INSTRUCTION MANUAL

SIMHYB: a simulation software for the study of the evolution of hybridizing populations

Version 1.0

Álvaro Soto de Viana[1]

David Rodríguez Martínez

Unai López de Heredia Larrea[1]

[1]GI Genética, Fisiología e Historia Forestal.
Dpto. Sistemas y Recursos Naturales.
ETSI Montes, Forestal y del Medio Natural
Universidad Politécnica de Madrid.

http://gfhforestal.com/software

July, 2017

CONTENTS

OVERVIEW

SIMHYB is a Java-based software for the simulation of mixed hybridizing populations. The program is intended for the analysis of the effect of the different demographic and adaptive parameters on the evolution of these populations. Census size of each species, number of intermediate specific classes, directional fertility among them and fitness coefficient of each class can be defined by the user. Inheritance of fitness and ageing effect are also taken into account. The software generates individuals of known pedigree, allowing their traceability throughout the generations. SIMHYB yields for each simulated generation an output file easily convertible to an input for STRUCTURE (Pritchard JK, Stephens M, Donnely P (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945-959), one of the most popular softwares for the Bayesian analysis of populations.

SIMULATION PROCESS

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 intraspecific variability, ε_{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.

Each cycle or generation includes four phases: 1) reproduction, 2) ageing, 3) selection, and 4) standardization. For **1) reproduction**, no spatial restriction is considered in the current version of SIMHYB, i.e., pollen from any individual can reach the flowers of any other tree with the same 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. This way, at pollination, 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 mitochondria 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 formula:

$$f = w_h f_h + w_{sp} f_{sp}$$

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

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, ε_h is a variability parameter, to include variability among full-sibs, and w_h and w_{sp} are weights, established by the user.

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

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

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).

3) Selection 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. Finally, **4) standardization** is performed, so that fitness

coefficients vary 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).

GETTING STARTED

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.

Download and unzip the package and copy the executable file, `SIMHYB.jar`, and the `simhyb.properties` file in the desired location of your computer. Every time you launch a simulation, this file will be overwritten with the parameters and instructions of the current simulation. It is therefore recommended to save this file together with the results for your own records. You don't have to worry about the content of the first copy of the file, included in InstallSIMHYB.rar since it will be overwritten, but the executable needs to have this file to begin the simulations. If your computer has a Linux based OS of your computer, we recommend modifying the path of the input files by editing the following lines in the simhyb.properties file:

`specificCategoryFile=`/YOURPATH/specificclasses.txt

`alleleFrequenciesFile=`/YOURPATH/allelefrequiencies.csv

`selfIncompatibilityLocusFile=`/YOURPATH/selfincompatibility.csv

`fertilityTableFile=`/YOURPATH/fertility.txt

Before launching the program, create a folder named logs at the same directory where the executable is located.

