GENETHAPLO: a java program to analyse the genome-wide multilocus genetic structure of predominantly selfing or clonal populations

Multilocus genotypes provide valuable information about mating systems (Jullien et al. 2019). Four software packages were previously developed to identify individuals originating from clonal reproduction using their multilocus genotype: MLGSIM (Stenberg et al. 2003); GENOTYPE and GENODIVE (Meirmans and Van Tienderen 2004), GENECLONE (Arnaud-Haond and Belkhir 2007) and poppr (Kamvar et al. 2014). Yet, none of these programs is specifically designed to identify individuals reproducing by selfing, in particular to detect repeated multilocus genotypes within a population and through time (or space) and recognize potential recombinants, formed by rare outcrossing events.

GENETHAPLO is a program written in Java with four modules:

1. A module to convert the format of a dataset
2. A module to filter the dataset
3. A module to analyse the genetic diversity
4. A module to analyse the multilocus genetic structure

*Formatting the data file*

The first line of the data file is a header line describing the content of each column, i.e. the name of the population, of the sub-population, of the individual and of each locus. Each following line provides the genotype of an individual at the specified loci. The individuals should be sorted so that populations and sub-populations are grouped together in consecutive lines.

Example:

temp,pop,Individu,ENPB1,MTIC59L,MTIC37C,MTIC126,FMTBN,MTIC243,MTIC40,MTIC86,MTIC252,TPG85C,TP36B,B12F1,TPC56E,TPC95G,TPC63A

pop,1987,F20089-1987-001,278278,110110,86086,99099,198198,118118,128128,157157,142142,133133,196196,98098,172172,88088,190190

pop,1987,F20089-1987-003,278278,110110,86086,99099,198198,118118,128128,157157,142142,133133,196196,98098,172172,88088,190190

pop,1987,F20089-1987-004,278278,110110,86086,99099,198198,118118,128128,149149,142142,133133,198198,98098,172172,88088,190190

…

pop,2009,F20089-2009-006,274280,97110,86086,99099,166166,118118,134134,126126,145145,131131,188188,83083,182182,136136,192192

pop,2009,F20089-2009-007,278278,110110,86086,99099,198198,118118,128128,149149,142142,133133,198198,98098,172172,88088,190190

pop,2009,F20089-2009-008,274280,97097,95095,99099,188188,118118,128128,155155,145145,141141,188188,86086,166166,86086,192192

*Module 1: format conversion*

This module takes a dataset in the read2snp (Uricaru et al. 2014) format and converts it to a format suitable for GenetHaplo, as detailed above.

*Module 2: data filter*

This module allows to filter out the loci, and individuals, having a percentage of missing data exceeding a specific threshold (given by the user). The two output files are i) a reduced dataset and ii) the list of the loci and individuals that have been removed. The percentage of missing data before and after filtering is also provided.

*Module 3: genetic diversity*

This module computes the key descriptors of genetic diversity classically used in population genetics studies. A first table summarizes the average number of individuals, alleles, the expected and observed heterozygosity and the FIS. The selfing rate is also calculated from the FIS for each sub-population. The module also provides these descriptors of diversity per locus and a table of allele frequencies for each sub-population.

*Module 4: multilocus genetic structure*

This module comprises three steps:

1. Grouping individuals according to their multilocus genotypes (thereafter called MLG): This module is based on a graph algorithm, where each node is an individual and nodes are connected when the individuals share the same MLG. An error rate can be specified by the user to allow grouping MLGs that differ at less than a given proportion of loci. This avoids over-splitting the MLGs due to genotyping errors or recent mutations. The module also takes into account missing data that can generate uncertainties. In case of missing data, it is possible for an individual to have a genotype compatible with several MLG. In such a case, the individual is randomly assigned to one of the possible MLG groups based on a random draw where each MLG group has a probability of being chosen that is proportional to its size. The output files provide i) the list of all individuals with the MLG to which they are assigned ii) the list of all identified MLGs with their frequency in each sub-population, their residual heterozygosity, defined as the proportion of heterozygous loci out of the total number of loci without missing values, and the number of missing values in each MLG.
2. Estimating genetic distance between MLGs: The genetic distance between two MLGs is estimated as the number of alleles that differ between the two synthetic MLGs divided by the total number of alleles without missing data in these two MLGs. This module generates a distance matrix as well as a histogram depicting the pairwise distance distribution.
3. Identifying recombinant MLGs: This module uses the genetic distances to rapidly identify putative recombination events between MLGs. A MLG is a candidate recombinant between two other MLGs (thereafter called “parental MLGs”) if the sum of the allele differences between it and its two putative parents equals the number of allele differences between these two parental MLGs. Only the MLGs that are represented by at least two individuals can be considered as potential parents. The output file provides a list of potential families, with the details of pairwise genetic distances.

