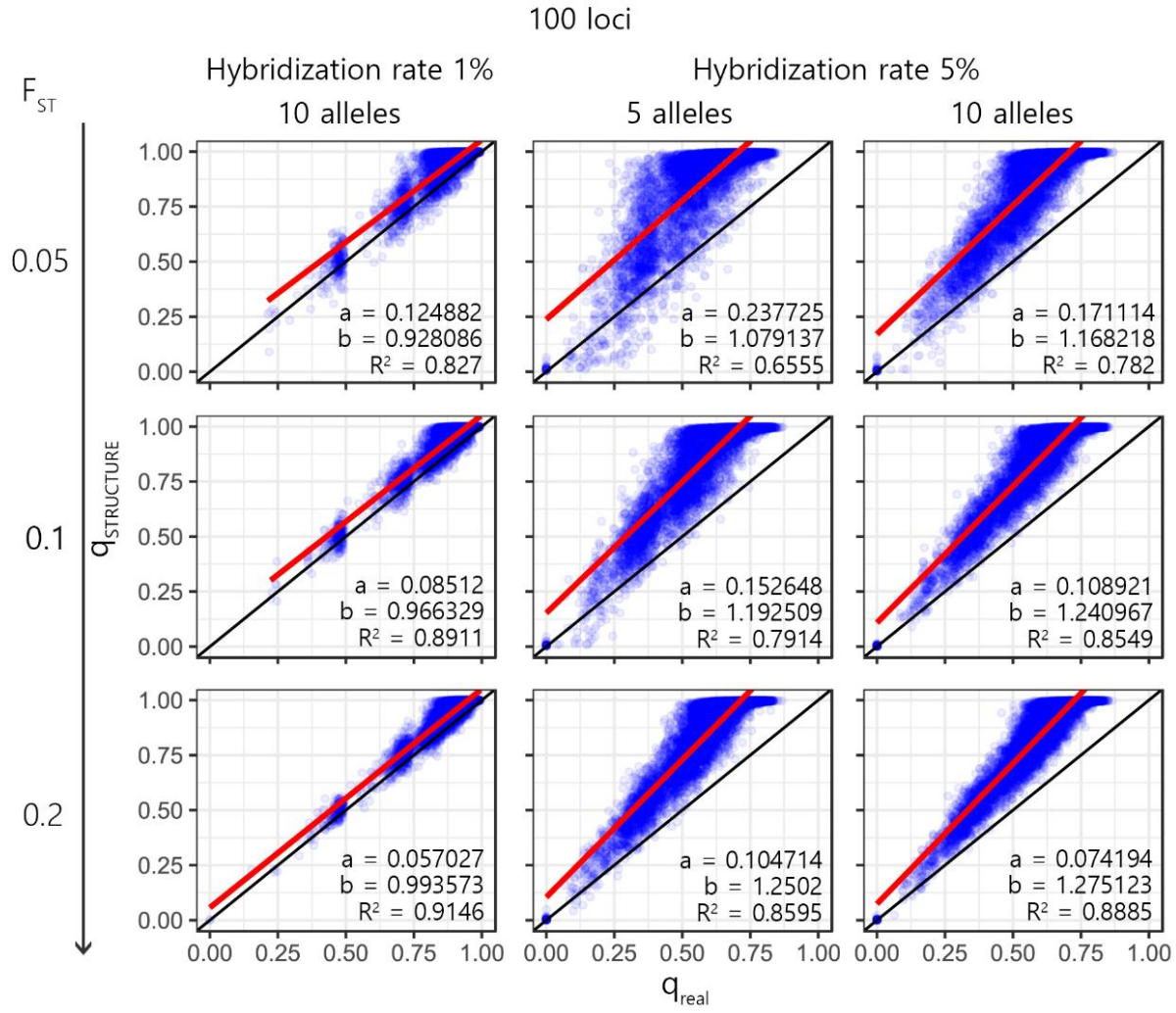

Figure S6. Density plots for $q_{STRUCTURE}$ - q_{real} for individuals from population A resulting from the simulations with F_{ST}= 0.05 between the ancestral populations. (**a)** hybridization rate of 1%; (**b)** 5%.

