RABBIT 2.0

Chaozhi Zheng

Biometris

Wageningen University and Research

Wageningen, The Netherlands

August 30, 2018

Contents

1 Introduction

RABBIT v2.0 has three key functions: magicReconstruct for haplotype reconstruction, magicImpute for genotype imputation, and magicMap for linkage map construction. The target mapping population can be bi-or multi-parental with founders being inbred or outbred. They have three common required arguments: genotype data, model, and population design; magicMap has an additional required argument to specify the number of linkage groups. Each function has many options, and each option is given in form of `optionname ->optionvalue`, where `optionvalue` is default value.

The RABBIT software is freely available from the web site: `https://github.com/chaozhi/RABBIT.git`

1.1 Citing RABBIT

If you use RABBIT in your analyses and publish your results, please cite the appropriate article.

The citation for RABBIT's haplotype reconstruction is
ZHENG, C., M. P. BOER, and F. A. VAN EEUWIJK, 2015 Reconstruction of genome ancestry blocks in multiparental populations. Genetics 200: 1073-1087.

The citation for RABBIT's genotype imputation is
ZHENG, C., M. P. BOER, and F. A. VAN EEUWIJK, 2018 Accurate genotype imputation in multiparental populations from low-coverage sequence. Genetics 210: 71-82.

The citation for RABBIT's map construction is
ZHENG, C., M. P. BOER, and F. A. VAN EEUWIJK, 2018 Construction of genetic linkage maps in multiparental populations. Submitted.

1.2 Genotype data

We denote the genotype data of a mapping population by a data structure called `magicsnp`. The input argument can be either data matrix or data file in CSV format. A valid `magicsnp` file is composed of three main parts: genetic map, founder genotypes, and offspring genotypes.

The `magicsnp` data matrix will look something like this:

nfounder	4						
marker	SNP1	SNP2	SNP3	...	SNP998	SNP999	SNP1000
chromosome	1	1	1	...	X	X	X
pos(cM)	0.12	0.23	1.2	...	95.1	98.6	99.3
founder1	2	N	1	...	2	1	2
...							
founder4	1	2	2	...	2	2	N
offspring1	12	22	11	...	1N	NN	22
offspring2	11	NN	2N	...	2	1	N
...							
offspring100	1N	22	12	...	11	12	2N

where quotes of strings are not shown. The quotes are not contained in the CSV file, and they will be automatically added after importing data.

- Row 1: The 1st element is somewhat arbitrary descriptive string. The 2nd element denotes the number of founders.
- Rows 2–4: Genetic map. The 1st elements of rows 2–4 are somewhat arbitrary descriptive strings. The 2–end elements of row 2 are the marker IDs that are unique. The 2–end elements of row 3 are chromosome (linkage group) IDs that are string or integer. The sex chromosome must be labelled by "X" or "x". The 2–end elements of row 4 are marker positions in cM, which must be non-decreasing within a linkage group. For map construction by magicMap, the chromosome IDs and marker positions are set to "NA".
- Rows 5–end: Genotypes of founders and offspring. Founders precede offspring, with boundary being determined by the number of founders in row 1.

All markers are assumed to be bi-allelic. The genotypes in row 5–end can be represented in two possible formats, but not a mixture of them for a given data file. The first representation is called genotypes, taking possible values 1, 2, "N", 11, 12, 22, "1N", "2N", and "NN". Here "N" denotes a missing allele, and the three genotypes (or alleles) 1, 2, and "N" are only for fully inbred founders or X chromosomes of males (e.g. offspring2). The second representation is allelic depths, denoted by "c1|c2", where c1 and c2 are the number of reads for alleles 1 and 2, respectively.

All input genotypes are assumed to be unphased. However, phased input called genotypes are allowed, where 21, "N1", and "N2" are equivalent to 12, "1N", and "2N", respectively.

1.3 Model

The second argument model describes the dependence of maternally and paternally derived chromosomes in an offspring, and it must be "depModel", "indepModel", or "jointModel". In general, we may set model to "depModel" for a homozygous population, and set to "indep-Model" for a heterozygous population. The general "jointModel" is preferred but in cost of somewhat computational time.