*Running the program*

This java program can be launched from a command prompt, in the folder where the modules are stored, using the command java -jar module.jar, where module.jar should be replaced by the corresponding module name.

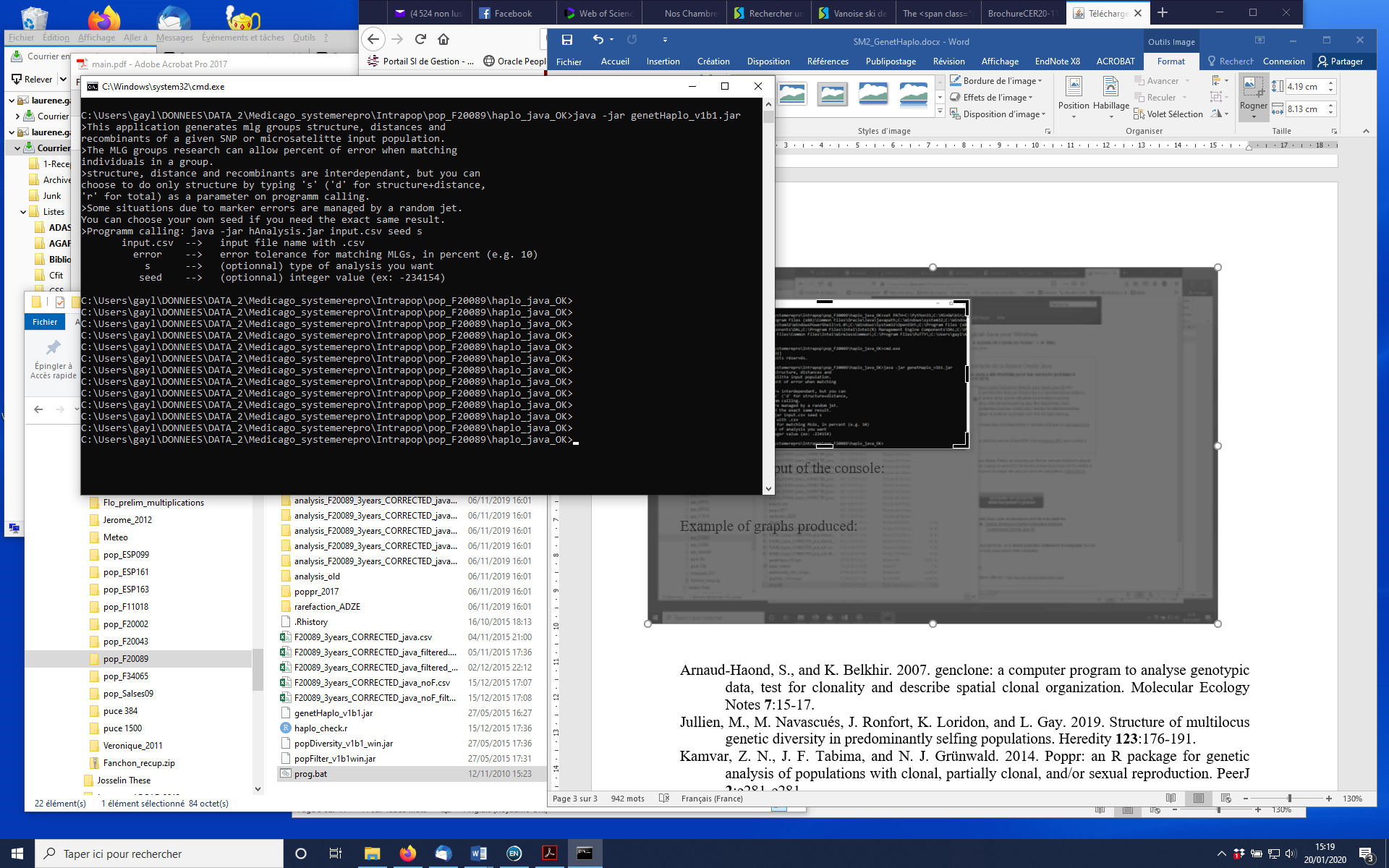
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Module | Function | Module name | Arguments | Example |
| Module 1 | format conversion | atgcTo12\_v1.jar | infile | java -jar atcgTo12 SNP\_EPO\_seuil\_160.csv |
| Module 2 | data filter | popFilter\_v1b1win.jar | infile  + threshold of missing data for individuals and loci | java -jar popFilter\_v1b1win.jar F20089.csv 25 35 |
| Module 3 | genetic diversity | popDiversity\_v1b1\_win.jar | infile | java -jar popDiversity\_v1b1\_win.jar F20089.csv |
| Module 4 | multilocus genetic structure | genetHaplo\_v1b1.jar | infile  + error rate  + type of analysis\*  + random seed | java -jar atcgTo12 F20089.csv 5 d 0 |

