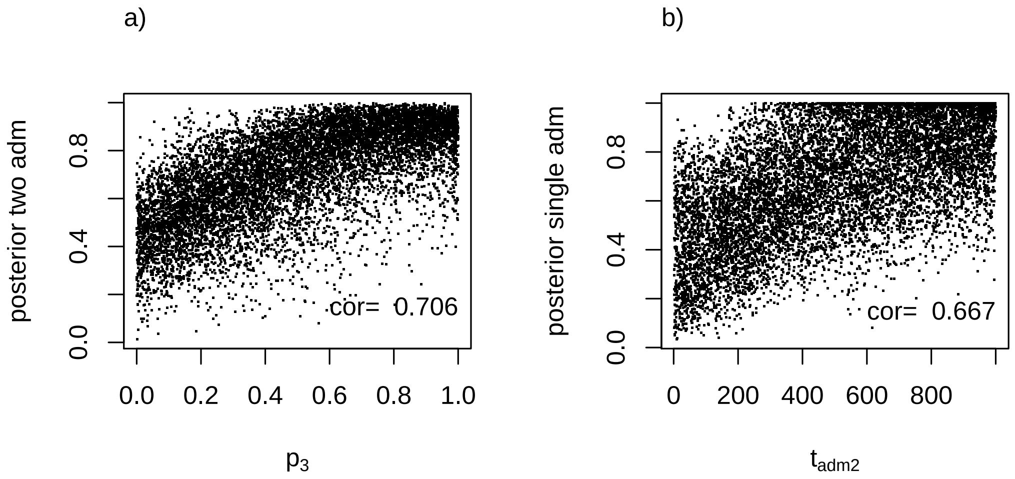
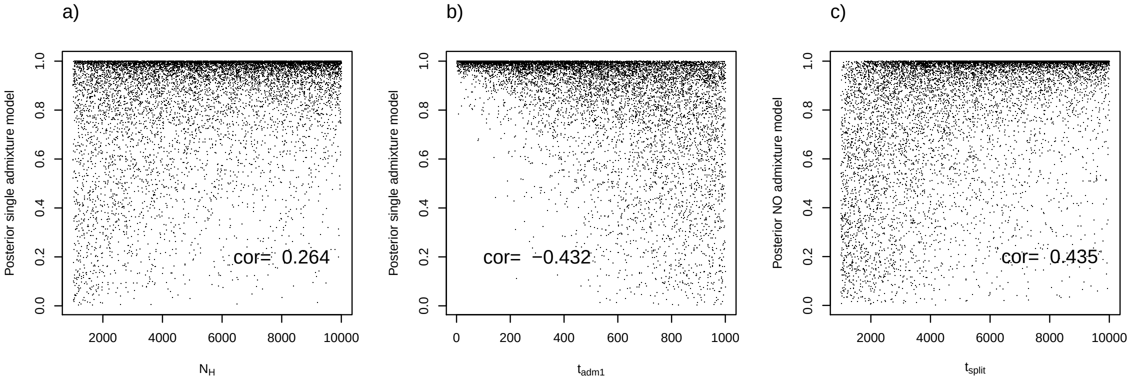
Supplementary Material

Population divergence with or without admixture: selecting models using an ABC approach

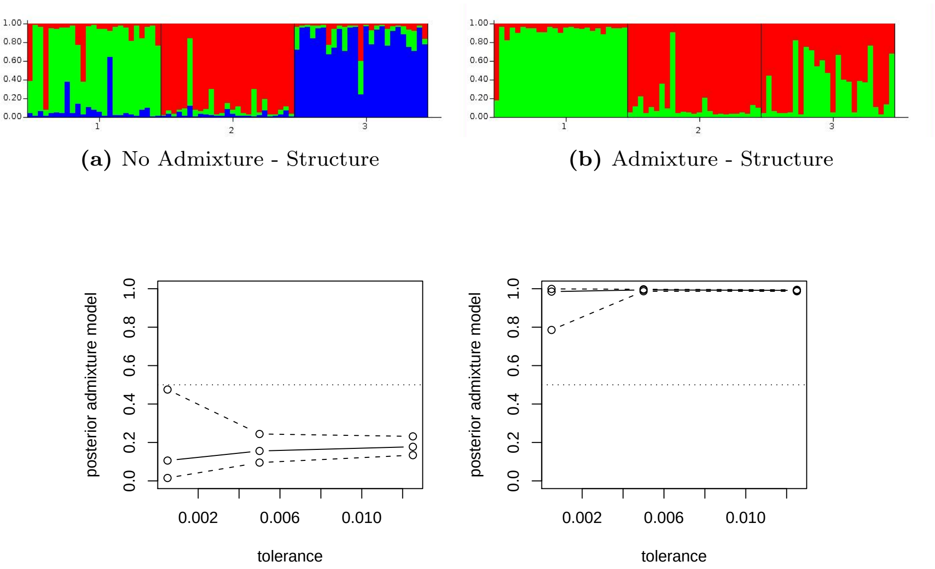
Vitor C. Sousa, Mark A. Beaumont, Pedro Fernandes,
Maria M. Coelho, Lounès Chikhi