It is also recommended to place all the needed input files in the same directory where the executable is located.
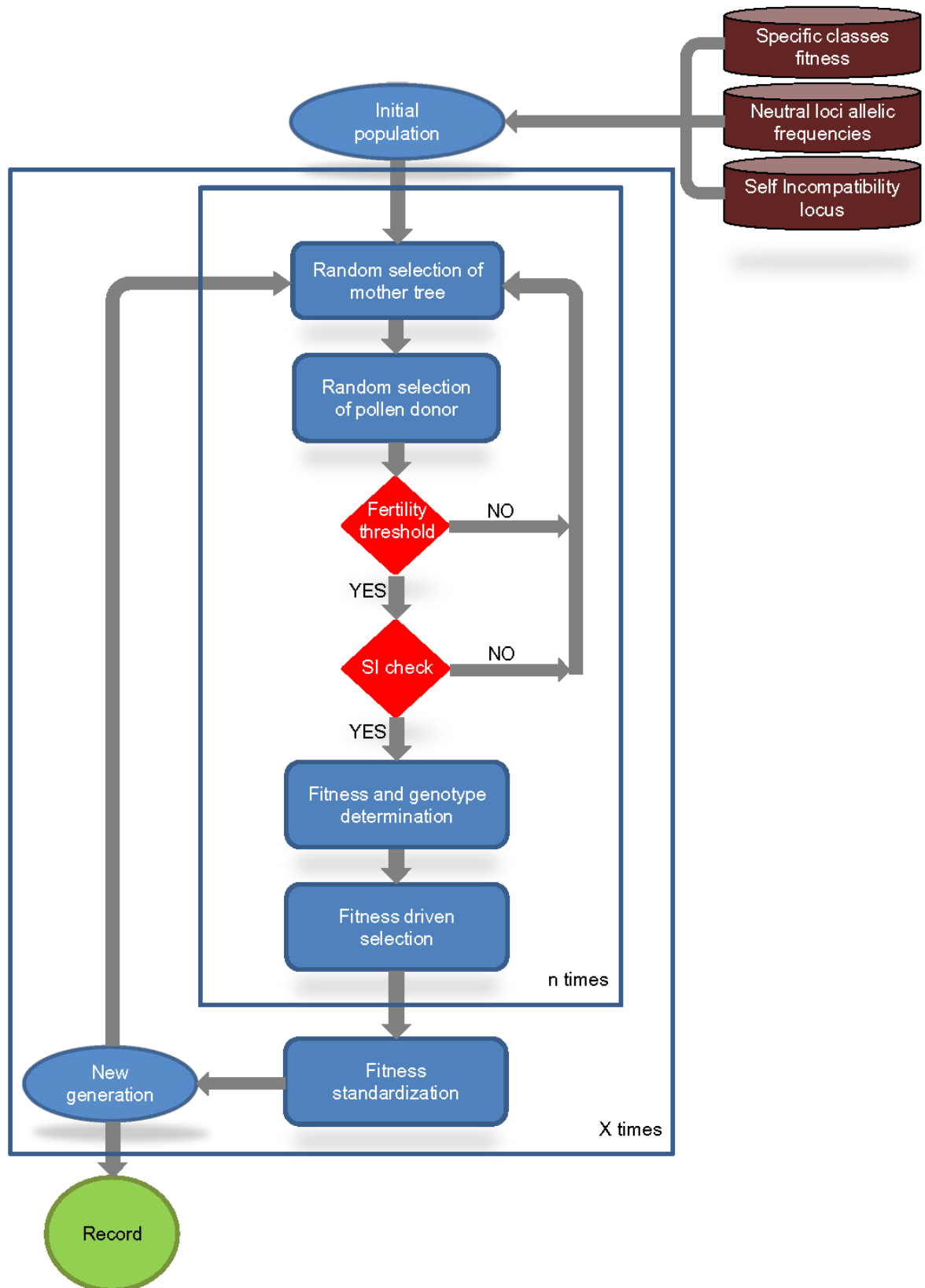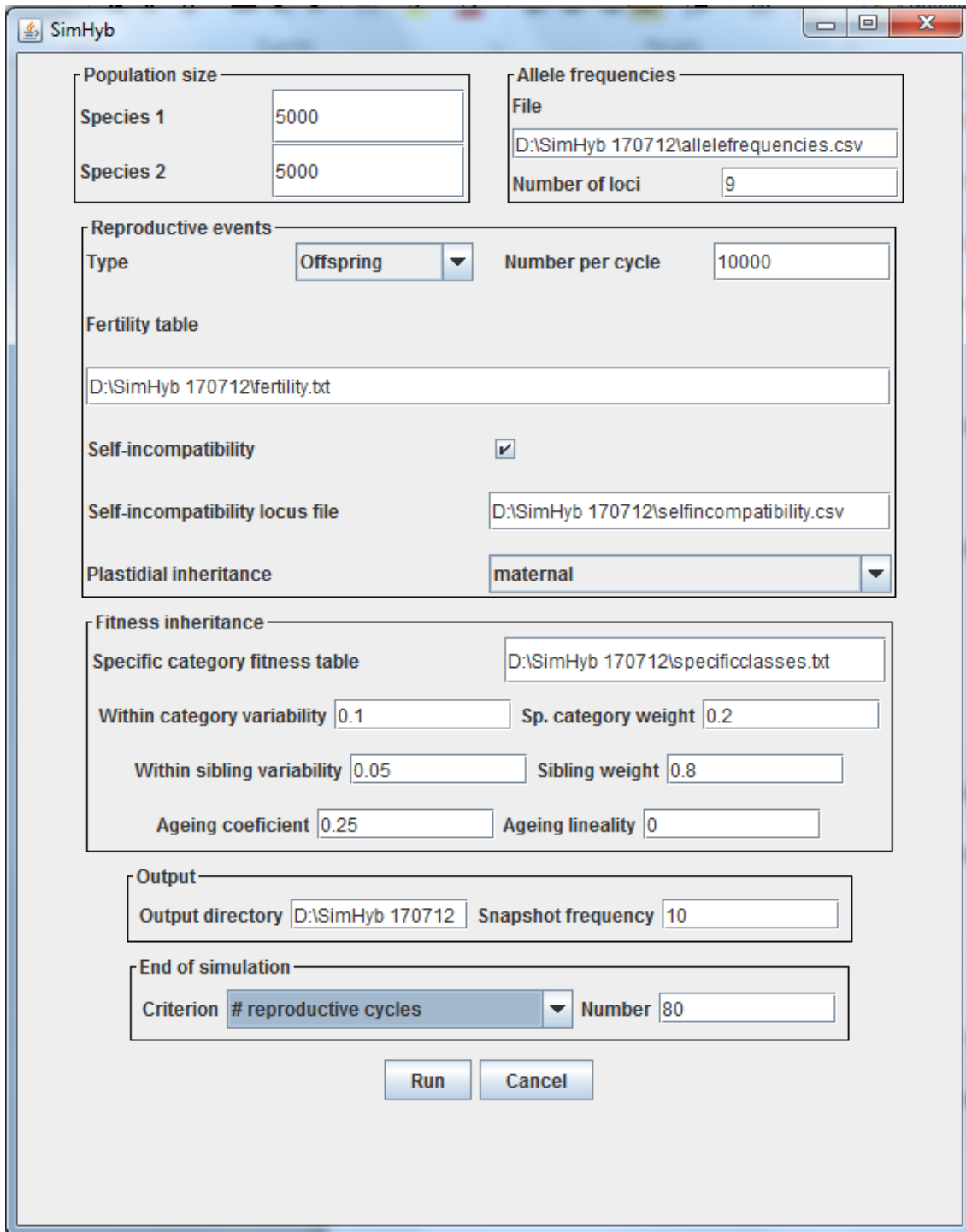
Figure 1. SimHyb flowchart. X represents the number of reproductive cycles and n is the number of reproductive events (matings or offspring) per cycle.