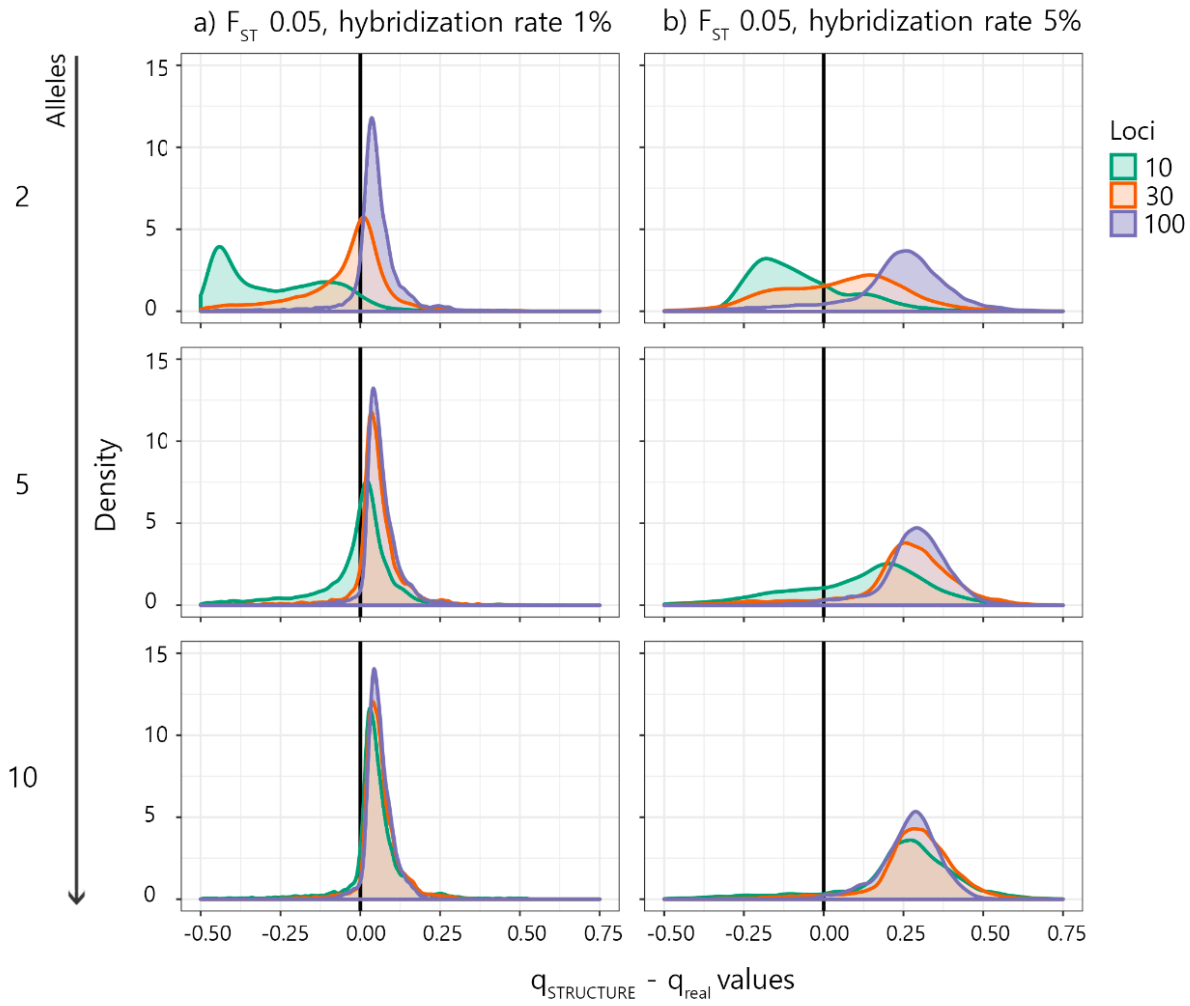

Figure S7. Density plots for $q_{STRUCTURE}$ - q_{real} for individuals from population A resulting from the simulations with F_{ST}= 0.2 between the ancestral populations. **(a)** hybridization rate of 1%; **(b)** 5%.

