iForest
Biogeosciences and Forestry

Research Article
doi: 10.3832/ifor2569-011
vol. 11, pp. 99-103

INCOTW - Sassari, Italy (2017)
**"International Congress on Cork Oak Trees and Woodlands"**
Guest Editors: Piermaria Corona, Sandro Dettori

# SimHyb: a simulation software for the study of the evolution of hybridizing populations. Application to *Quercus ilex* and *Q. suber* suggests hybridization could be underestimated

**Álvaro Soto,**
**David Rodríguez-Martínez,**
**Unai López De Heredia**

**We present SimHyb, a Java-based software for the simulation of mixed hybridizing populations. The software incorporates user-defined initial parameters and input files to account for the initial census size of two species in a closed population, the number of intermediate specific classes, the directional fertility among specific classes, the fitness coefficients for each specific class, the inheritance of fitness, and the degree of ageing and self-incompatibility of the individuals. All these demographic and adaptive parameters can be modified by the user to analyze their effect on the evolution of the mixed population. SimHyb allows the traceability of each individual, whose pedigree is also recorded. For each simulated generation the software yields an output file that is easily convertible to an input for Structure, one of the most popular softwares for the Bayesian analysis of populations. Application of SimHyb to simulate *Quercus ilex* and *Q. suber* hybridizing populations, and further analysis with Structure, reveals that advanced introgressed individuals are very often misclassified with the currently available set of nuclear microsatellite markers, so that introgression between these two species could have been underestimated in previous studies. However, we provide a simple parameter based on Structure results to identify the directionality of pollination in the progeny of a known mother tree.**

**Keywords: Hybridization, Introgression, Simulations, Molecular Markers, *Quercus suber*, *Quercus ilex***

## Introduction

Hybridization and introgression have presumably played a key role in the evolution of plants and, probably to a minor extent, of animals (Mallet 2005). Gene transfer can provide the hybridizing species with new adaptive skills, making them able to endure new environmental stresses or to colonize new habitats (Rieseberg et al. 2003, Petit et al. 2004), or can even lead to the appearance of new species (Seehausen 2004). Identification of hybrid individuals has traditionally been based on phenotypic characters. However, such identification is not always obvious, and in the last decades genetic information based on molecular markers has been used for this purpose. Depending on the molecular marker of choice, inferences about evolutionary or population processes can be obtained at different time scales. Molecular markers from organellar DNA have been successfully used to infer ancient hybridization in plants, for instance, in the case of *Quercus ilex* and *Q. suber* (Jiménez et al. 2004, Lumaret et al. 2005, Magri et al. 2007). Nuclear molecular markers combined with Bayesian approaches have been used for the identification of hybrid individuals in populations, according to the posterior probability of each individual to belong to any of the parental species genetic clusters. This approach was applied by Burgarella et al. (2009) in sympatric *Q. ilex* – *Q. suber* populations, estimating a current introgression rate of <2%. Later on, this procedure has been applied in studies for many other species, including animals (Neaves et al. 2010, Bogdanowicz et al. 2012, Malde et al. 2017).

Before proceeding to the quantification of introgression rates, the detection power and accuracy of the set of markers used to identify hybrid individuals needs to be estimated. According to this detection power, a threshold to classify a particular individual as a hybrid can further be established. This task is usually performed using virtual individuals whose actual specific category is accurately known. The aforementioned works usually employ software such as HybridLab (Nielsen et al. 2006) for this purpose. In these types of programs, the generation of a virtual hybrid genotype is performed by drafting an allele from each genetic pool (parental species) at each locus, according to the specific allele frequencies provided by the user. The same procedure

GI Genética, Fisiología e Historia Forestal, Dept. Sistemas y Recursos Naturales, ETSI Montes, Forestal y del Medio Natural, Universidad Politécnica de Madrid, Madrid (Spain)

@ Álvaro Soto (alvaro.soto.deviana@upm.es)

