CHROMIBD

Author: Tom DRUET (tom.druet@uliege.be)

Introduction

CHROMIBD models a set of “target chromosomes” as a mosaic of “reference chromosomes”. These chromosomes need to be previously phased and the program relies only on phased (non-missing) markers in target and reference chromosomes. The program works only with bi-allelic markers such as SNPs.

The initial implementation works within a genealogy and the set of reference chromosomes are the “parental chromosomes” of the target chromosome. In that case, the program computes IBD probabilities between the target chromosome and a set of parental chromosomes (the method works best without "phantom parental chromosomes"). This method is described in:

T. Druet and F. Farnir (2011) Modeling of identity-by-descent processes along a chromosome between haplotypes and their genotyped ancestors. Genetics 188(2):409-19.

The version 1.2 allows also to work with unrelated individuals and without genealogy. In that case, the model is similar to IMPUTE (Marchini et al., 2007) and describes a target haplotype as a mosaic of a set of reference haplotypes. The method is described in:

P. Faux, P. Geurts and T. Druet (2019) An extra-trees framework for modeling haplotypes as mosaic of reference haplotypes (under review).

Citations

If you use the program in a published analysis, please cite the following publication:

T. Druet and F. Farnir (2011) Modeling of identity-by-descent processes along a chromosome between haplotypes and their genotyped ancestors. Genetics 188(2):409-19.

When using the --unrelated option, please also cite:

P. Faux, P. Geurts and T. Druet (2019) An extra-trees framework for modeling haplotypes as mosaic of reference haplotypes (under review).

Data format

Files are provided per chromosome (marker map and haplotyes). One run must be performed per chromosome. The format required for the input files:

1. **pedigree file**: animal sire dam. Numbers must be consecutive integers and elder individuals must have the lowest numbers.
2. **marker file**: one line per marker, marker number (consecutive number starting from 1), marker name (50 characters at maximum and no blanks within the name) and marker position (in cM or in Mb).
3. **haplotypes file**: one line per phase with individual number (6 positions), space, phase origin (1 for paternal and 2 for maternal), space and haplotype with two positions for each marker (no space between successive markers) – alleles are coded as integer (0 for missing or unphased markers). With the --unrelated option, the same format is used but target and reference haplotypes are internally recoded with their position in their respective files.
4. **allelic frequencies file**: marker number (the same as in marker file), frequency of allele 1, frequency of allele 2.

Running CHROMIBD

To run CHROMIBD, you need to type:

./CHROMIBD --method

with eventually additional arguments. The ‘method’ argument must come first and determines which method you use. To run CHROMIBD with the algorithm using the pedigree and the parental haplotypes as described in Druet and Farnir (2011) you need to use the --pedigree argument. To describe target chromosomes as mosaic of unrelated reference haplotypes (Faux et al., 2019), use the --unrelated argument.

Running CHROMIBD --pedigree

Running CHROMIBD with the --pedigree argument is equivalent to the initial implementation and uses the method described in Druet and Farnir (2011). If not further arguments are provided the program will asks interactively for the following elements (as initially):

Name of file with known haplotypes (can be empty or a non-existing file)
Name of pedigree file
Name of marker file
Name of the file with allelic frequencies
Model with or without inbreeding

Alternatively, the user can provide these elements with the command line with the following arguments:

CHROMIBD --pedigree --haplotypes name1 --ped name2 --map name3 --freqs name4 --inbreeding 1

where:

name1 is the name of the file with known haplotypes,
name2 is the name of the pedigree file,
name3 is the name of the marker file,
name4 is the name of the file with allelic frequencies,
and 1 or 0 after --inbreeding indicate respectively inbreeding is used or not.

Output file

IBD probabilities are stored in a file called "HiddenAncestors". Each line contains:

target_chromosome ID / parental chromosome ID / marker number / IBDprobabilities.

Chromosome ID are equal to twice the individual number minus one for the paternal chromosome and equal to twice the individual number for the maternal chromosome (e.g, chromosomes 1 & 2 for individual 1, chromosomes 3 & 4 for individual 2, …, chromosomes 99 & 100 for individual 50). For phantom chromosome, the ID "-1" is used.

The output file is large because it contains number of (target chromosomes) x (number of parental chromosomes per target chromosome) x (number of markers) lines. If size of the output file is a problem, the user can modify the program and use IBD probabilities for their application without creating an IBD output file. For instance, if the user is interested by imputation, he can add a subroutine to transform IBD probabilities in genotype probabilities.

Advice

The method computes IBD probabilities for each target chromosome independently. The results would be identical if instead of processing all target chromosomes in one run, the target chromosomes are each processed in an independent run. Therefore, IBD probabilities from different target chromosomes can be computed on different computers (in a cluster). In addition, this feature can also be used to avoid large output files. For instance, if the user is interested by imputation, he can run one target chromosome, use IBD probabilities to perform imputation and then erase the IBD probabilities output file from that chromosome.

Running CHROMIBD --unrelated

When running CHROMIBD with the --unrelated argument, the following arguments must be provided:

--refs followed by the name of the file containing the reference haplotypes,
--targets followed by the name of the file containing the target haplotypes,
--map followed by the name of marker file.

By default, the program runs both the forward-backward and the Viterbi algortihms. To run only the forward-back or the Viterbi algorithm use the --algortihm argument followed respectively by FB or VTB.

In addition, the user can provide two parameters:

--ngen followed by the number of generations determining in combination with the genetic distances between markers the frequency of 'recombination' between reference segments (4 by default).

--gerr followed by the probability to observe a difference between the target chromosome and the reference haplotype (0.001 by default).

Alternatively, the user can provide the value of Ne:

--Ne followed by a value for Ne. In that case, --ngen and --gerr are estimated as in IMPUTE (Marchini et al., 2007):

ngen = 4*Ne / N (with being the number of reference haplotypes) and gerr = θ / 2(θ + N) where θ is equal to:

$$\left(\sum_{1}^{N-1} \frac{1}{i}\right)^{-1}$$

Output file

With the --unrelated option, the target and reference haplotypes are internally recoded with their position in their respective files (e.g., 1 for the first haplotype from the target haplotype file, 3 for the third haplotype in the reference haplotype file, etc.). These numbers are used in the output files.

With the forward-backward algorithm, the following files are created:

- imputed.dose: the dosages obtained as the genotypes observed in the reference haplotypes weighted by the state probabilities from the forward-backward algorithm (the first column is the internal haplotype ID, the order in the target haplotype files);
- imputed.bestgeno: the file contains the genotypes that have the largest probability (the first column is the internal haplotype ID, the order in the target haplotype files);
- imputed.bestref: the file contains the genotypes observed on the reference with the highest probability (the first column is the internal haplotype ID, the order in the target haplotype files);
- mosaic.fb: the mosaic of reference haplotypes with the highest local state probability (first the internal target haplotype number, followed by the reference haplotypes number at each marker position).

With the Viterbi algorithm, two files are created:

- mosaic.vtb: the most likelily mosaic of reference haplotypes obtained with the Viterbi algortihm (first the internal target haplotype number, followed by the reference haplotypes number at each marker position);
- imputed.vtb: the genotypes observed on the most likely mosaic from the Viterbi algortihm (the first column is the internal haplotype ID, the order in the target haplotype files).