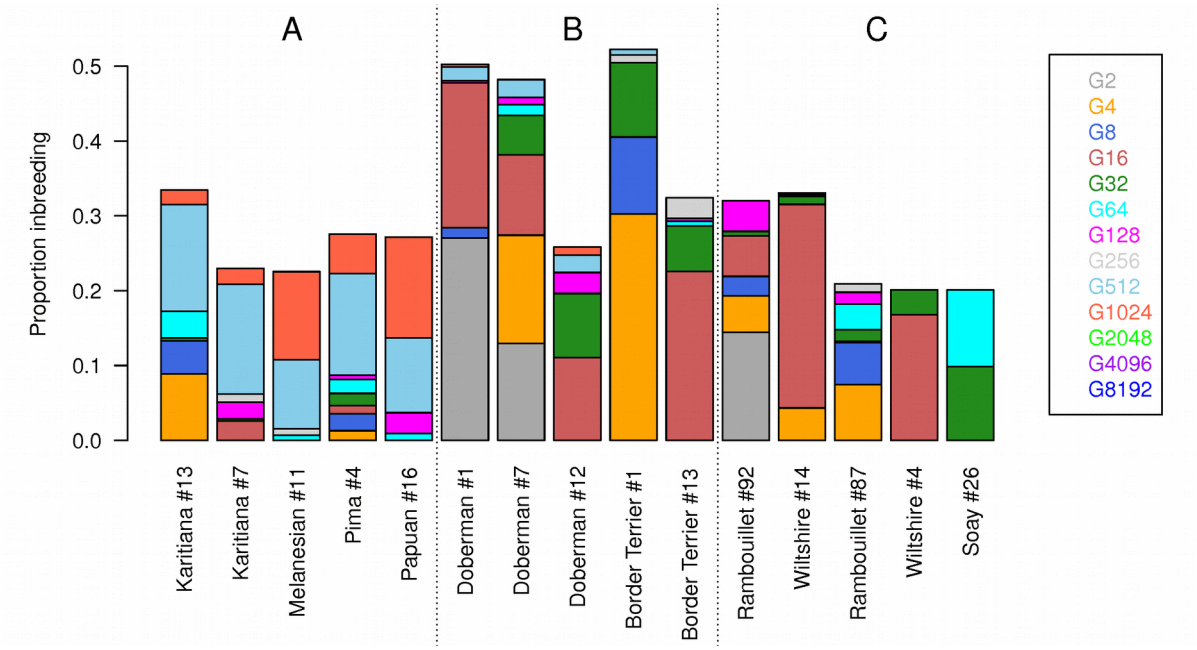
ZooRoH user's manual



September 2017

ZooRoH: a program for age-based partitioning of individual inbreeding using an exponential mixture model

Author: Tom DRUET (tom.druet@ulg.ac.be)

Introduction

Citation

If you use ZooRoH.f90 in a published analysis, please cite the following reference:

Druet T. and Gautier M. (2017). A model-based approach to characterize individual inbreeding at both global and local genomic scales. *Molecular Ecology* doi:10.1111/mec.14324

Compilation

To obtain an executable version of ZooRoH, the source code must be compiled with a Fortran compiler as for instance GFORTRAN or Intel Fortran (note that efficiency varies with compilers).

gfortran -ffree-line-length-0 ZooRoH.f90 -o ZooRoH
ifort ZooRoH.f90 -o ZooRoH (for Intel compiler)

Running ZooRoH

Input files

By default, ZooRoH reads a genotype file with a format similar to the "Oxford GEN" format with one line per marker and individuals in columns (GEN format). The five first columns (space or tab separated) contain information on the marker:

1) Chromosome number
2) SNP ID or Marker name (max 50 characters)
3) Marker physical position in bp
4) First marker allele (max 50 characters)
5) Second marker allele (max 50 characters)

Regarding the map position, we assume that 1 Mb = 1 cM to convert the physical map to genetic distances. If you have a genetic map, we recommend using it on the same scale (cM x 10^6).

