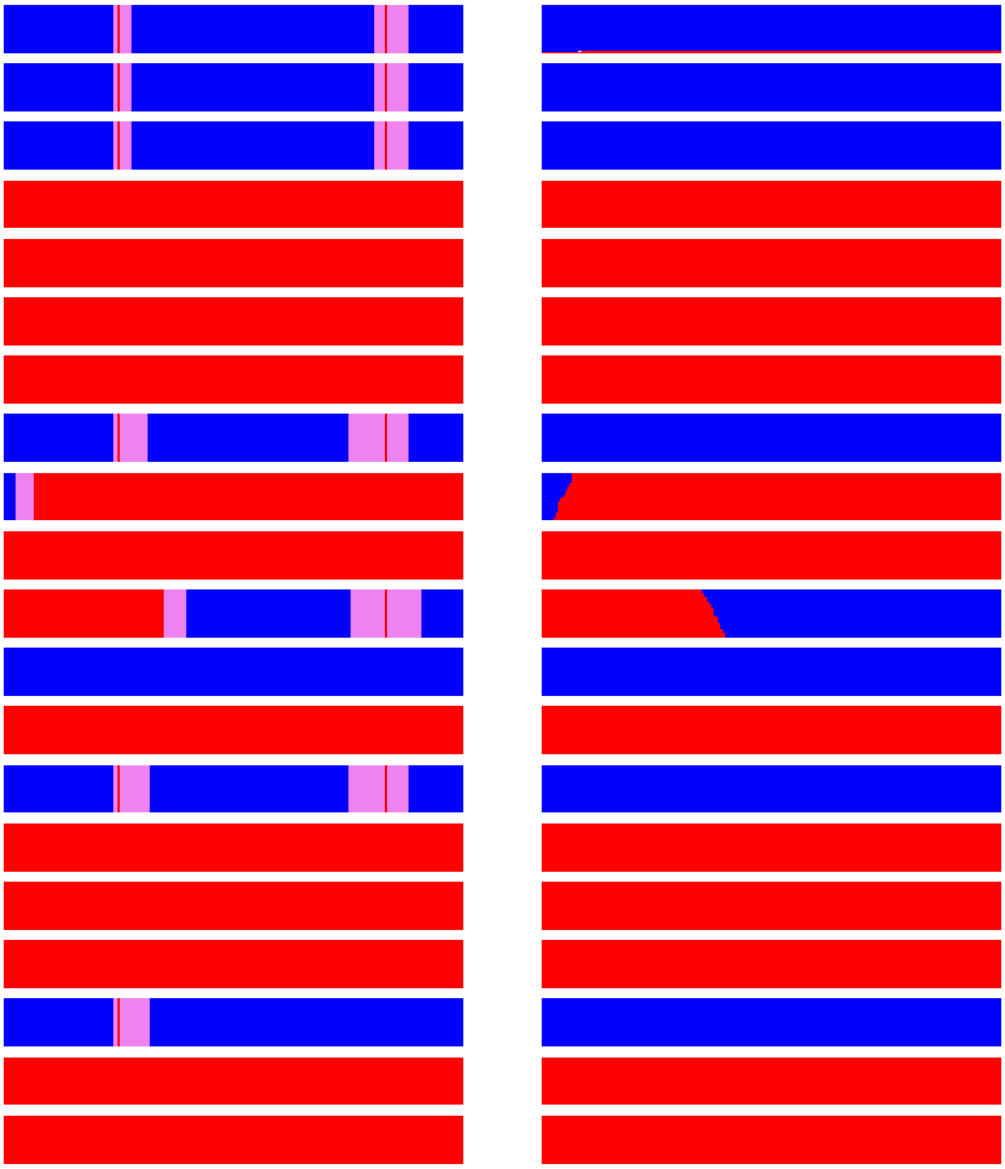
LINKPHASE3 user's manual



LINKPHASE3: a program for haplotype reconstruction using pedigree information in the PHASEBOOK package

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

Introduction

LINKPHASE3 has similar properties than LinkPHASE but this new version is more robust to genotyping and map errors and is also faster. It uses familial information (genotyped parents or several genotyped offspring) to reconstruct haplotypes and works particularly well for large half-sib families. Haplotype reconstruction is performed only when strong information is available and markers can remain unphased. As LinkPHASE, LINKPHASE3 doesn't use linkage disequilibrium information. LINKPHASE3 is designed for bi-allelic markers. Genotype cleaning is adviced prior to run the program.