INPUT FILES AND PARAMETERS

When SIMHYB.jar is executed, and before starting the simulation, an interface window appears where the user must define the simulation parameters and input files (Figure 2).
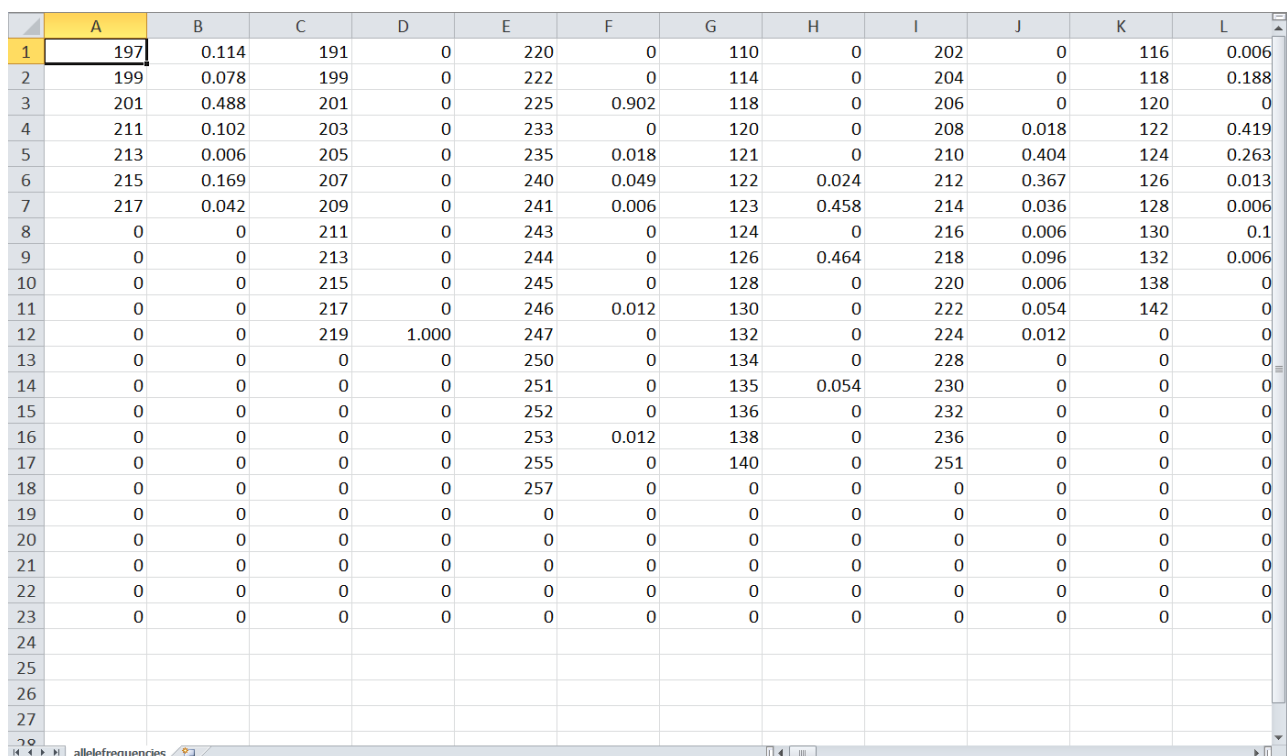

Figure 2. The SimHyb interface. When the simulation is launched, after selecting the input files and parameters, a pop-up window appears.

Population size

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.

Neutral loci

The user must also provide an `allelefrequencies.csv` file with the loci and the allele frequencies of each species. The format is exemplified in (Figure 2). The interface allows to consider just a subsample of the loci in the simulations, using the same input file.

	A	B	C	D	E	F	G	H	I	J	K	L
1	197	0.114	191	0	220	0	110	0	202	0	116	0.006
2	199	0.078	199	0	222	0	114	0	204	0	118	0.188
3	201	0.488	201	0	225	0.902	118	0	206	0	120	0
4	211	0.102	203	0	233	0	120	0	208	0.018	122	0.419
5	213	0.006	205	0	235	0.018	121	0	210	0.404	124	0.263
6	215	0.169	207	0	240	0.049	122	0.024	212	0.367	126	0.013
7	217	0.042	209	0	241	0.006	123	0.458	214	0.036	128	0.006
8	0	0	211	0	243	0	124	0	216	0.006	130	0.1
9	0	0	213	0	244	0	126	0.464	218	0.096	132	0.006
10	0	0	215	0	245	0	128	0	220	0.006	138	0
11	0	0	217	0	246	0.012	130	0	222	0.054	142	0
12	0	0	219	1.000	247	0	132	0	224	0.012	0	0
13	0	0	0	0	250	0	134	0	228	0	0	0
14	0	0	0	0	251	0	135	0.054	230	0	0	0
15	0	0	0	0	252	0	136	0	232	0	0	0
16	0	0	0	0	253	0.012	138	0	236	0	0	0
17	0	0	0	0	255	0	140	0	251	0	0	0
18	0	0	0	0	257	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0
24												
25												
26												
27												

allelefrequencies

Figure 2. An example of the allelefrequencies.csv file, opened in Windows Excel. Each loci is represented by two columns. The first one includes the allele names (sizes in bp in this example) and the second one its frequencies in the species. All the loci for the first species are presented in consecutive columns and then, in the same order, they appear for the other species. Please notice the columns do not have headings.