After the first five columns, each value represents the genotype coded as the number of copies of

allele 1 carried by the individual (0 for homozygotes with the second allele, 1 for heterozygotes, 2 for homozygotes with the first allele and 9 for missing genotypes).

Alternatively, ZooRoH can read genotype probabilities (GP), genotype likelihoods on phred scale (GL) or read depth for both alleles (AD). In case of GP or GL, three values are given per individual (three columns) corresponding to genotype probabilities or phred likelihoods for genotypes 11 (homozygotes with allele 1), 12 (heterozygotes) and 22 (homozygotes with allele 2). With AD, two columns are expected per individual, allelic depth for allele 1 and allelic depth for allele 2.

Parameters file

Name of the genotype file and parameters are provided in the parameter file that must be called “param.txt”. Each parameter or option is defined with a precise key (the list of keys is provided here below) preceded by "#". When information is associated with the key, it must be provided on the next line. Some keys are optional and default values are used if the key is not mentioned.

#NUMBER_OF_CLASSES
Specify the number of HBD and non-HBD classes (one value K)

#RATE_PARAMETERS
Specify for each class the rate of the exponential distribution associated to each class. The rate is approximately equal to the size of the corresponding inbreeding loop (this is an approximation and represents only a qualitative measure). One line with K (the number of classes) values separated by space.

#HBD_INDICATORS
Determine which classes are HBD (1) and non-HBD (0). One line with K (the number of classes) values (either 1 for IBD and 0 for non-IBD) separated by space.

#STARTING_MIXING_PROPORTIONS
Give the starting mixing proportions for each class (should sum to 1.0). One line with K (the number of classes) values between 0 and 1.

#TRANSITION_MATRIX (optional, default = identity matrix)
A K x K matrix (K being the number of classes) determining which transitions are possible (1) or not (0). The matrix determines if transitions from class X to class Y are allowed. With this, the user can specify that all transitions are possible (identity matrix), that transitions from a state to the same state are not allowed (0's on the diagonal), that transitions from HBD-classes to other HBD-classes are not allowed, etc. By default, all transitions are allowed. K lines of K values (0 or 1) separated by space are expected.

#ERROR_RATE
Give the error rate associated with genotypes.

#ESTIMATE_RATES (optional, default = no)
This option is used when the user wants to estimate the rate associated with each defined class (by default the model uses the rates specified by the user). If the option is used, two values must be provided, the minimum rate (don't put values below 1) and the maximum rate (the value is a function of the marker density and informativity).

#ONE_RATE (optional, default = no)
No value expected. This option is used in case only one HBD class and one non-HBD class are

defined. In that case, with the ONE_RATE option, the same R is estimated for both classes (only for estimation purpose; for pre-defined R's, values provided by the user are used).

#NUMBER_OF_INDIVIDUALS
Specify the number of individuals in the file (one value expected)

#INPUT_FILE
Specific the name of the input file (one name expected, max 50 characters)

#ANALYSIS_RANGE / optional, default = 1 to max)
To run to model on a subset of individuals. The user gives the first and the last individual to analyse. Then the program analyzes all individuals between these bounds. Two values expected.

#MINMAF (optional, default = 0.d0))
Skip markers with a MAF lower than the threshold (one value expected)

#FREQUENCIES (optional, default = estimated)
By default, frequencies are estimated from the data set. The user can also provide the name of a file containing allele 1 frequencies. The file is a single column with the frequencies. One name expected (max 50 characters)

#ITERATIONS (optional, default = 1000)
Specify the number of iterations of the EM algorithm per individual.

#INPUT_FORMAT (optional, default = GEN)
Possible values: GEN GP GL AD. The formats are described in the input files section.