Received: Jul 28, 2017 - Accepted: Jan 12, 2018

**Citation:** Soto A, Rodríguez-Martínez D, López De Heredia U (2018). SimHyb: a simulation software for the study of the evolution of hybridizing populations. Application to *Quercus ilex* and *Q. suber* suggests hybridization could be underestimated. iForest 11: 99-103. – doi: 10.3832/ifor2569-011 [online 2018-01-31]

*Communicated by: Piermaria Corona*

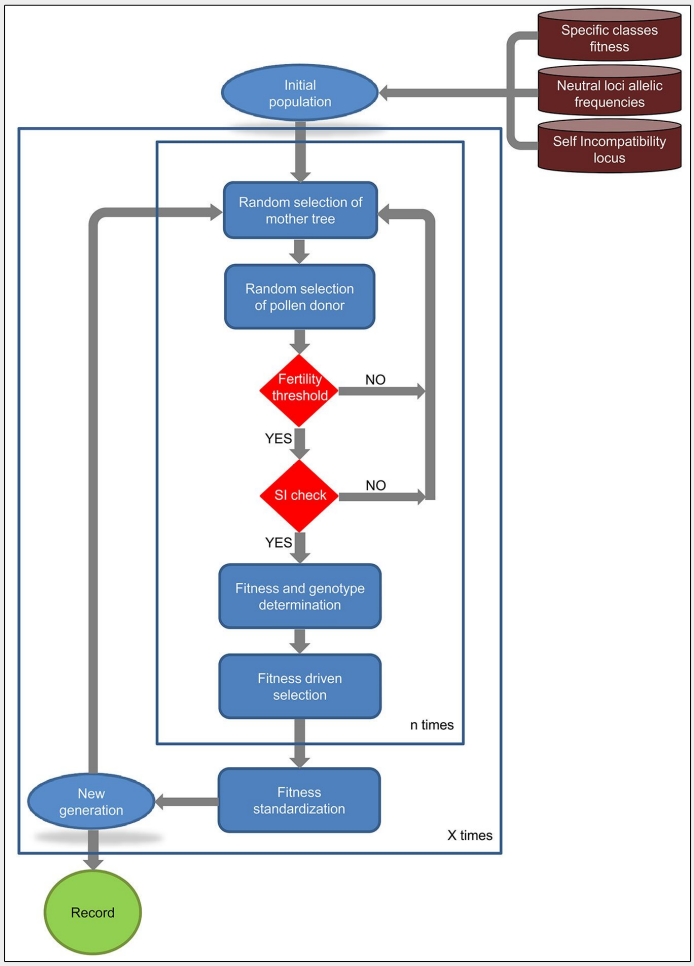


**Fig. 1** - Overview of the workflow of SIMHYB.

is followed to obtain F2 or backcrosses, considering the F1 hybrid pool as a parental species.

Here we present SIMHYB, a software that follows a different approach, since it allows the simulation of the evolution of a mixed hybridizing population throughout generations, as well as the complete traceability of each individual pedigree. Thus, the expected contribution of each parental species to the genome of any individual is perfectly known. The program also allows the analysis of the effect of demographic and selective factors on the evolution of the population. SIMHYB provides an output easily convertible to an input file for STRUCTURE (Pritchard et al. 2000), one of the most commonly used softwares for the Bayesian analysis of populations and detection of hybrids.

As a case study, we have simulated *Q. ilex* and *Q. suber* hybridizing populations, considering nine widely used nuclear microsatellites. Further analysis of the SIMHYB output individuals with STRUCTURE have revealed that introgression between these species could have been underestimated in previous studies.

## Material and methods

### *Conceptual approaches*

SIMHYB was programmed in Java 8, and runs in any computer with an OS that allows for Java (https://www.java.com/), or OpenJDK (http://openjdk.java.net/): Linux/Unix, Microsoft Windows, Mac OS X, and other platforms. The software is available at: http://www.gfhforestal.com/software.

