SimHyb 2: Practical teaching evolutionary forces in Population Genetics using traceable virtual pedigrees

SimHyb 2: Enseñanza práctica de las fuerzas evolutivas en Genética de Poblaciones mediante pedigríes virtuales rastreables

Álvaro Soto[1], David Rodríguez-Martínez, Unai López de Heredia[1]
alvaro.soto.deviana@upm.es, rodriguezm.david@gmail.com, unai.lopezdeheredia@upm.es

[1]Departamento de Sistemas y Recursos Naturales.
Educational Innovation Group "Arboreto de Montes'
Universidad Politécnica de Madrid
Madrid, España

Abstract- Practical approaches have become a standard in many scientific disciplines, including Population Genetics. By analyzing appropriate datasets the students can draw conclusions about the evolution of populations with higher efficiency than if based exclusively on theoretical lessons. However, preparing appropriate datasets is a hard task and a wrong selection can spoil a well-aimed practice. Here we present SimHyb 2, a software intended for application in Population Genetics teaching. It simulates the course of a mixed population under user-defined reproductive and evolutionary conditions. It is very suitable for project-based-learning (PBL) approaches: the students obtain directly their own datasets in different evolutionary scenarios and the outputs can be easily adapted for downstream analysis with other popular tools as GenAlEx or Structure. SimHyb 2 is the only program available providing traceable individual pedigrees, so it is very appropriate for preparing datasets for parentage analysis, spatial genetic structure or conservation genetics study cases.

Keywords: Population Genetics, Project-Based Learning, Simulation Software

Resumen- Los enfoques prácticos son un estándar en muchas disciplinas científicas, incluida la Genética de Poblaciones. Analizando datos apropiados, los alumnos pueden extraer conclusiones sobre la evolución de las poblaciones más eficazmente que basándose exclusivamente en lecciones teóricas. Desafortunadamente, preparar estos datos es arduo y un error puede arruinar una práctica bien intencionada. Presentamos SimHyb 2, un software diseñado para la enseñanza de Genética de Poblaciones. Simula el desarrollo de una población mixta en condiciones reproductivas y evolutivas definidas por el usuario. Es muy adecuado para el Aprendizaje Basado en Proyectos: los estudiantes obtienen directamente sus propios conjuntos de datos en diferentes escenarios y los resultados pueden adaptarse fácilmente para su análisis posterior con otras herramientas populares como GenAlEx o Structure. SimHyb 2 es el único programa disponible que proporciona pedigríes trazables, siendo muy apropiado para preparar casos para análisis de parentesco, estructura genética espacial o genética de la conservación.

Palabras clave: Genética de Poblaciones, Aprendizaje Basado en Proyectos, Software de Simulación

1. Introduction

As many other disciplines, teaching population genetics has noticeably benefited in the last decades from computational power and the development of specialized software tools. Since the late 1990's and 2000's, a large amount of *ad hoc* programs were developed by scientists to calculate diversity and differentiation parameters; for spatial genetic structure analysis or for parentage analysis. Nevertheless, non-user-friendly input requirements often hampered their wide application in the classroom. In recent years, ready-to-use applications have been developed implementing many of these analysis in a very accessible way, as the MS Excel Add-in GenAlEx (Peakall and Smouse, 2012). Notwithstanding, providing the adequate datasets for practical exercises, illustrating the desired features, is always a dull and unappreciated task for the teacher. This is particularly true for population genetics dynamics.

In this work, we present SimHyb 2, a software to simulate the evolution of hybridizing populations (or species) under user-defined conditions. SimHyb 2 constitutes a useful tool to explore and illustrate the effect of the main evolutionary forces (migration, drift, selection, reproductive success). Contratrily to other software, such as HybridLab (Nielsen et al., 2006), which produces genotypes only based on allele frequencies, SimHyb 2 works with virtual *individuals*, allowing them to mate under user-defined conditions. It is the only software, to our knowledge, providing a thoroughly traceable pedigree for each virtual individual. SimHyb 2 yields outputs easily transferable for analysis with other popular software such as GenAlEx, or Structure.

A. Educational objectives

SimHyb 2 helps students understand evolutionary forces, and is particularly appropriate for Project Based Learning (PBL). The students can produce their own genotype datasets with SimHyb 2, to be directly used in further analyses in the project. Due to the stochastic procedures included in different phases of the simulation process, outputs will never be

DOI: 10.26754/CINAIC.2023.0048

identical (in terms of genotypes, pedigrees, survival…), which can be also a desirable feature from the teacher's point of view.