\* the type of analysis for the module 4 can be:

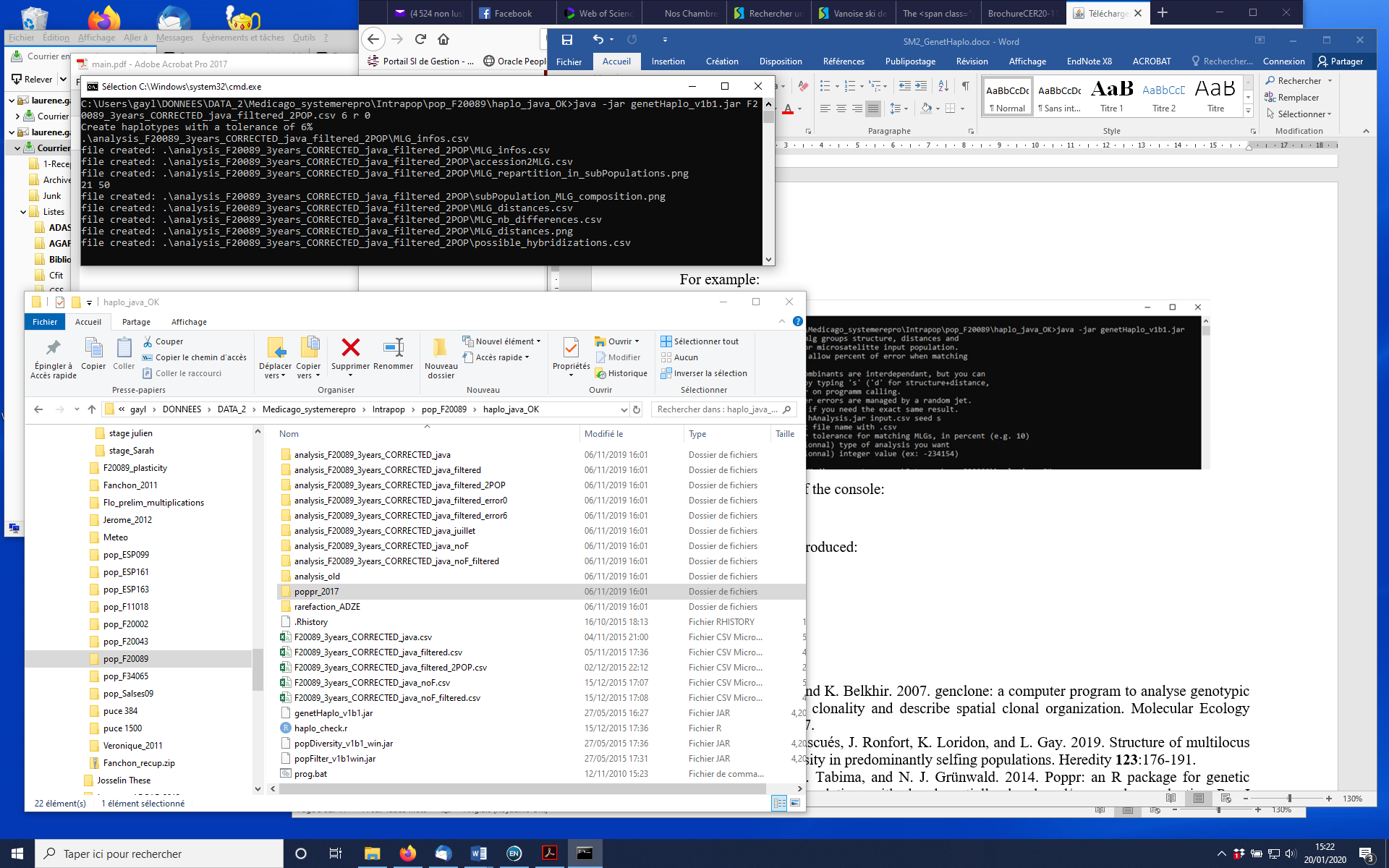
* only MLG groups (no argument)
* MLG groups and distances (d as an argument, as shown in the example)
* MLG groups + distance + potential recombinants (r as an argument, as shown in the example)