The user can define can define as many intermediate specific classes as desired, intermediate between the two pure species included in the original population, and can define the limits of each class, according to the so called "species coefficient", in a `specificclasses.txt` file. In this file, each row corresponds to a specific class, defined by a number, a name for the class, its boundaries in terms of the species coefficient, and the fitness coefficient corresponding to the specific class (f_{sp}). The route must be specified in the line Specific category fitness table. See an example in Figure 3.
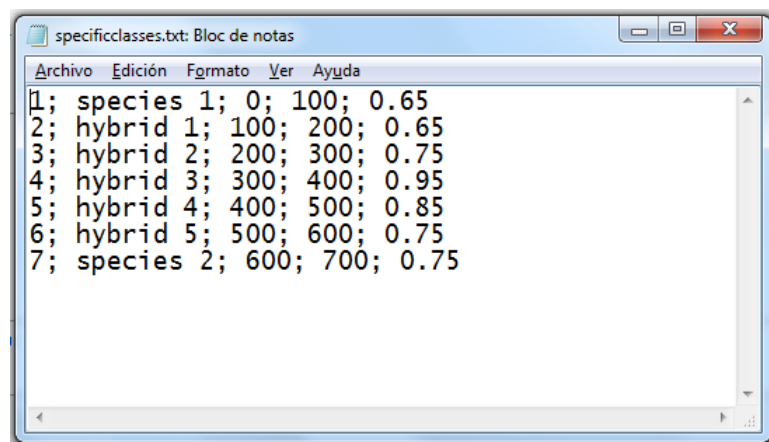
Figure 3. An example of the specificclasses.txt file, opened here in Windows Notepad. In each row, the first figure is the identification of the specific category, followed by its description; the next two numbers correspond to the category boundaries, according to the specific coefficient; tha last number is the fitness coefficient associated to the specific category. In this example, 7 categories are defined, two corresponding to the parental species (1 and 7) and with 5 intermediate, hybrid categories.

Fertility

Fertility within and between specific classes takes into account the pollination direction. Figure 4 shows an example of the `fertility.txt` file. Each row corresponds to a specific class, acting as mother tree, while each column corresponds to a specific class acting as pollen donor.

Chloroplast inheritance

Chloroplast can be inherited from the mother, as in most angiosperm species, or from the father, as in most gymnosperms.

Self-incompatibility

The user can select the option of considering self-incompatibility in the simulations. The current version of SIMHYB only takes into account gametophytic self-incompatibility. Figure 5 shows an example of the `selfincompatibility.csv` file.

End of simulation

The user can select to finish the simulation when one of the chloroplasts is completely replaced by the other one, when a specified number of specific categories remain in the population or after a desired number of cycles. When the simulation finishes, a pop-up window appears.

	A	B	C	D	E	F	G	H	I	J	K
1	#	1 (father)	2	3	4	5	6	7			
2	1	1	1	1	0.8	0.3	0	0			
3	2	1	1	1	1	0.8	0.5	0.3			
4	3	0.9	1	1	1	1	0.8	0.5			
5	4	0.9	1	1	1	1	1	0.8			
6	5	0.9	0.9	1	1	1	1	1			
7	6	0.8	0.9	1	1	1	1	1			
8	7	0.8	0.9	1	1	1	1	1			
9											
10											
11											
12											
13											
14											
15											
16											
17											
18											
19											
20											
21											
22											
23											
24											
25											
26											

fertility

Figure 4. An example of the fertility.txt file, opened here in Windows Excel. This example corresponds to 7 specific classes (two parental species, 1 and 7, and 5 intermediate, hybrid categories). In columns, pollen donors, and in rows, mother trees. Please notice asymmetric fertility is allowed.