LINKPHASE3 also identifies cross-overs and determines the haplotype origins (which parental haplotypes are inherited). Specific rules are available to phase sex-chromosomes and it is possible to provide sex-specific maps as input. The program has also options to estimate the length of the genetic maps (per sex). For data set with more markers than individuals (e.g. whole genome-sequencing), the program also accept formats with columns corresponding to individuals.

Citation

If you use LINKPHASE3.f90 in a published analysis, please cite the following publication:

T. Druet and M. Georges (2015) LINPHASE3: an improved pedigree-based phasing algorithm robust to genotyping and map errors. Bioinformatics doi:10.1093/bioinformatics/btu859.

Compilation

To obtain an executable version of LINKPHASE, the source code must be compiled with a Fortran compiler as for instance GFORTRAN or Intel Fortran. From my experience, compilation with Intel Fortran results in much faster executable. To compile:

gfortran LINKPHASE3.f90 -o LINKPHASE3
ifort LINKPHASE3.f90 -o LINPHASE3 (for Intel compiler)

Running LINKPHASE3

Recommendation

Please carefully read the printed information and check whether the information is correct (for instance the number of individuals, markers, etc). Errors indicate that the files are uncorrectly red or a problem in the parameter file.

Input files

As other programs from the PHASEBOOK package, LINKPHASE3 works per chromosome (marker and genotype files must be split per chromosome). Three input files are required: a pedigree file, a genotype file and a marker file. The format of the files is the same as for other programs from the PHASEBOOK package but is a bit more flexible. Fields in the genotype file needs only to be separated with a white space (no need of a fixed number of positions / digits). Marker number don't need to be consecutive.

1) Pedigree file. It contains three columns with: individual ID / father ID / mother ID. ID Numbers must be integers (not alphanumeric) and elder individuals must have the lowest numbers (***the pedigree must be sorted***). Ideally, the highest ID should not be too large.

2) Marker file. Three columns and one line per marker, with: marker number / marker name (50 characters at maximum and no white space within the name) / marker position (in cM). Markers must be orderer according to map position. For some species, when no genetic map is available, it can be assumed that 1 cM equals 1 Mb. Two markers can't have the same position; if two markers have the same position, CO is impossible between these markers and the MCS.txt file will contain problems.

3) Genotype file. One line per individual with: individual ID / 2 alleles for marker1 / 2 alleles for marker 2 / ... (2 x N alleles per line, where N is the number of markers). Alleles must be coded “1” and “2”, “0” for missing genotypes. See the option #COLUMN for an alternative format.

Parameters file / options

Name of files and options are red in the parameter file which must be called “linkin.txt” which has the following format:

```
#PEDIGREE_FILE
pedfilename
#GENOTYPE_FILE
genotypefilename
#MARKER_FILE
markerfilename
#HALFSIB_PHASING
yes
#HMM_PHASING
yes
#N_TEMPLATES
50
#CHECK_PREPHASING
yes
```

Lines starting by '#' cannot be changed and must be kept in the same order. Other lines contain

either names of input files or information for option. For most options (except N_TEMPLATES), the possible answers are 'yes' or 'no'.

HALFSIB_PHASING option. If 'yes' the program uses linkage information to reconstruct haplotypes of parents based on segregation of marker alleles in offsprings (similar to STEP2 in Druet and Georges (2010)). If 'no' the programs uses only Mendelian segregation rules (STEP1).

HMM_PHASING option. If 'yes' the program runs the new hidden Markov model (HMM) described in Druet and Georges (2015) to improve haplotype reconstruction after the HALFSIB_PHASING. This improves the haplotype reconstruction only in presence of genotyping errors. This option can be used only if the #HALFSIB_PHASING option was used.

N_TEMPLATES option. If 'yes' the program performs within family imputation as described in Druet and Georges (2015) to improve haplotype reconstruction. The within family imputation is usefull for parent with few genotyped offsprings and without their own parent genotyped.