If no argument (infile or option) is added in the command, a short description of the script is displayed.

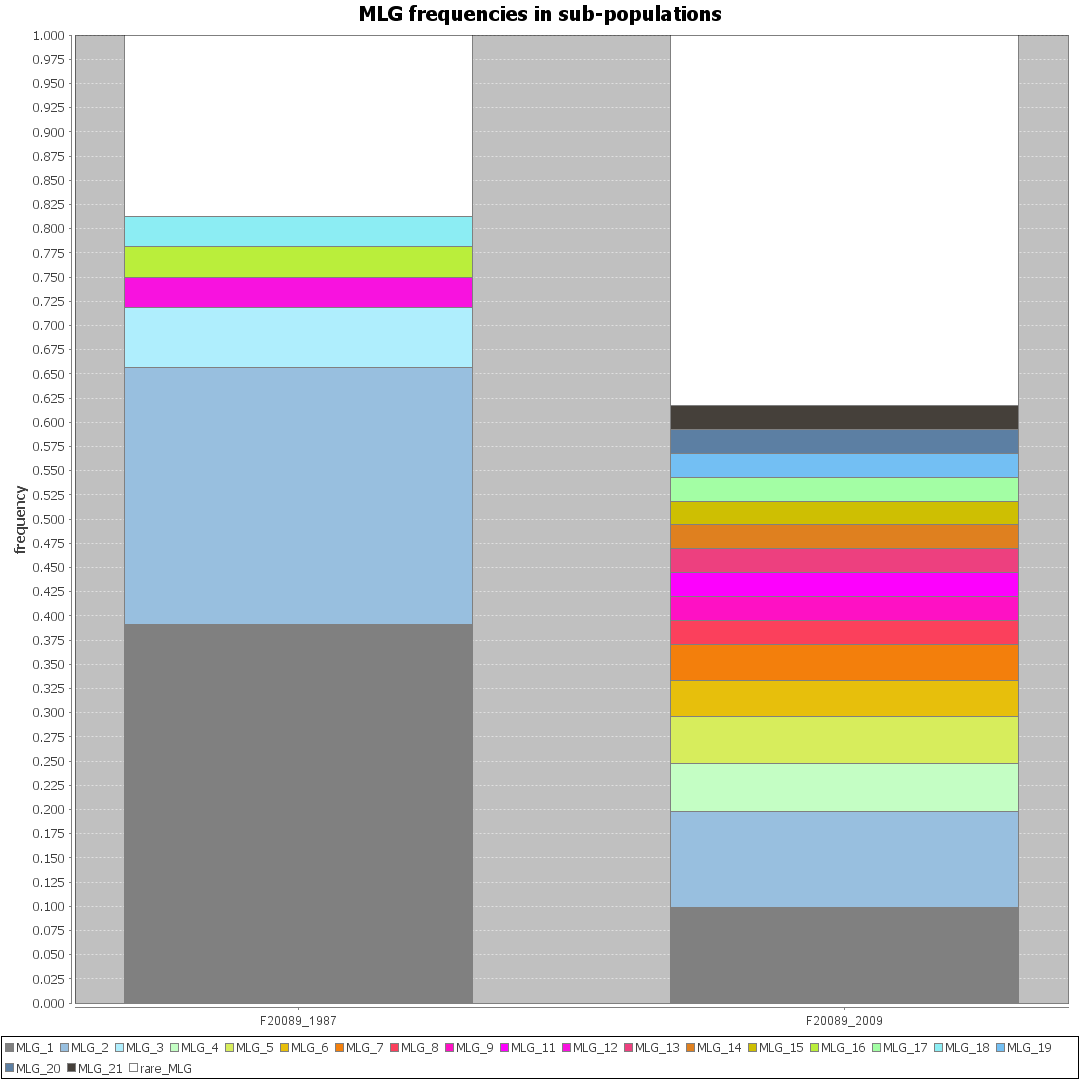
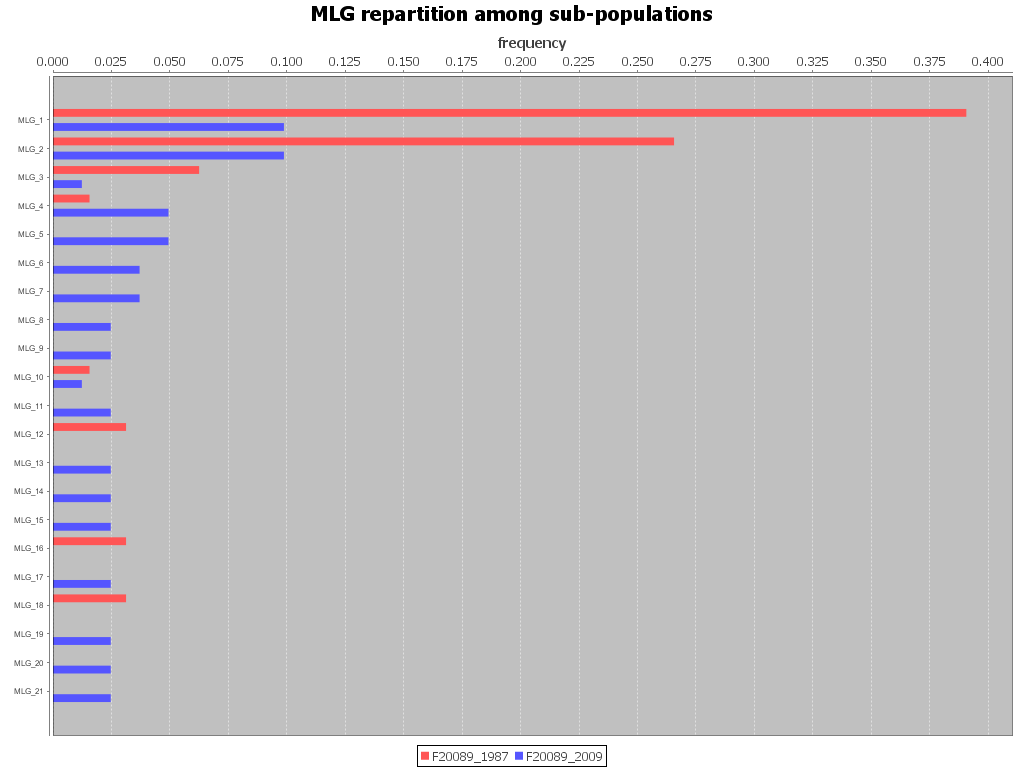
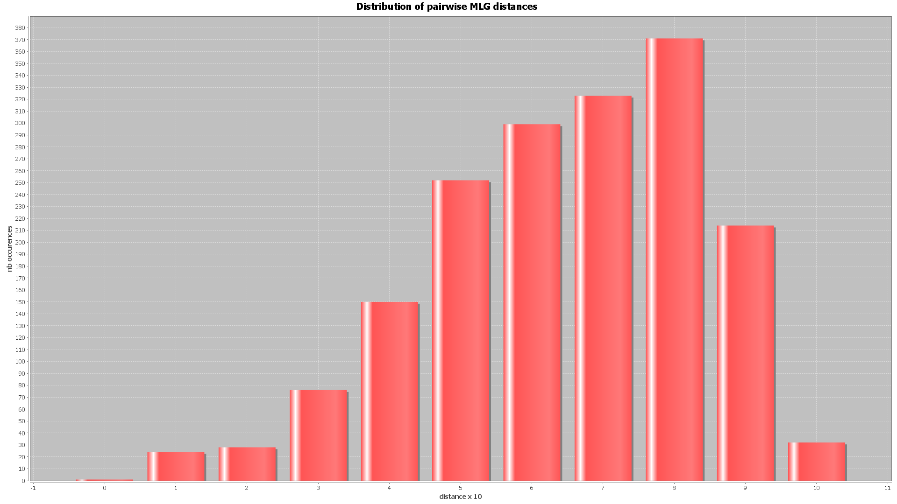
For example:



Example of output of the console:



Example of graphs produced:



Alcala, N., and N. A. Rosenberg. 2017. Mathematical Constraints on <em>F</em><sub>ST</sub>: Biallelic Markers in Arbitrarily Many Populations. Genetics **206**:1581-1600.

Arnaud-Haond, S., and K. Belkhir. 2007. genclone: a computer program to analyse genotypic data, test for clonality and describe spatial clonal organization. Molecular Ecology Notes **7**:15-17.

Edge, M. D., and N. A. Rosenberg. 2014. Upper bounds on FST in terms of the frequency of the most frequent allele and total homozygosity: The case of a specified number of alleles. Theoretical Population Biology **97**:20-34.

Frachon, L., C. Libourel, R. Villoutreix, S. Carrère, C. Glorieux, C. Huard-Chauveau, M. Navascués, L. Gay, R. Vitalis, E. Baron, L. Amsellem, O. Bouchez, M. Vidal, V. Le Corre, D. Roby, J. Bergelson, and F. Roux. 2017. Intermediate degrees of synergistic pleiotropy drive adaptive evolution in ecological time. Nature Ecology & Evolution **1**:1551-1561.

Golding, G. B., and C. Strobeck. 1980. Linkage disequilibrium in a finite population that is partially selfing. Genetics **94**:777-789.

Goudet, J. 2005. Hierfstat, a package for R to compute and test hierarchical F-statistics. Molecular Ecology Notes **5**:184-186.

Jakobsson, M., M. D. Edge, and N. A. Rosenberg. 2013. The Relationship Between <em>F</em><sub>ST</sub> and the Frequency of the Most Frequent Allele. Genetics **193**:515-528.

Jullien, M., M. Navascués, J. Ronfort, K. Loridon, and L. Gay. 2019. Structure of multilocus genetic diversity in predominantly selfing populations. Heredity **123**:176-191.

Kamvar, Z. N., J. F. Tabima, and N. J. Grünwald. 2014. Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. PeerJ **2**:e281-e281.

Keenan, K., P. McGinnity, T. F. Cross, W. W. Crozier, and P. A. Prodöhl. 2013. diveRsity: An R package for the estimation and exploration of population genetics parameters and their associated errors. Methods in Ecology and Evolution **4**:782-788.

Mehta, R. S., A. F. Feder, S. M. Boca, and N. A. Rosenberg. 2019. The relationship between haplotype-based *F*ST and haplotype length. Genetics **213**:281-295.

Meirmans, P. G., and P. H. Van Tienderen. 2004. genotype and genodive: two programs for the analysis of genetic diversity of asexual organisms. Molecular Ecology Notes **4**:792-794.

Navascués, M., A. Becheler, L. Gay, J. Ronfort, K. Loridon, and R. Vitalis. 2020. Power and limits of selection genome scans on temporal data from a selfing population. bioRxiv:2020.2005.2006.080895.

Nordborg, M. 2000. Linkage disequilibrium, gene trees and selfing: An ancestral recombination graph with partial self-fertilization. Genetics **154**:923-929.

Pollak, E. 1987. On the theory of partially inbreeding finite populations .1. Partial selfing. Genetics **117**:353-360.

Stenberg, P., M. Lundmark, and A. Saura. 2003. mlgsim: a program for detecting clones using a simulation approach. Molecular Ecology Notes **3**:329-331.

Uricaru, R., G. Rizk, V. Lacroix, E. Quillery, O. Plantard, R. Chikhi, C. Lemaitre, and P. Peterlongo. 2014. Reference-free detection of isolated SNPs. Nucleic Acids Research **43**:e11.

Weir, B. S., and C. C. Cockerham. 1984. Estimating F-statistics for the analysis of population sructure. Evolution **38**.

Wright, S. 1969. Evolution and the Genetics of Populations. Vol. II. The Theory of Gene Frequencies. University of Chicago Press, Chicago.