10

Figure 5. An example of the selfincompatibility.csv file

OUTPUT FILES

SIMHYB provides as output a `population.csv` file quite similar to a Structure input file. The first row includes the name of the first ten columns. The second row includes the name of the loci included in the genotypes. The next rows correspond to all the individuals of the population throughout the simulation. All the generations (yield of each reproductive cycle) are included in the file. Each row represents an individual, including its Id number, its specific coefficient, the Id numbers of its parent trees, its chloroplast and mitochondria lineage (only two options for each organelle, one per parental species), the current generation (one individual can appear in several rows, corresponding to its evolution across the generation), its birth and death generation and its standardized fitness in this particular generation. The remaining cells correspond to its genotype, including the self-incompatibility locus and the neutral loci, with two cells per locus.

To run this output in Structure, the user just needs to remove the first row (and the ones corresponding to the non-desired individuals or generations) and to take into account the "additional" columns in Structure interface.

Additionally, SıмHʏв provides intermediate results, while the simulation is still running, which include "snapshots" of the population at the frequency defined in SıмHʏв interface.

Figure 6 shows an example of an output file.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Id	Sp. Coef.	Father	Mother	Chloropl	Mitochor	Generati	Birth Ger	Death Ge	Fitness						
2	Self-inco	locus 1	locus 2	locus 3	locus 4	locus 5	locus 6	locus 7	locus 8	locus 9						
306	272	700	132	185	B	B	3	2	4	0.0678	SI12	SI2	197	197	203	207
307	262	525	137	128	B	B	3	2	4	0.07	SI3	SI1	197	215	207	211
308	269	700	167	128	B	B	3	2	4	0.1177	SI2	SI13	197	197	207	205
309	286	350	165	162	B	B	3	2	4	0.1413	SI6	SI10	197	197	219	219
310	297	525	172	167	B	B	3	2	4	0.1476	SI9	SI12	197	201	211	219
311	247	700	134	176	B	B	3	2	4	0.1838	SI7	SI13	197	197	203	207
312	210	700	170	139	B	B	3	2	4	0.1853	SI3	SI12	197	197	205	207
313	242	700	169	128	B	B	3	2	4	0.2024	SI2	SI13	197	197	205	205
314	338	350	159	144	B	B	3	3	4	0.2029	SI5	SI3	215	211	211	203
315	248	700	170	104	B	B	3	2	4	0.2031	SI2	SI3	197	197	203	207
316	305	525	107	168	B	B	3	3	4	0.2065	SI2	SI12	197	213	209	219
317	395	350	173	144	B	B	3	3	4	0.2279	SI7	SI3	215	201	219	219
318	399	525	173	287	B	B	3	3	4	0.2286	SI5	SI1	197	197	215	207
319	352	437	150	233	B	B	3	3	4	0.24	SI8	SI9	197	197	219	219
320	396	525	173	243	B	B	3	3	4	0.2415	SI7	SI12	197	197	203	207
321	288	350	192	193	B	B	3	2	4	0.2496	SI11	SI5	197	197	219	219
322	315	437	236	129	B	B	3	3	4	0.2791	SI1	SI12	215	197	203	207
323	118	700	93	56	B	B	3	1	4	0.2825	SI2	SI5	197	197	191	211
324	386	612	262	147	B	B	3	3	4	0.2871	SI3	SI1	197	197	207	207
325	356	525	179	243	B	B	3	3	4	0.2883	SI2	SI1	197	197	211	219
326	187	700	58	71	B	B	3	1	4	0.2891	SI11	SI2	197	197	203	211
327	202	525	173	199	B	B	3	2	4	0.2892	SI7	SI3	197	197	207	219
328	320	612	236	199	B	B	3	3	4	0.2935	SI5	SI3	197	201	201	207
329	254	525	123	159	B	B	3	2	4	0.3003	SI2	SI5	211	197	219	199

Figure 6. An example of SimHyb output file.

HOW TO CITE THIS PROGRAM

Soto A, Rodríguez-Martínez D, López de Heredia U (2017) SıмHʏв: a simulation software for the study of the evolution of hybridizing populations. Application to *Quercus ilex* and *Q. suber* suggest hybridization could be underestimated. iForest XX: XXXX