### *Input files and parameters*

Before starting the simulation, an interface window appears where the user must define the simulation parameters and input files (the Instructions Manual provided with the software includes examples of these input files). The user can define the population size for each species, considering only adult individuals. The size of the global population will be fixed throughout the simulation. The user must also provide the file with the loci and the allele frequencies of each species. Another important input file includes the specific classes and the fitness coefficient corresponding to each class. The user can define as many specific classes as desired, intermediate between the two pure species included in the original population, and can establish the limits of each class, according to the so-called "species coefficient". Fertility among the different specific classes, which takes into account the pollination direction, is also defined as an input. Maternal or paternal inheritance of the chloroplast is also defined in this interface.

### *Simulation run*

SIMHYB creates an initial population with the specified number of adult individuals per species. Each individual has a multilocus genotype, with alleles drafted according to the allele frequencies provided by the user. The fitness coefficient of each individual corresponds to the one defined in the input, considering also the intraspecific variability $\varepsilon_{sp}$. Once the initial population is created, SIMHYB simulates successive "generations", in which new individuals are incorporated to the reproductive population while other individuals disappear (Fig. 1).

Each cycle or generation includes four phases: (1) reproduction; (2) ageing; (3) selection; and (4) standardization. No spatial restriction is considered in the current version of SIMHYB for reproduction (phase 1), *i.e.*, pollen from any individual can reach the flowers of any other tree with equal probability. The spatial positions of the individuals are not reckoned in SIMHYB and, accordingly, there is no isolation by distance. The probability of obtaining a viable cross is firstly determined by the fertility coefficients defined in the input file "fertility.txt", according to the specific classes of the mother tree and the pollen donor. After that, effective pollination is limited by self-incompatibility, if this option is selected. SIMHYB considers gametophytic self-incompatibility driven by a single locus. Doing so, one of the alleles at this locus from the pollen donor is randomly selected and compared with both alleles of the mother tree. After passing these barriers, the new individual will carry at each locus one of the alleles of each parent, randomly selected, and the chloroplast and mitochondrial DNA of the corresponding parent, as specified by the user. The species coefficient of the new individual will be the average of the parents' values, so it becomes a precise estimation of the contribution of each genetic pool (pure species) to the genome of the individual. A new fitness coefficient will be assigned to the new individual according to the following formulas (eqn. 1, eqn. 2):

$$f = w_h f_h + w_{sp} f_{sp} \quad (1)$$

$$f_h = \frac{f_{mother} + f_{father}}{2} + \varepsilon_h \quad (2)$$

where $f_{sp}$ is the fitness coefficient corresponding to the specific class of the new individual, $f_h$ is the fitness inherited from the parent trees, $\varepsilon_h$ is a variability parameter, to include variability among full-sibs, and $w_h$ and $w_{sp}$ are weights, established by the user.

Ageing (phase 2) affects each individual fitness every cycle, according to this formula:

$$f_t = f_{t-1}(1-a)b + f_0(1-a)^t(1-b) \quad (3)$$

where $f_t$ is the fitness coefficient $t$ cycles after birth, $f_0$ is the initial fitness, at the time

of birth, $a$ is the ageing coefficient, and $b$ is the linearity coefficient, which varies between 0 and 1. However, the user should be cautious with high values of $b$, since they can lead to virtually eternal individuals (see below).

Selection (phase 3) takes place at this moment, and the $N$ (population size, defined by the user) individuals with the highest fitness coefficients are selected and remain in the reproductive population, while the other individuals die. In the last step of the simulation, standardization (phase 4), the fitness coefficients are normalized between 0 and 1 (this way, if high $b$ and $w_h$ have been selected, certain individuals can keep very high fitness values along the generations).

*Output files*

The user can select to finish the simulation when one of the chloroplasts is completely replaced by the other one or after a desired number of cycles. SIMHYB provides as output a picture of the complete evolution of the population, as a series of "snapshots" after each reproductive cycle in a single *.csv* file where each individual appears as a row. The first row includes the headings of the first ten columns, while the second row includes the name of the loci included in the genotypes. The next rows correspond to the individuals in each generation ("snapshot" of the population). Each individual appears in a row, including its ID, species coefficient, father, mother, chloroplast, mitochondria, current generation, generation of birth, generation of death, fitness coefficient at the current generation, and the genotype, including the diploid self-incompatibility locus and the neutral loci defined by the user. This way, the output file is easily convertible into a STRUCTURE input file, simply removing the non-desired rows. Columns 3 to 12 could be considered as extra-columns in STRUCTURE analysis.