1.4 Population design

The population design information is specified by the third argument popdesign, which can be either mating schemes or a pedigree file in CSV format. Consider four-way recombinant inbred lines with two generations of selfing, the popdesign in form of mating schemes is given by `{"Pairing", "Pairing", "Selfing", "Selfing"}`. And the valid pedigree file will look something like this

Pedigree-Information	DesignPedigree			
Generation	MemberID	Gender	MotherID	FatherID
0	1	0	0	0
0	2	0	0	0
0	3	0	0	0
0	4	0	0	0
1	5	0	1	2
1	6	0	3	4
2	7	0	5	6
3	8	0	7	7
4	9	0	8	8
Pedigree-Information	SampleInfor			
OffspringID	MemberID	Funnelcode		
Offspring1	9	3-1-4-2		
Offspring2	9	1-2-4-3		
...				
Offspring100	9	2-4-1-3		

which is composed of two parts: design pedigree and sample information, and they are separable via the key string "Pedigree-Information".

In the design pedigree, the members are ordered so that parents are always above children. All members are labelled uniquely by natural number starting from 1. Founders are always in the beginning, and their parents are set to 0. The generation is non-decreasing starting from 0. The gender takes values 1 for female, 2 for male, and 0 for hermaphrodite or non-applicable. The gender is non-applicable if there are no sex chromosomes in `magicsnp`. The founder 1 corresponds to the first row of the genotype data (i.e. row 5 of `magicsnp`), and so on.

In the sample information, the offspring IDs must be the same as those in `magicsnp`. For non funnel based population design, the funnel code is always in the natural ordering (e.g. `1-2-3-4`), and the member IDs for offspring are different from each other.

2 RABBIT for haplotype reconstruction

2.1 Command line

The Mathematica command line used for haplotype reconstruction is given by

```
magicReconstruct[magicsnp, model, popdesign, options]
```

where the three required arguments are explained in the Introduction. If genetic map is unknown, please use magicMap to construct it. The overlapping markers at the same positions in `magicsnp` are jittered. If there are too many missing founder genotypes or founders are outbred and unphased, please use magicImpute to impute and/or phasing founder genotypes.

2.2 Options

founderAllelicError -> 0.005 to specify the allelic error probability in founders.

offspringAllelicError -> 0.005 to specify the allelic error probability in offspring.

isFounderInbred -> True to specify whether the founders are completely inbred. If isFounderInbred -> False, the founder genotypes are assumed to be phased.

sequenceDataOption -> optionvalue to specify options for sequence data with allelic depth. The default optionvalue = {isOffspringAllelicDepth -> Automatic, minPhredQualScore -> 30}.

isOffspringAllelicDepth -> Automatic to specify whether genetic data are allelic depths or called genotypes. By default, the form is detected automatically from input `magicsnp`.

minPhredQualScore -> 30 to specify the minimum of Phred quality scores among all markers.

outputFileID -> "" to specify the stem of output filenames.

isPrintTimeElapsed -> True to specify whether to print information such as running time.

reconstructAlgorithm -> "origPathSampling" to specify the alogrithm for haplotype reconstruction, and the option value must be "origPathSampling", "origPosteriorDecoding", or "origViterbiDecoding".

sampleSize -> 1000 to specify the number of posterior sampling when reconstructAlgorithm -> "origPathSampling", and it has no effects for other option values of reconstructAlgorithm.

2.3 Output files

The magicReconstruct returns a single output file, and it is transformed into a user-friendly summary file by

```
saveAsSummaryMR[outputfile, summaryfile]
```

where the `summaryfile` will be over written if it exists.

The `summaryfile` contains two key parts: ancestral genotype probabilities and ancestral haplotype probabilities for all offspring at all markers if reconstructAlgorithm -> "origPosteriorDecoding". Otherwise, it contains optimal ancestral origin path if reconstructAlgorithm ->"origViterbiDecoding", or independent sampled ancestral origin paths if reconstructAlgorithm -> "origPathSampling".

2.4 Visualization

The functions `plotAncestryProbGUI[summaryfile, trueFGLdiplofile,options]` or `plotAncestryProbGUI[summaryfile,options]` returns an animate for visualizing posterior probability if reconstructAlgorithm -> "origPosteriorDecoding". The summaryfile is the outputfile returned by saveAsSummaryMR, and trueFGLdiplofile gives the true ancestral origins if input data are simulated.