2. Context & description

SimHyb 2 is intended for practical teaching of Population Genetics, at the undergraduate or master level. It is programmed in Java 16, 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, user manual and input examples are available at https://github.com/GGFHF/SimHyb2

A. Overview

SimHyb 2 simulates the evolution of a population of constant size with up to two different diploid genetic groups (we will refer to them as "species", for simplicity), which may hybridize, depending on the conditions defined by the user. The population consists of a fixed number of individuals, each of them identified by a vector including pedigree, a so-called "specific coefficient" (representing the expected contribution of each species to the individual genome), the complete genotype for a user-defined number of loci and other information. SimHyb 2 was initially intended for vascular plants as model species, so it considers both chloroplast and mitochondria.

In each cycle, in order to produce offspring, pairs of individuals are selected randomly for mating. Individuals are considered hermaphrodite, so they can act both as mother and as father. If the parent pair passes the species-specific fertility and the self-incompatibility filters, a viable offspring individual is produced, and its genotype is defined, drafting alleles from the parents. A fitness value is also assigned to the individual, based on its specific composition and on the particular parental values. According to user specifications, immigrant individuals can also incorporate to the population at this stage.

Then, individuals from previous cycles are aged, reducing their fitness values. After ageing, individuals with the lowest fitness are removed, so that the population remains constant. Finally, before initializing the next cycle, fitness values of survivors are standardized between 0 and 1. Overview of SIMHYB 2 functioning is depicted in Figure 1.

3. Results

One of the main applications of SimHyb 2 is the illustration of the effects of evolutionary forces on the genetic pool and allelic frequencies of a population, and the generation of appropriate datasets for academic exercises and further analysis with other popular software tools such as GenAlEx or Structure (Pritchard et al., 2000). Other commonly used software, such as HybridLab (Nielsen et al., 2006), only allows the production of virtual F1 or F2 hybrid individuals. It drafts alleles according to their frequencies in the parental populations (species) to complete the genotype of hybrid individuals, and requires the calculation of frequencies in the hybrid output to be included as an input in a following cycle if further introgression levels are desired. On the contrary, SimHyb 2 uses the original allele frequencies only in the construction of the first generation of the population, and later on, those individuals reproduce among them, according to user-defined rules. In doing so, SimHyb 2 provides traceable pedigrees of individuals with known, different introgression levels. Therefore, SimHyb 2 outputs provide the genotypes and information to prepare datasets for exercises on parentage analysis, on spatial genetic structure, or for Conservation Genetics practical cases. SimHyb 2 can also be used for research purposes, as did the previous version, to assess hybridization and introgression processes (López de Heredia et al., 2018, 2020) or to check suitability of markers for different purposes (Cosín-Roldán *et al.*, 2023).

Output files are provided as *.csv* files. The first row register the headings of the first 10 columns, which include different information of each individual (see below). The second row includes the name of the nuclear, diploid loci, starting with the SI locus and followed by the neutral loci. Third and successive rows include the virtual individuals. The first 10 positions include the following information: 1) Individual ID (integer), 2) Specific coefficient (numeric character), 3) Father individual ID, 4) Mother individual ID, 5) Chloroplast (A or B), 6) Mitochondria (A or B), 7) Generation (integer; cycle of the simulation), 8) Birth Generation (integer; cycle in which the individual is added to the population), 9) Death Generation (integer; cycle in which the individual is removed from population; "-1" for survivors), 10) individual fitness value in that generation (number, between 0 and 1). The next positions register the diploid genotype of the individual, starting with the self-incompatibility locus (alphanumeric characters, two alleles) and neutral loci (alphanumeric characters, two alleles per locus).

A. Implementation and results achieved

SimHyb 2 has been successfully applied to teaching the subjects "Conservation and Breeding of Forest Genetic Resources" (Master in Forest Engineering) and "Forest Genetics" (Degree in Forest Engineering) at the Universidad Politécnica de Madrid during the courses 2021-2022 and 2022-2023 (with success rates above 84%), as well as in a pilot experience in the subject "Evolution" (Degree in Biological Sciences) at the Universidad de Córdoba during the course 2022-2023.

4. Conclusions

SimHyb 2 is a useful tool to illustrate the effect of evolutionary driving forces in the dynamics and allele frequencies of populations, specially suitable for project-based-learning approaches. SimHyb 2 outputs can be easily adapted for analysis with other commonly used population genetics tools. Since it provides traceable pedigrees and actual values of expected introgression for virtual individuals, it can be used both in teaching and in research, for parentage analysis or to explore hybridization processes or test suitability of a marker set for these purposes.

SimHyb 2 was initially intended for long-living plants, with over-lapping generations, user can set a large ageing coefficient in order to get "annual" (or, at least, non-overlapping) generations. Application for animal species is more difficult in its current version, not only due to the inclusion of chloroplast (which could be simply overlooked), but mainly because individuals are considered hermaphrodite. We aim to incorporate the appropriate modification for dioecious species in a following version.