*Case study: hybridization and introgression in the Quercus ilex × suber complex*

In order to confirm the full operability of the software, a simulation study was performed on the *Q. ilex* and *Q. suber* hybridizing complex employing nine microsatellites that are commonly used to address hybridization in these species: MSQ4, MSQ13 (Dow et al. 1995), QpZAG9, QpZAG15, QpZAG36, QpZAG46, (Steinkellner et al. 1997), QrZAG7, QrZAG11 and QrZAG20 (Kampfer et al. 1998). We used the allele frequencies obtained for the pure *Q. ilex* and *Q. suber* individuals identified in Burgarella et al. (2009), covering the whole distribution range of *Q. suber*. We run several simulations with a variable number of input parameters to produce pure and hybrid individuals of known pedigree. The output of the simulations was downstreamed to STRUCTURE 2.3.4 (Pritchard et al. 2000), and the individual $q$ values of the offspring, and the ratio between the offspring and the maternal $q$ were calculated to infer the detection power of the aforementioned microsatellites in identifying hybrids and advanced introgression. For STRUCTURE analyses we used the admixture model assuming independent allele frequencies, a burn-in of 50,000 steps followed by 100,000 iterations, as in Burgarella et al. (2009). Results do not vary significantly across multiple runs and with longer burn-in/iteration cycles. Statistical analyses were performed using the R software (R Core Team 2013).

**Tab. 1** - Specific classes defined in the simulation, with the corresponding range for the SIMHYB specific coefficient. The expected values (average and range, in brackets) for each category of the $q_s$ provided by STRUCTURE are also included.

| Specific category | SIMHYB specific coefficient | STRUCTURE expected $q_s$ |
|---|---|---|
| 1 (*Q. suber*) | [0, 100] | 0.982 (0.858 - 1) |
| 2 | (100, 200] | 0.786 (0.715 - 0.858) |
| 3 | (200, 300] | 0.643 (0.572 - 0.715) |
| 4 | (300, 400) | 0.500 (0.429 - 0.572) |
| 5 | [400, 500) | 0.358 (0.286 - 0.429) |
| 6 | [500, 600) | 0.215 (0.143 - 0.286) |
| 7 (*Q. ilex*) | [600, 700] | 0.072 (0 - 0.143) |

**Tab. 2** - Example of assignation of simulated individuals to specific categories according to STRUCTURE $q_s$. Simulated population includes 5% of first generation hybrids (specific category 4) and 5% of advanced introgressed individuals (categories 2 and 6).

| Specific coefficient | True specific category | STRUCTURE-assigned specific category (%) | | | | | | | Total number |
|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
| 0 | 1 | 100.00 | - | - | - | - | - | - | 6355 |
| (0, 100) | 1' | - | - | - | - | - | - | - | - |
| (100, 200] | 2 | 39.94 | 42.49 | 16.43 | 1.13 | - | - | - | 353 |
| (200-300] | 3 | - | - | - | - | - | - | - | 0 |
| (300, 400) | 4 | - | - | 17.00 | 70.25 | 12.61 | 0.14 | - | 706 |
| [400, 500) | 5 | - | - | - | - | - | - | - | 0 |
| [500, 600) | 6 | - | - | - | 10.20 | 33.15 | 34.84 | 21.81 | 353 |
| [600, 700) | 7' | - | - | - | - | - | - | - | - |
| 700 | 7 | - | - | - | - | - | 0.13 | 99.87 | 6356 |

## Results and discussion

*Case study: hybridization and introgression in the Quercus ilex × suber complex*

The simulation of the evolution of mixed, hybridizing *Q. ilex* and *Q. suber* populations, confirmed the full operability of SIMHYB. In the simulations reported here, we defined seven specific classes. A species coefficient of 0 was assigned to pure *Q. suber* individuals in the initial population, and 700 to pure *Q. ilex* individuals (Tab. 1). The output of SIMHYB provided hybrid and