Supplementary Figure 1: Posterior probability of the correct model as a function of the parameters used to simulate the datasets in the comparisons Single *vs* Two-event admixture models. a) Posterior probability for the two-event admixture model (two adm) as a function of the admixture contribution p_3 (datasets simulated under the two adm model). Increasing the admixture contribution of P_3 increases the probability of identifying the two adm model as the correct model. Note that when p_3 tends to zero, the two adm model approaches the single adm model. b) Posterior probability for the single-event admixture model (single adm) as a function of the time of admixture t_{adm} (datasets simulated under the single adm model). Increasing the time since admixture event increases the probability of identifying the single adm model as the correct model. The figures illustrate that the probability of identifying the correct model as the most likely depend on the parameter values. Each point corresponds to the posterior probability obtained for a given dataset. In total, 10,000 datasets with 20 independent loci were generated according to the models in figure 1f) and 1e), respectively. Results obtained with the logistic regression for a tolerance level of 0.0125. The Pearson correlation coefficients are shown in the plots.



Supplementary Figure 2: Posterior probability of the correct model as a function of the parameters used to simulate the datasets in the comparisons Single *vs* No admixture models. a) Posterior probability for the single-event admixture model (single adm) as a function of the effective size of the hybrid population N_H (datasets simulated under the single adm model). Increasing the effective size (lower drift) of the hybrid population increases the ability of identify the single adm as the correct model. b) Posterior probability for the single-event admixture model as a function of the time of admixture t_{adm_1} (datasets simulated under the single adm model). The older the time of admixture (t_{adm_1}) the most likely is that the NO admixture model is incorrectly selected as the most likely model. c) Posterior probability for the population split without admixture model (NO admixture) as a function of the t_{split} (datasets simulated under the NO admixture model). The older the time of split (t_{split}) the most likely is that the NO admixture model is correctly selected as the most likely model. The figures illustrate the fact that the probability of identifying the correct model as the most likely depend on the parameter values. Each point corresponds to the posterior probability obtained for a each dataset. In total, 10,000 datasets with 5 independent loci were generated according to the models in figure 1b), 1b) and 1a), respectively. Results obtained with the logistic regression for a tolerance level of 0.0125. The Pearson correlation coefficients are shown in the plots.