ACKNOWLEDGEMENTS

This research was funded by the Universidad Politécnica de Madrid (grant number IE23-1303) and the Spanish Ministry of Science and Innovation (grant 288 number PID2019-110330GB-C22).

REFERENCES

Cosín-Roldán, J., López de Heredia, U. and Soto, Á. (2023). Highly informative SNP markers for routine individual identification and certification of forest reproductive material in *Q. ilex* and *Q. suber* and their hybrids. *Forest Ecology and Management* 544: 121243 DOI: 10.1016/j.foreco.2023.121243

López de Heredia, U., Sánchez, H. and Soto, Á. (2018). Molecular evidence of bidirectional introgression between *Quercus suber* and *Quercus ilex*. *iForest - Biogeosciences and Forestry* 11: 338-343. DOI: 10.3832/ifor2570-011

López de Heredia, U., Mora-Márquez, F., Goicochea, P.G., Guillardín-Calvo, L., Simeone, M.C. and Soto, Á. (2020). ddRAD sequencing-based identification of genomic boundaries and permeability in *Quercus ilex* and *Q. suber* hybrids. *Frontiers in Plant Science* 11: 564414. DOI: 10.3389/fpls.2020.564414

Nielsen, E.E.G., Bach, L.A. and Kotlich, 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

Peakall, R. and Smouse, P.E. (2012). GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research. Bioinformatics 28: 2537-2539. DOI: 10.1093/bioinformatics/bts460

Pritchard, J.K., Stephens, M. and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155(2): 945-959. DOI: 10.1093/genetics/155.2.945