CHECK_PREPHASING option. If 'yes' the program runs the HMM described in Druet and Georges (2015) after STEP1 (the use of Mendelian segregation). This allows to correct some haplotype reconstruction errors resulting from genotyping errors in parents. It seems preferable to detect such potential errors before running HALFSIB_PHASING, which might 'propagate' the errors. This option is mostly useful if a parent has its own parent genotyped.

On simulated data, use of all options improved haplotype reconstruction. Using LINKPHASE3 without the last three options results very similar to LinkPHASE_2.3.

Additional options

The linkin.txt parameter file must contain the previously mentioned information in the specified order. In addition, the user can add some optional features.

#SEXCHROM

Indicates that this is a sex-chromosome. In that case, the marker file must contain a fourth mandatory column indicating whether the marker is on the X-specific part with the letter “X” or on the pseudo-autosomal region (PAR) with the letter “P”. In the case this option is used, the pedigree file must also contain a fourth column indicating the sex of the individual (1 for males (heterogametic); 2 for females). Individuals with unknown sex will be ignored. In the heterogametic sex, a null haplotype with allele “9” is modeled and transmitted to offspring from the same sex.

Rules used to model the sex-chromosomes are described in Murgiano et al. (2016).

#SEXMAP

This option allows you to specify two genetic maps in the parameter file. In that case, the second map must simply be provided as an additional column just after the first map (and before the chromosome information in case the #SEXCHROM option is used).

#COLUMNS

With this option, an alternative format is used for the genotype file, with columns corresponding to individuals and lines to markers. The first line must contains the IDs of the individuals separated by a space (“ “). Genotypes are coded as 0, 1 and 2 (for 11, 12, 22). Other values are considered as missing.

#ITERATIONS
This option indicates that the user wants to estimate the genetic maps (one per sex) with the EM algorithm presented in Zhang et al. (2020). The number of iterations must be provided in the next line. The method takes into account informativeness and is better that the standard outputs that provide simple estimates of genetic lengths based on observed cross-overs. Indeed, some cross-overs might be missed due to low informativeness and the cross-overs are identified conditionally on the provided map. With the iterative method, the map is updated.

#GENO_ERROR
This option allows the user to specify the error term used in the emission probabilities from the hidden Markov model (see details in the paper). This values indicates the probability that the allele observed on the haplotype transmitted by the parent and the haplotype inherited by its offspring are different. This could be due to genotyping errors or local pre-phasing errors. By default, this value is set to 0.001.

Output files

By default, the program has four main output files with inferred haplotypes (phases), inheritance patterns (origins.txt), identified cross-overs (recombinations) and number of recombinations per offspring (nrec.txt). LINKPHASE3 uses STEP2 (haplotype reconstruction of parents based on marker allele segregation in offsprings) and HMM_PHASING option only if the parent has three or more genotyped offspring or if one of the corresponding grandparent (parent's parent) is also genotyped. Consequently, offspring in families with one or two halfsibs and without corresponding grand-parents genotyped are not included in the origins.txt, recombinations and nrec.txt files.

1) 'phases' has two lines per individual with individual ID / origin of haplotype / corresponding haplotype (serie of marker alleles with '0' when the marker is unphased). The origin is 1 for paternal and 2 for maternal haplotypes (when no information is available (e.g. no genotyped parent), paternal and maternal haplotypes are randomly labelled). With the #COLUMN option, that file is transposed (with IDs on first line, origin of haplotype on the second and one phase per column).

2) 'origins.txt' has the same format as 'phases' except that haplotypes are replaced by inheritance patterns (1 when the paternal haplotypes of the parent is inherited and 2 when the maternal haplotype is inherited). Inheritance patterns are identified thanks on informative markers (markers phased in both parent and offspring and heterozygous in the parent). With the #COLUMN option, that file is transposed (with IDs on first line, parental origin (parent ID) on the second, etc).

3) 'recombinations' is a list of identified cross-overs (CO) with offspring ID / parent ID / flanking informative marker 1 / flanking informative marker 2. IDs of flanking markers correspond to the position (line number) in the map file.

4) 'nrec.txt' contains offspring ID / parent ID / sex of parent / number of offsprings in the halfsib-family / parent phasing information / mate genotype information / number of heterozygous genotypes in the parent / number of homozygous genotypes in the parent / number of informative markers for CO identification / number of identified CO. The sex of parent is 1 for males and 2 for females. The 'parent phasing information' is 1 if the parent was phased in Step 1 based on Mendelian segregation rules (one of the corresponding grand-parents is genotyped). The 'mate genotype information' indicates whether the second parent is genotyped (1) or not (0).