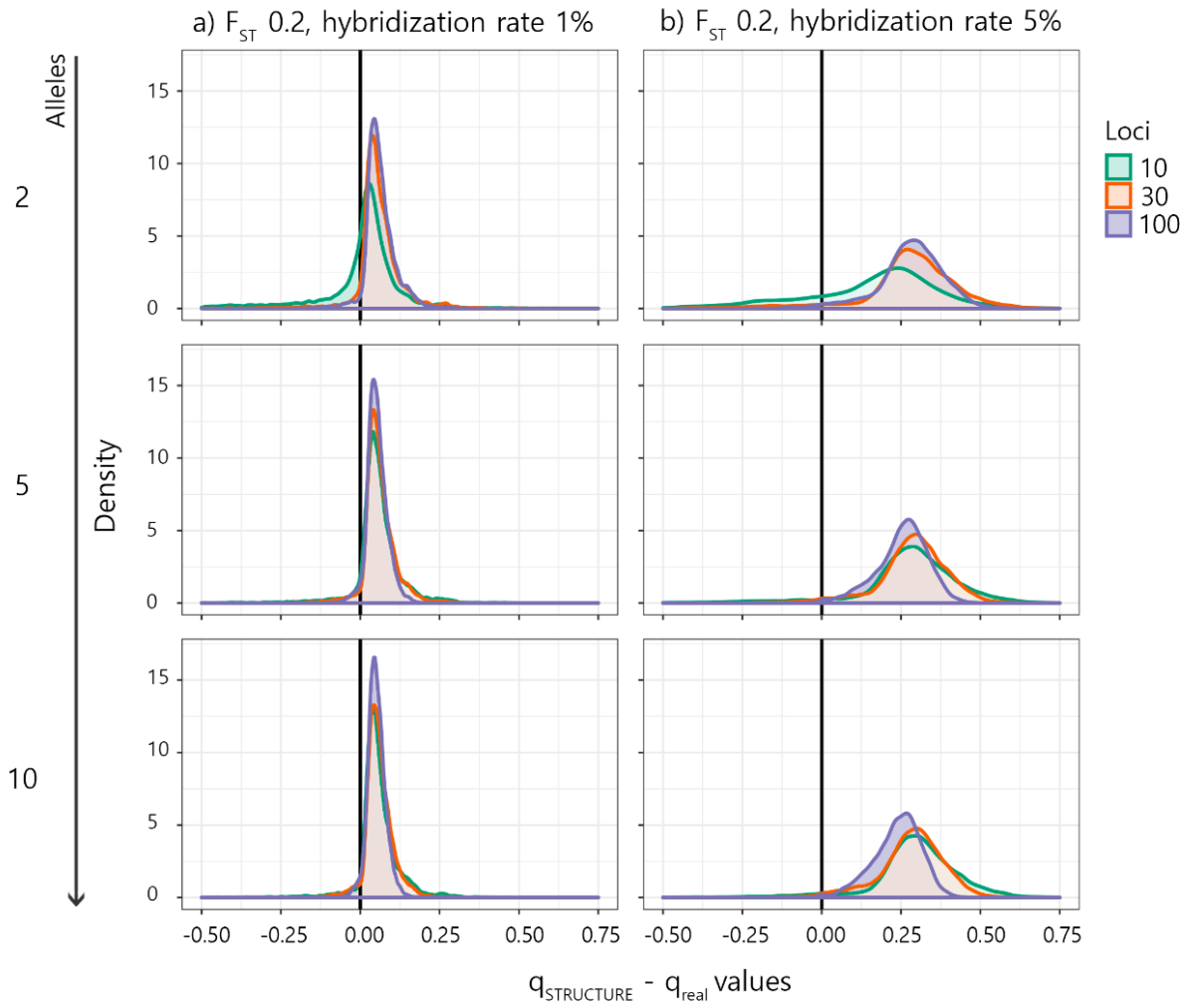

Figure S8. Comparison of $q_{STRUCTURE}$ and q_{real} when 10% of the individuals from the target population are sampled from the ancestral population and are used as reference.
Simulations were carried out with F_{ST} = 0.1 and 30 loci, varying the number of alleles per marker and hybridization rate, and using or not the options POPINFO and POPALPHAS. The combined use of the two options (last column) increased the accuracy of the estimates of *q*.