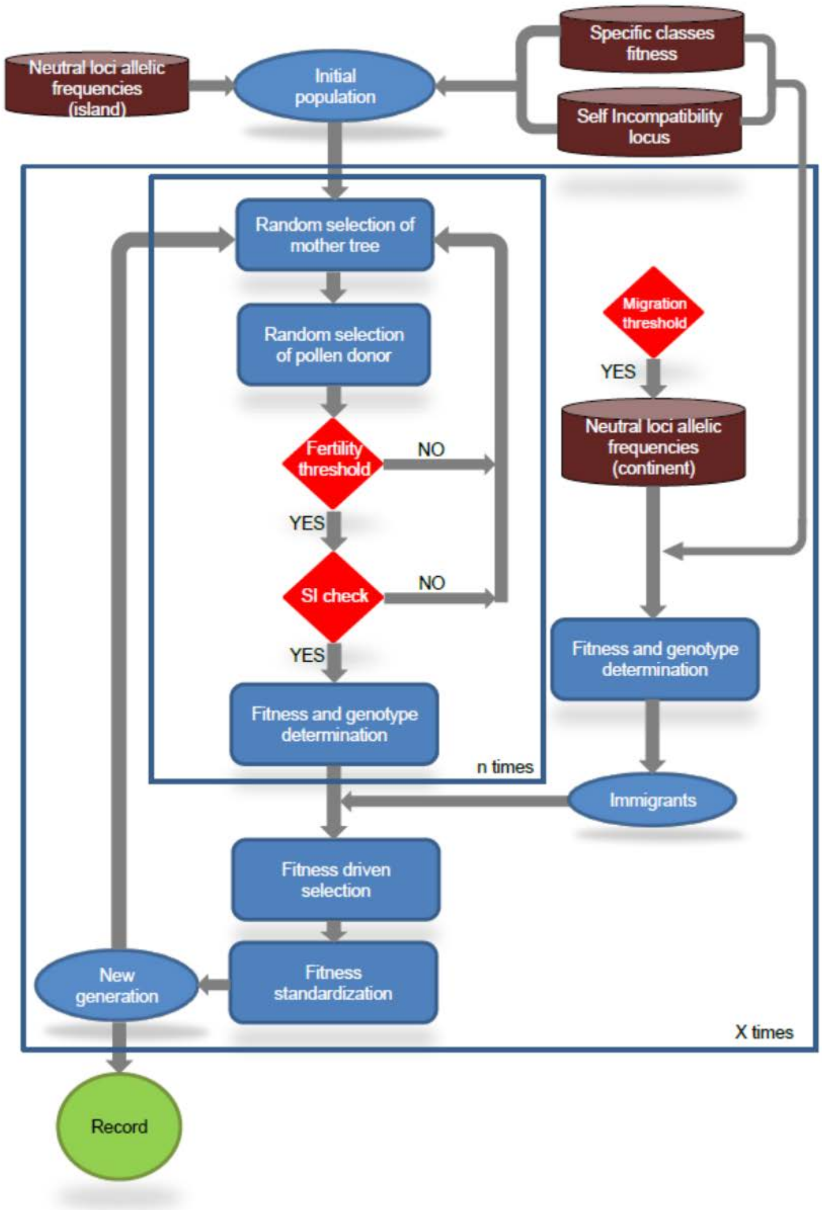

Figure 1. Overview of SIMHYB 2 functioning.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Id	Sp. Coef.	Father	Mother	Chloropla	Mitochon	Generatic	Birth Gen	Death Ge	Fitness					
2	Self-incoı	locus 1	locus 2	locus 3	locus 4	locus 5	locus 6	locus 7	locus 8						
1151	830	700	398	516	B	B	5	5	16	0.6251	SI12	SI3	211	201	215
1152	819	700	410	361	B	B	5	5	16	0.6291	SI5	SI5	211	211	215
1153	755	350	52	393	B	B	5	4	17	0.63	SI8	SI1	201	211	213
1154	752	350	533	239	A	A	5	4	17	0.6302	SI3	SI12	201	201	219
1155	770	350	156	417	B	B	5	4	17	0.6328	SI2	SI12	201	199	213
1156	836	700	412	495	B	B	5	5	16	0.6433	SI3	SI3	211	201	215
1157	841	700	352	591	B	B	5	5	16	0.6443	SI3	SI1	201	211	219
1158	757	350	468	47	A	A	5	4	17	0.6446	SI8	SI2	211	211	217
1159	798	350	739	568	B	B	5	4	17	0.6454	SI1	SI5	199	211	213
1160	784	350	37	512	B	B	5	4	17	0.6493	SI1	SI5	211	197	215
1161	792	525	305	727	A	A	5	4	17	0.6517	SI11	SI7	201	211	217
1162	820	700	354	537	B	B	5	5	16	0.6524	SI5	SI2	211	199	215
1163	786	350	51	397	B	B	5	4	17	0.6544	SI11	SI2	211	201	215
1164	818	700	653	551	B	B	5	5	16	0.6547	SI7	SI9	201	201	213
1165	753	350	268	349	B	B	5	4	17	0.655	SI9	SI2	201	211	219
1166	778	350	374	273	A	A	5	4	17	0.6553	SI11	SI1	197	211	217
1167	834	700	325	329	B	B	5	5	16	0.6561	SI3	SI1	211	201	215
1168	797	350	329	287	A	A	5	4	17	0.6623	SI12	SI6	197	211	217
1169	808	700	398	329	B	B	5	5	16	0.6629	SI2	SI12	211	201	215
1170	793	350	401	226	A	A	5	4	17	0.6631	SI2	SI5	199	201	219
1171	763	350	545	35	A	A	5	4	17	0.6634	SI13	SI3	197	211	219
1172	837	350	266	365	B	B	5	5	17	0.8972	SI2	SI8	201	201	213
1173	811	350	546	18	A	A	5	5	17	0.9001	SI1	SI12	197	211	217
1174	838	525	422	742	B	B	5	5	17	0.9018	SI3	SI1	211	201	219
1175	849	350	329	173	A	A	5	5	17	0.9029	SI1	SI1	201	211	217
1176	842	350	348	18	A	A	5	5	17	0.9032	SI5	SI12	197	201	219
1177	824	525	398	624	B	B	5	5	17	0.9086	SI2	SI7	211	201	215
1178	823	306	159	754	B	B	5	5	17	0.9174	SI3	SI6	211	211	217
1179	816	350	213	334	B	B	5	5	17	0.9213	SI1	SI5	201	201	213
1180	847	350	643	740	A	A	5	5	17	0.9228	SI9	SI13	201	199	213
1181	846	350	554	280	A	A	5	5	17	0.9238	SI6	SI3	197	197	219
1182	800	350	59	594	B	B	5	5	17	0.9247	SI1	SI5	201	201	213
1183	825	350	710	638	A	A	5	5	17	0.9258	SI2	SI5	201	197	217
1184	843	350	165	429	B	B	5	5	17	0.9373	SI3	SI2	201	197	215
1185	804	350	67	449	B	B	5	5	17	0.938	SI2	SI8	201	201	215
1186	815	350	285	400	B	B	5	5	17	0.9415	SI1	SI13	201	201	215
1187	810	350	204	496	B	B	5	5	18	0.944	SI5	SI8	211	197	217
1188	809	525	714	571	B	B	5	5	18	0.9458	SI3	SI7	211	201	213
1189	802	525	434	730	A	A	5	5	18	0.9479	SI13	SI12	199	201	217
1190	840	350	709	212	A	A	5	5	18	0.9484	SI2	SI5	197	211	215

Figure 2. Output of a SIMHYB 2 simulation, including individual genotypes.