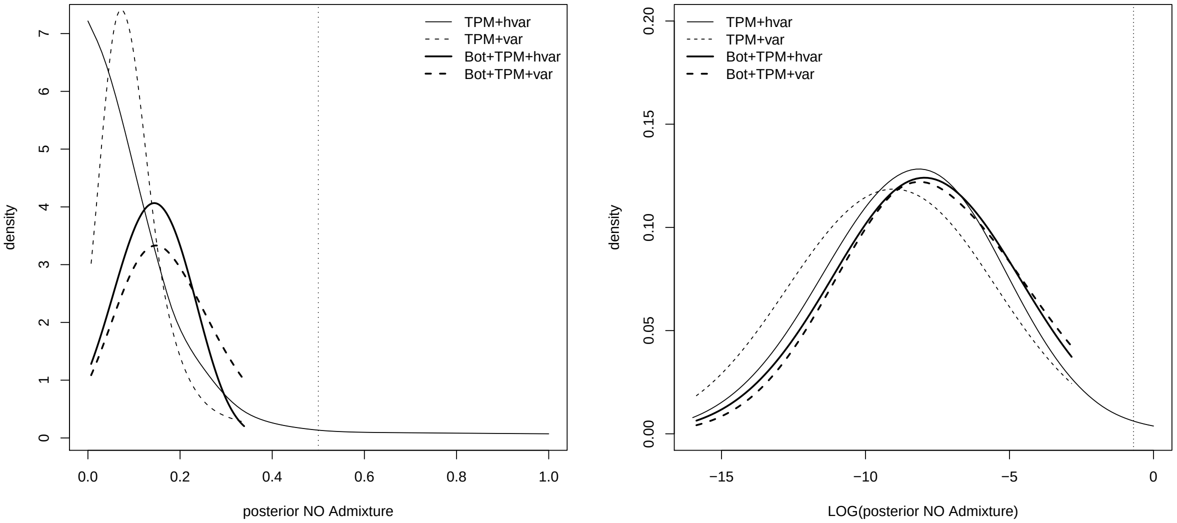
(a) No Admixture - Structure

(b) Admixture - Structure

(c) No Admixture - ABC

(d) Admixture - ABC

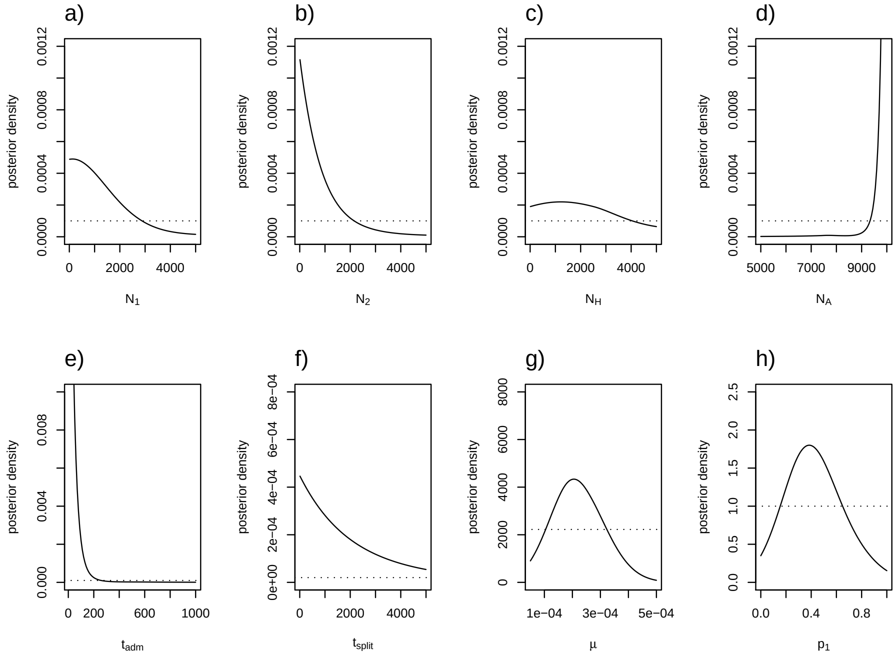
Supplementary Figure 3: Comparison of the ABC model choice and STRUCTURE results for datasets with five loci simulated under the population split with and without admixture models (single adm and NO admixture, respectively), shown in Figures 1a) and 1b). a) STRUCTURE results for the dataset simulated under the NO admixture model. Several admixed individuals identified in all populations, especially in population 1, P_1. This could have been interpreted as evidence for an admixture event, but it is actually due to shared ancestral polymorphism. b) STRUCTURE results for the dataset simulated under the admixture model. In population 3, P_H, several individuals show admixed genotypes. This is expected as the data was simulated assuming that population 3 resulted from an admixture event between population 1 and population 2. Note however, that one individual in population 1, and one individual from population 2, are incorrectly identified as admixed. c) ABC model choice results for the dataset simulated under the NO admixture model. The posterior probability of the admixture model is plotted as a function of the tolerance (solid line). Results obtained with the multinomial regression with 95% Confidence Intervals shown as dashed lines. The posterior probability for the admixture model is lower than ≈ 0.20 for most tolerance levels, suggesting that the NO admixture model has a higher posterior probability (≈ 0.80). The ABC method correctly identifies the NO admixture model has the most likely. d) ABC model choice results for the dataset simulated under the admixture model. The posterior probability of the admixture model is plotted as a function of the tolerance (solid line). Results obtained with the multinomial regression with 95% Confidence Intervals shown as