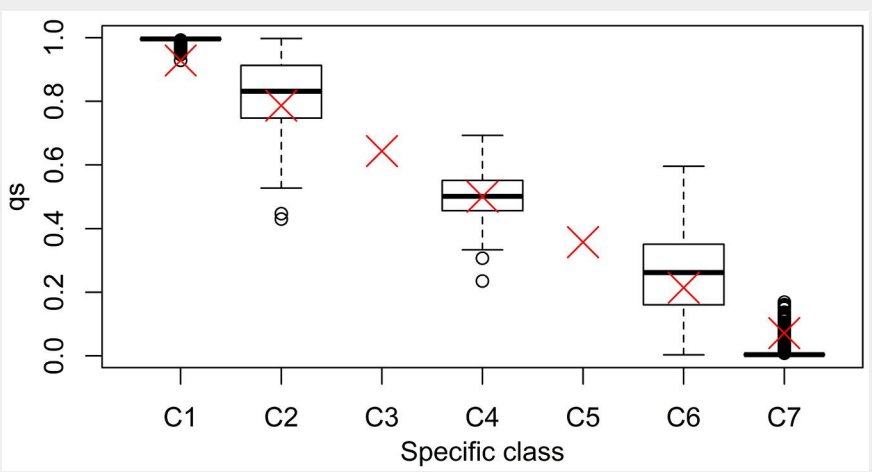


**Fig. 2** - Boxplot of the distribution of STRUCTURE-assigned $q_s$ to virtual individuals, according to their specific class. The red cross represents the expected $q_s$ value.

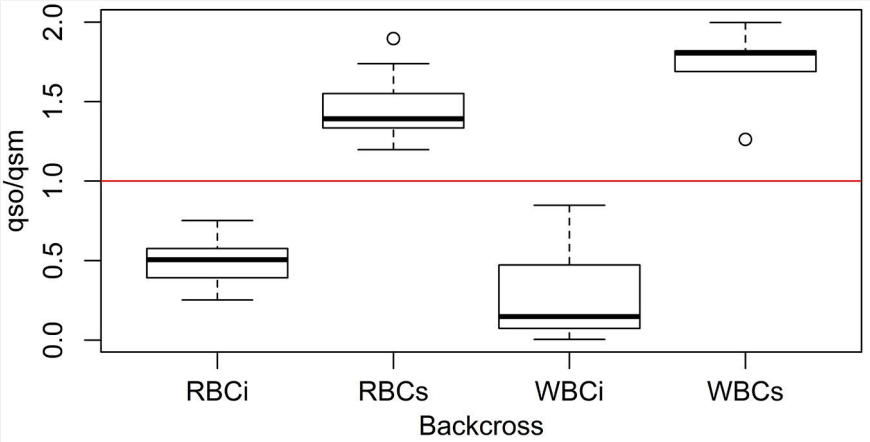


**Fig. 3** - Boxplot of the distribution of the ratio $q_s$(offspring)/$q_s$(mother) for simulated backcrosses. (RBCi): backcrosses F1 hybrid × *Q. ilex* rightly classified according to STRUCTURE $q_s$; (RBCs): backcrosses F1 hybrid × *Q. suber* rightly classified according to STRUCTURE $q_s$; (WBCi): backcrosses F1 hybrid × *Q. ilex* wrongly classified according to STRUCTURE $q_s$; (WBCs): backcrosses F1 hybrid × *Q. suber* wrongly classified according to STRUCTURE $q_s$. Values below 1 (red line) indicates the hybrid has been pollinated by *Q. ilex* while values above 1 indicates pollination by *Q. suber*.