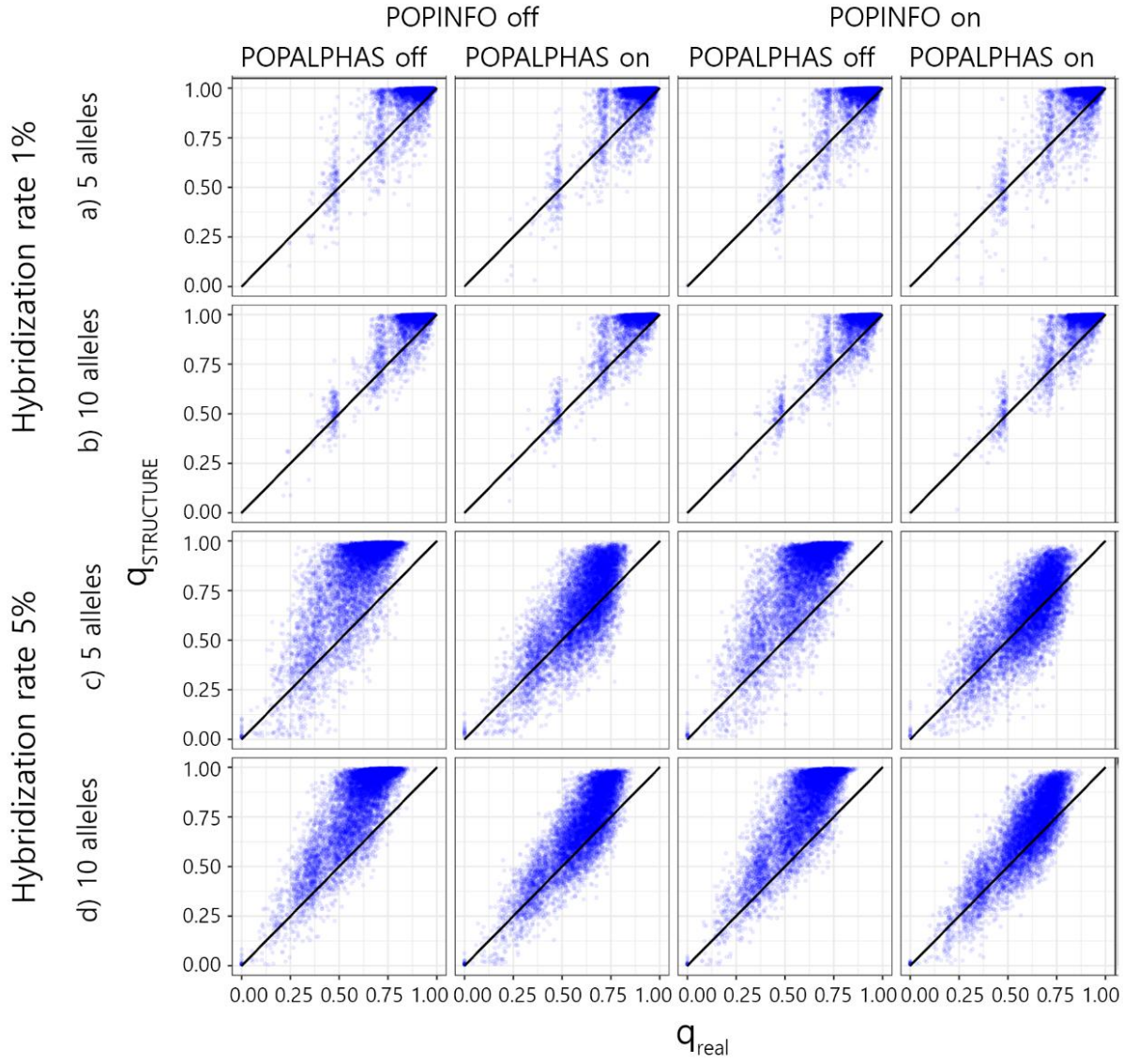

Figure S9. Comparison of *q* values estimated for a dataset of common quails without and with the inclusion of non-admixed reference individuals. Although some outlier points appear below the diagonal, the majority of them lay above the line, showing that *q* values estimated in the absence of non-admixed individuals are generally larger than *q* values obtained when including reference samples.