dashed lines. The posterior probability is close to ≈ 1, suggesting that the admixture model is more likely. The ABC method correctly identifies the admixture model as the most likely. The NO admixture dataset was simulated with the following parameters: $N_1 = 2500$, $N_2 = 1250$, $N_3 = 2500$, $N_A = 5000$, $t_{split} = 1000$, $\mu = 0.0001$ (ms command line: ms 150 5 -t 2 -I 3 50 50 50 -n 1 0.5 -n 2 0.25 -n 3 0.50 -ej 0.05 3 1 -ej 0.05 2 1 -en 0.05 1 1). This dataset was analysed with STRUCTURE (a) and the ABC model choice method (c). The single-event admixture dataset was simulated with the following parameters: $N_1 = 3750$, $N_2 = 1250$, $N_H = 2500$, $N_A = 5000$, $t_{adm_1} = 50$, $t_{split} = 1000$, $\mu = 0.0001$, $p_1 = 0.3$ (ms command line: ms 150 5 -t 2.0 -I 3 50 50 50 -n 1 0.75 -n 2 0.25 -n 3 0.50 -es 0.0025 3 0.3 -ej 0.0025 4 2 -ej 0.0025 3 1 -ej 0.05 2 1 -en 0.05 1 1). This dataset was analysed with STRUCTURE (b) and the ABC model choice method (d). The population labels are shown for the STRUCTURE results, where $P_1 = 1, P_2 = 2, P_H = 3$. The hybrid population is the third one. The STRUCTURE v2.3 (Pritchard *et al.*, 2000; Falush *et al.*, 2003) analyses were performed with the following settings: 30,000 MCMC iterations were performed after a 10,000 iterations burnin, and with the default settings of STRUCTURE v2.3. The admixture model was assumed, inferring the admixture parameter alpha, assuming this was the same for all populations. The correlated allele frequency model was assumed, with different F_{ST} for each population, assuming a Dirichlet prior $D(1, ..., 1)$ for the allele frequencies. We ran the STRUCTURE program from K values ranging from 1 to 4. For each K value 3 different runs were performed. We then looked at the estimate of the likelihood of the data for each K value Ln(P(D—K)) to select the most likely value of K. For the dataset with admixture this was K=2, and for the No admixture model this was K=3.