When running the HMM_PHASING option, additional files are obtained.

1) emission_parents.txt, similar to 'phases' but haplotypes of marker alleles are replaced by emission probabilities of allele 1. It contains only parents (no offspring) phased with the HMM_PHASING

option. With the #COLUMN option, that file is transposed (with IDs on first line, origin of haplotype on second).

2) origins_hmm.txt, similar to 'origins.txt' but inheritance patterns are not based on 'informative markers' but corresponds to the inheritance probabilities obtained from the HMM (see Druet and Georges (2015)). With the #COLUMN option, that file is transposed (with IDs on first line, parental origin (parent ID) on the second, etc).

3) recombinations_hmm, similar to 'recombinations' but a CO is identified based on informative markers (markers phased in both parent and offspring and heterozygous in the parent) AND inheritance probabilities of the HMM. These inheritance probabilities must switch from < 0.001 to > 0.999 or vice versa. In simulations studies, these identified CO were better than those in 'recombination' file.

4) nrec_hmm.txt, similar to 'nrec.txt' but with CO identified in the 'recombinations_hmm' file (and not 'recombination' file). The last column represents an alternative way to count CO that must still be validated and should not be used at the present time.

5) MCS.txt, the 'Map Confidence Score' as described in Druet and Georges (2015). The file gives information to identify potential map errors. It contains the marker number (position in the marker file), the position of the marker in Morgans, the position of the next marker, the recombination rate estimation from the EM algorithm between the markers, the number of used parents, the genotyping error rates in parents, the number of used offsprings, the within haplotype allelic entropy measure and the map confidence score. If the CHECK_PREPHASING option is used, the score combines genotyping error rates and entropy measures from both the CHECK_PREPHASING and HMM_PHASING steps. In that file, recombination rates are obtained based on cross-overs counts and averaged for both sexes. The RRm.txt and RRf.txt contain the equivalent recombination rates (same approach) estimated respectively in males and females only.

6) genotyping_errors.txt indicates potential genotyping errors based on rules described in Druet and Georges (2015): individual ID, marker number (order in the marker file), step when the errors was identified (after Step2 'prephasing' or after Step3 'hmmphasing'), an indicator whether the individual is a 'parent' or an 'offspring' in the family, the genotype of the individual, probability or error, the likelihood of the four parental haplotype configurations (11, 12, 21 and 22), the number of halfsibs in the family. The identified genotyping errors can result from map errors. The user should first check in the MCS files which markers are incorrectly mapped.

7) convergence_per_parent.txt: information on convergence of HMM for each parent.

8) emap.txt: when the #ITERATIONS option is used to estimate the map length, a file emap.txt is produced with the sex-specific genetic maps. It contains the number of the marker interval (column #1), the phyiscal position of the first marker in bp/10,000,000 (column #2), the physical position of the second marker (column #3), the estimated genetic distance (in Morgans) in males (column #4) and the estimated genetic distance (in Morgans) in females (column #5).

References

T. Druet and M. Georges (2010) A Hidden Markov Model combining linkage and linkage disequilibrium information for haplotype reconstruction and QTL fine mapping. Genetics 184: 789-798.

T. Druet and M. Georges (2015) LINPHASE3: an improved pedigree-based phasing algorithm robust to genotyping and map errors. Bioinformatics 31(10):1677-9.

L. Murgiano, D.P. Waluk, R. Towers, N. Wiedemar, J. Dietrich, V. Jagannathan, M. Drögemüller, P. Balmer, T. Druet, A. Galichet, M.C. Penedo, E.J. Müller, P. Roosje, M.M. Welle, T. Leeb (2015) An Intronic MBTPS2 Variant Results in a Splicing Defect in Horses with Brindle Coat Texture. G3 6(9):2963-70.

Z. Zhang, N.K. Kadri, E. Mullaart, R. Spelman, S. Fritz, D. Boichard, C. Charlier, M. Georges, T. Druet (2020). Genetic architecture of individual variation in recombination rate on the X chromosome in cattle. Heredity 125:304-316.