#OUTPUT (optional, default = 'no')
Possible values are ALL, SUM or IND. By default, HBD probabilities are not provided at each marker position. The user can require the HBD probability (obtained by summing HBD probabilities from all HBD classes) for each marker position per individual with the SUM option. This generate one file called SUMHBDp.txt with as many columns as individuals and as many lines as markers (the first three columns provide marker information). Similarly, the user can require a file for each HBD-class. Then the program generates several files with the same structure (as many as the number of HBD classes). The names of the files are HBDp_RX.txt with X being the number of the HBD class (one for the first HBD class, two for the second, etc).

The ALL and SUM options require to store all the HBDp (for all individuals and all markers). This increases memory requirements (and storage). With the IND format, one output file is created with K+4 columns per marker: the individual number, the chromosome number, the marker position, the local class number (the hidden state estimated at the marker by the Viterbi algorithm) and the K class probabilities (the K probabilities to belong to each of the hidden states estimated by the Forward-Backward algorithm). The file contains per individual as many lines as markers. Individuals are printed sequentially (first individual 1 for all markers, then individual 2, etc). There is no need to store all HBDp probabilities with that format because information is printed after each individual is processed.

Output files

The following information is printed on the screen for each individual after completing the desired number of iterations: number of the individual (position in the GEN file), log(likelihood), AIC, BIC, estimated mixing proportions (K values) and rates (K values). AIC and BIC are estimated

using the number of free parameters (P = 2*K-1 if both mixing proportions and rates are estimated, P = K-1 if only mixing proportions are estimated and P = 2 for a model with 2 classes with identical rate) and n records (where n is the number of markers).

HBDclasses_MeanProb.txt contains for each individual, the position of the individual in the input file, the estimated proportion of the genome in each HBD or non-HBD class (K values), the total HBD proportion (sum of HBD-classes) and the homozygosity.

HBDclasses_MixingCoef.txt contains for each individual, the position of the individual in the input file, the estimated mixing proportion for each HBD or non-HBD class (K values).

HBDclasses_Rates.txt contains for each individual, the position of the individual in the input file, the estimated rates for each HBD or non-HBD class (K values).

HBDclasses_CountsF.txt contains for each individual, the position of the individual in the input file, the estimated number of segments associated to HBD or non-HBD class (K values).

HBDsegments.txt contains a list of all HBD segments identified with the Viterbi algorithm. HBD segments are defined as continuous streches of markers assigned to any of the HBD classes. For each segment, the following information is provided: position of the individual in the genotype file, chromosome, number of first marker, number of last marker, position of the first marker, position of the last marker, length of the fragment, number of SNPs of the segment, number of heterozygous SNPs in the segment.

HBDsegments_per_class.txt contains a list of all HBD segments (defined per HBD class) identified with the Viterbi algorithm. Segments are defined as continuous streches of markers assigned to the same HBD class. For each segment, the following information is provided: position of the individual in the genotype file, chromosome, number of first marker, number of last marker, position of the first marker, position of the last marker, length of the fragment, number of SNPs of the segment, number of heterozygous SNPs in the segment, number of the HBDclass (the number of the hidden state).

SUMHBDp.txt (when the ‘SUM’ output option is used) contains one line per marker with the marker number, its position, the chromosome and then one column per individual (only those that have been analyzed if the #ANALYSIS_RANGE option was used) with its associated total HBD probability at the marker position.

HBDp_RX.txt (when the ‘ALL’ output option is used) is similar to SUMHBDp.txt but with the HBD probability associated to the X^{th} HBD-class.

LocalHBDp.txt is obtained when the ‘IND’output option is used. It contains NSNPs lines per individual (NSNPs x NIND in total where NSNPs is the number of markers and NIND the number of individuals). One line reports the individual number (position in the GEN file), the chromosome number, the marker position, the local class number (the hidden state estimated at the marker by the Viterbi algorithm) and the K class probabilities (the K probabilities to belong to each of the hidden states estimated by the Forward-Backward algorithm).

References

Druet T. and Gautier M. (2017). A model-based approach to characterize individual inbreeding at both global and local genomic scales. *Molecular Ecology* doi:10.1111/mec.14324.