introgressed individuals whose pedigree could be easily reconstructed, and whose species coefficient provided a direct quantification of the contribution of each species to its genome. Tab. 1 includes the expected values of *q*, the probability of an individual to belong to the *Q. suber* cluster, provided by STRUCTURE. This parameter can be used to detect hybrid individuals (Burgarella et al. 2009) and is frequently considered a proxy of the contribution of one species to the genome. Burgarella et al. (2009) reported a good performance of STRUCTURE and this set of markers for the detection of first generation hybrids and first backcrosses, in terms of efficiency and accuracy (*sensu* Vähä & Primmer 2006). On the contrary, we have detected a high proportion of wrong assignations for advanced introgressed individuals (hybrid individuals corresponding to further backcrosses). In a scenario of a low hybridization rate as the one reported by Burgarella et al. (2009) a high percentage of advanced introgressed individuals could go unnoticed with this set of markers. Tab. 2 and Fig. 2 summarize the results obtained for a simulated population with a 5% of first generation hybrids and a 5% of advanced introgressed individuals, highlighting the wrong assignations particularly for this latter group. Results for other simulated populations are provided in Tab. S1 to Tab. S4 in Supplementary material.

However, since SIMHYB provides individuals of known pedigree, we performed direct comparisons of the *q* parameter provided by STRUCTURE for different individuals and for their parents. Our results show that the ratio $q_{offspring}/q_{mother}$ helps to assess the specific identity of the pollen donor. If this ratio is higher than 1, the pollen donor should have been a *Q. suber* individual (or, at least, an individual with a higher proportion of *Q. suber* genome than the mother tree), while, if the *q* ratio is lower than one, the pollen donor belongs to the *Q. ilex* group. Fig. 3 shows an example of how this parameter behaves in the case of backcrosses rightly and wrongly assigned based exclusively in the STRUCTURE *q* parameter.

## Conclusion

SIMHYB is a user friendly software for the simulation of hybridizing populations, and the output provided is easily convertible into an input file for STRUCTURE, one of the most popular software packages for the analysis of admixture using genotypic data. It can be used both for research and for educational purposes, to analyze the effect of the different factors influencing the evolution of hybridizing populations, such as population size, inter- and intraspecific fertility, directionality in the effective pollination, selective advantage (fitness), ageing, etc.

Application to *Q. ilex* and *Q. suber* with a set of commonly used nuclear microsatellite markers has revealed that advanced introgression could have been missed in previous studies, so that rates of current hybridization between these species could have been underestimated. Nevertheless, we have found that ratio of the *q* parameter provided by STRUCTURE for a seedling and for its mother tree provides a good indication of the specific identity of the pollen donor.

## Acknowledgements

AS & ULH conceived the idea; AS designed the program and drafted the manuscript; DRM programmed the software. All the authors have contributed to the final manuscript.

Development of this software has been funded by the project AGL2015-67495-C2-2-R (Spanish Ministry of Economy and Competitiveness)

## References

Bogdanowicz W, Piksa K, Tereba A (2012). Hybridization hotspots at bat swarming sites. PLoS ONE 7 (12): e53334. - doi: 10.1371/journal.pone.0053334

Burgarella C, Lorenzo Z, Jabbour-Zahab R, Lumaret R, Guichoux E, Petit RJ, Soto A, Gil L (2009). Detection of hybrids in nature: application to oaks (*Quercus suber* and *Q. ilex*). Heredity 102: 442-452. - doi: 10.1038/hdy.2009.8

Dow B, Ashley M, Howe H (1995). Characterization of highly variable (GA/CT)n microsatellites in the bur oak, *Quercus macrocarpa*. Theoretical and Applied Genetics 91: 137-141. - doi: 10.1007/BF00220870

Jiménez P, López De Heredia U, Collada C, Lorenzo Z, Gil L (2004). High variability of chloroplast DNA in three Mediterranean evergreen oaks indicates complex evolutionary history. Heredity 93: 510-515. - doi: 10.1038/sj.hdy.6800551

Kampfer S, Lexer C, Glössl J, Steinkellner H (1998). Characterization of $(GA)_n$ microsatellite loci from *Quercus robur*. Hereditas 129: 183-186. - doi: 10.1111/j.1601-5223.1998.00183.x

Lumaret R, Tryphon-Dionnet M, Michaud H, Sanuy A, Ipotesi E, Born C, Mir C (2005). Phylogeographical variation of chloroplast DNA in cork oak (*Quercus suber*). Annals of Botany 96: 853-861. - doi: 10.1093/aob/mci237