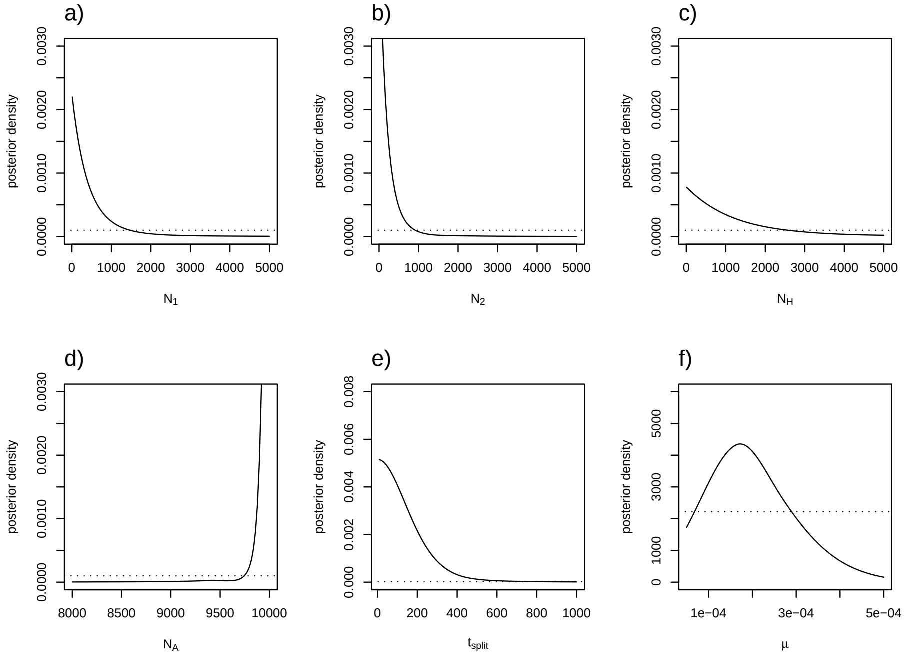
(a) ABC Rejection

(b) ABC Regression

Supplementary Figure 4: Robustness of ABC model choice to deviations from the model assumptions. Effect of the two phase mutation model (TPM), variation of mutation rate among loci (hvar - high variance, var - low variance) and bottlenecks (Bot) in the ability to detect the admixture model as the most likely. a) Distribution of the posterior for the NO admixture model obtained with the ABC rejection analyses of datasets simulated under an admixture model with violations of the model assumptions; b) Distribution of the LOG posterior for the NO admixture model obtained with the ABC regression (Beaumont, 2008). Note that the LOG scale was used because the probabilities were close to zero, as expected if the method was correctly identifying the admixture model as the most likely. The fact that the freshwater fish data suggested a recent population bottleneck (Sousa *et al.*, 2008), and that loci could have different mutation rates and not fit the single stepwise (SMM) mutation model lead to examine whether these deviations from the model assumptions could lead to a false detection of the NO admixture model as the most likely. Datasets with five independent loci were simulated according to a single admixture model with two parental populations (Figure 1b), allowing mutation rate to vary among loci according to a gamma distribution, following a TPM mutation model in which 85% of mutations were single step mutations, and the remaining mutation sizes were sampled from a geometric distribution with mean 2.0. The values of the parameters were chosen to reflect parameters similar to the ones that fit the fish data (see Supplementary Figure 5): $N_1 = N_2 = N_H = 1000$, $N_A = 10000$, $t_{adm} = 100$, $t_{split} = 1000$, $\mu = 1.5 \times 10^{-4}$, $p_1 = 0.3$. Mutation rate variation was allowed assuming a gamma distribution with mean 1.5×10^{-4}. We examined the effect of increasing the variance in the gamma distribution by considering a distribution with high variance (hvar) GammaDist($a = 5, b = 1.5 \times 10^{-4}/5$), and lower variance (var) GammaDist($a = 50, b = 1.5 \times 10^{-4}/50$). The curves show the effect of: (i) TPM mutation model (TPM); (ii) variation of the mutation rate among loci, either high variance (hvar) or low variance (var); (iii) effect of a 10-fold population bottleneck, 10 generations ago in all populations (bot) Each curve was obtained with the analysis of 100 simulated datasets. The ABC model choice procedure seems to be robust to these deviations, given that the posterior probability of the NO admixture is lower than ≈ 0.25 with the ABC Rejection (a), and close to zero with the ABC regression (b). The ms (Hudson, 2002) commands used to simulated these datasets were: ms 150 500 -t tbs -I 3 50 50 50 -es 0.0025 3 0.3 -ej 0.0025 4 2 -ej 0.0025 3 1 -ej 0.025 2 1 -en 0.025 1 1, for the TPM+hvar and TPM+var; and ms 150 500 -t tbs -I 3 50 50 50 -n 1 0.1 -n 2 0.1 -n 3 0.1 -es 0.0025 3 0.3 -ej 0.0025 4 2 -ej 0.0025 3 1 -ej 0.025 2 1 -en 0.025 1 1, for the Bot+TPM+hvar and Bot+TPM+var. The microsat program of Hudson (2002) was modified to simulated data according to a TPM mutation model.



Supplementary Figure 5: Posterior distribution for the parameters of the admixture model estimated for the *Iberochondrostoma lusitanicum* dataset. Parameters as the ones in Figure 1b. a) effective size of parental population 1 (N_1); b) effective size of parental population 2 (N_2); c) effective size of hybrid population (N_H); d) effective size of ancestral population (N_A); e) time of admixture (t_{adm_1}); f) time of split (t_{split}); g) mutation rate per generation (μ); h) admixture contribution from population P_1 (p_1). Estimates obtained with the ABC rejection and regression method based on 1,000,000 simulations, with a tolerance level of 0.01. Prior distributions shown as a horizontal dotted line. Prior distributions according to prior set 1 (see main text for details).



Supplementary Figure 6: Posterior distribution for the parameters of the population split without admixture (no admixture) model estimated for the *Iberochondrostoma lusitanicum* dataset. Parameters as the ones in Figure 1a. a) effective size of population 1 (N_1); b) effective size of population 2 (N_2); c) effective size of population 3 (N_3); d) effective size of ancestral population (N_A); e) time of split (t_{split}); f) mutation rate per generation (μ). Estimates obtained with the ABC rejection and regression method based on 1,000,000 simulations, with a tolerance level of 0.01. Prior distributions shown as a horizontal dotted line. Prior distributions according to prior set 1 (see main text for details).

References

Beaumont MA (2008). *Simulation, Genetics, and Human Prehistory*, McDonald Inst of Archeological, Cambridge, vol. McDonald Inst of Archeological, chap. Joint determination of tree topology and population history, pp. 134–154.

Falush D, Stephens M, Pritchard J (2003). Inference of population structure using multilocus genotype data linked loci and correlated allele frequencies. *Genetics* **164**: 1567–1587.

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

Pritchard JK, Stephens M, Donnelly P (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–59.

Sousa V, Penha F, Collares-Pereira MJ, Chikhi L, Coelho MM (2008). Genetic structure and signature of population decrease in the critically endangered freshwater cyprinid *Chondrostoma lusitanicum*. *Conserv Genet* **9**: 791–805.