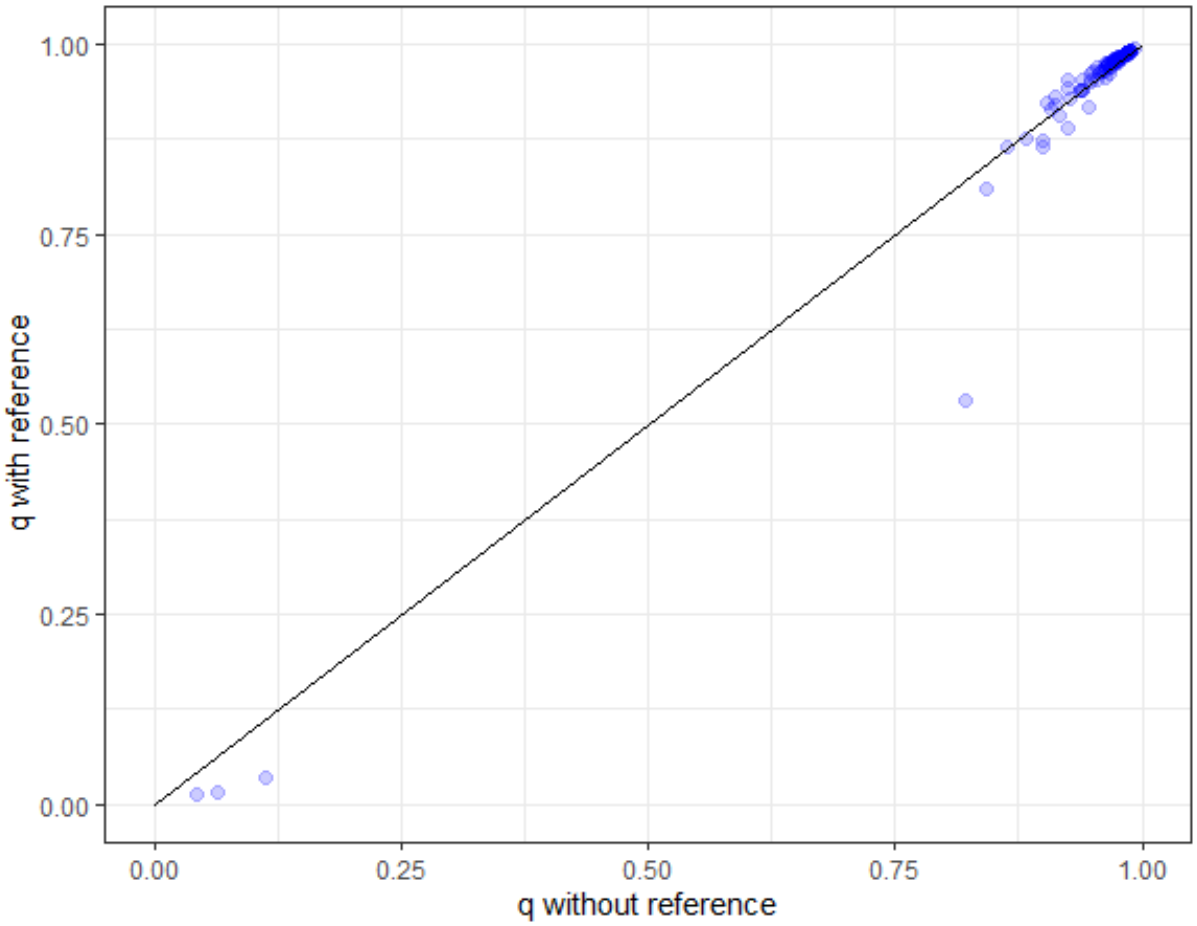

Figure S10. Comparison of *q* estimates obtained with ADMIXTURE, Ohana, sNMF and STRUCTURE. The data derived from simulations with FST=0.1 and 100 loci with 2 alleles. The different programs suffer from different biases. For these simulations, STRUCTURE offers the worse results, especially when the hybridization rate is high.