Magri D, Fineschi S, Bellarosa R, Buonamici A, Sebastiani F, Schirone B, Simeone MC, Vendramin GG (2007). The distribution of *Quercus suber* chloroplast haplotypes matches the palaeogeographical history of the western Mediterranean. Molecular Ecology 16: 5259-5266. - doi: 10.1111/j.1365-294X.2007.03587.x

Malde K, Seliussen BB, Quintela M, Dahle G, Besnier F, Skaug HJ, Olien N, Solvang HK, Haug T, Skern-Mauritzen R, Kanda N, Pastene LA, Jonassen I, Glover KA (2017). Whole genome resequencing reveals diagnostic markers for investigating global migration and hybridization between minke whale species. BMC Genomics 18 (1): 83. - doi: 10.1186/s12864-016-3416-5

Mallet J (2005). Hybridization as an invasion of the genome. Trends in Ecology and Evolution 20: 229-237. - doi: 10.1016/j.tree.2005.02.010

Neaves LE, Zenger KR, Cooper DW, Eldridge MDB (2010). Molecular detection of hybridization between sympatric kangaroo species in south-eastern Australia. Heredity 104: 502-512. - doi: 10.1038/hdy.2009.137

Nielsen EE, Bach LA, Kotlick P (2006). Hybridlab (version 1.0): a program for generating simulated hybrids from population samples. Molecular Ecology Notes 6: 971-973. - doi: 10.1111/j.1471-8286.2006.01433.x

Petit RJ, Bialozyt R, Garnier-Gere P, Hampe A (2004). Ecology and genetics of tree invasions: from recent introductions to Quaternary migrations. Forest Ecology and Management 197: 117-137. - doi: 10.1016/j.foreco.2004.05.009

Pritchard JK, Stephens M, Donnely P (2000). Inference of population structure using multilocus genotype data. Genetics 155: 945-959. [online] URL: http://www.genetics.org/cont

ent/155/2/945
R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. [online] URL: http://www.R-project.org/
Rieseberg LH, Raymond O, Rosenthal DM, Lai Z, Livingstone K, Nakazato T, Durphy JL, Schwarzbach AE, Donovan LA, Lexer C (2003). Major ecological transitions in wild sunflowers facilitated by hybridization. Science 301: 1211-1216. - doi: 10.1126/science.1086949
Seehausen O (2004). Hybridization and adaptive radiation. Trends in Ecology and Evolution 19: 198-207. - doi: 10.1016/j.tree.2004.01.003
Steinkellner H, Fluch S, Turetschek E, Lexer C, Streiff R, Kremer A, Burg K, Glössl J (1997). Identification and characterization of $(GA/CT)_n$-microsatellite loci from *Quercus petraea*. Plant Molecular Biology 3: 1093-1096. - doi: 10.1023/A:1005736722794
Vähä JP, Primmer CR (2006). Efficiency of model-based Bayesian methods for detecting hybrid individuals under different hybridization scenarios and with different number of loci. Molecular Ecology 15: 63-72. - doi: 10.1111/j.1365-294X.2005.02773.x

## Supplementary Material

**Tab. S1** - Example of assignation of simulated individuals to specific categories according to STRUCTURE $q_s$. Simulated population includes 5% of first generation hybrids (specific category 4).

**Tab. S2** - Example of assignation of simulated individuals to specific categories according to STRUCTURE $q_s$. Simulated population includes 10% of first generation hybrids (specific category 4).

**Tab. S3** - Example of assignation of simulated individuals to specific categories according to STRUCTURE $q_s$. Simulated population includes 5% of first generation hybrids (specific category 4) and 18% of advanced introgressed individuals (categories 2 and 6).

**Tab. S4** - Example of assignation of simulated individuals to specific categories according to STRUCTURE $q_s$. Simulated population includes 0.5% of first generation hybrids (specific category 4) and 29.5% of advanced introgressed individuals (categories 1', 2, 6 and 7').

**Link:** Soto_2569@suppl001.pdf