Besides the options for ListAnimate and ListPlot of Mathematica, there are two additional options: `isPlotGenoProb -> Automatic` specifices whether to visualize ancestral genotype probabilities or ancestral haplotype probabilities, and `linkageGroupSet->All` specifies the set of linkage groups to be visualized. For example, `linkageGroupSet->{1,3,4}` means that only results of the 1st, 3rd, and 4th linkage groups will be plotted.

3 RABBIT for genotype imputation

3.1 Command line

The Mathematica command line used for genotype imputation is given by

```
magicImpute[magicsnp, model, popdesign, options]
```

where the three required arguments are explained in the Introduction. magicImpute requires a genetic map, which can be obtained approximately from physical map by multiplying physical distances by a constant recombination rate.

3.2 Options

founderAllelicError -> 0.005 to specify the allelic error probability in founders.
offspringAllelicError -> 0.005 to specify the allelic error probability in offspring.
isFounderInbred -> True to specify whether the founders are completely inbred.
sequenceDataOption -> optionvalue to specify options for sequence data with allelic depth.

The default optionvalue = {isFounderAllelicDepth -> Automatic, isOffspringAllelicDepth -> Automatic, minPhredQualScore -> 30, priorFounderCallThreshold -> 0.99}.

isFounderAllelicDepth -> Automatic to specify whether genetic data of founders are allelic depths or called genotypes. By default, the form is detected automatically from input `magicsnp`.

isOffspringAllelicDepth -> Automatic to specify whether genetic data of offspring are allelic depths or called genotypes. By default, the form is detected automatically from input `magicsnp`.

minPhredQualScore -> 30 to specify the minimum of Phred quality scores among all markers.

priorFounderCallThreshold -> 0.99 to specify the threshold for prior calling of missing founder genotypes. Before founder genotype imputation, single locus calling of founder genotypes is performed if the posterior probability of the true genotype is greater than the threshold.

outputFileID -> "" to specify the stem of output filenames.

isPrintTimeElapsed -> True to specify whether to print information such as running time.

imputingTarget -> "All" to specify the imputing target, and it must be "Founders", "Offspring", or "All".

imputingThreshold -> 0.9 to specify an imputing threshold. A missing offspring genotype is imputed only if its posterior probability is greater than the threshold.

detectingThreshold -> 0.9 to specify a correction threshold. An observed genotype is corrected only if the posterior probability of the true genotype is greater than the threshold and is greater than the posterior probability of the observed genotype.

3.3 Output files

The magicImpute returns three output files in CSV format: `"stem_ErroneousGenotype.csv"`, `"stem_ImputedGenotype.csv"`, and `"stem_PosteriorProbability.csv"` if outputFileID -> "stem".

The file `"stem_ErroneousGenotype.csv"` saves the potential erroneous genotypes that are inconsistent between estimated genotypes and input genotypes in founders and offspring. If input genotypes are represented by allelic depth, the estimates are compared with single genotype calling.

The file `"stem_ImputedGenotype.csv"` is the same as a `magicsnp` file but with genotypic data being called and phased.

The file `"stem_PosteriorProbability.csv"` is the same as a `magicsnp` file except that a single genotype is represented by posterior probabilities. If the second argument model is set to "indepModel" or "jointModel" and the genotype does not belong to male X chromosome, it is represented like this "p11|p12|p21|p22" where p11, p12, p21, and p22 denotes the posterior probabilities of 11, 12, 21, and 22, respectively. If the second argument model is set to "depModel" or the genotype belongs to male X chromosome, it is represented like this "p11|p22" where p11 and p22 denotes the posterior probabilities of 11 (or 1 for male X) and 22 (or 2 for male X), respectively.

3.4 Visualization

`plotErrorPatternGUI[obsmagicsnp, estmagicsnp,truemagicsnp,options]` returns an user interface for visualizing the estimations of genotypes. The three required arguments correspond to observed, estimated, and true magicsnp, respectively. Genotypes are assigned one of statuses: "TrueCorrect" (genotype errors that are changed correctly), "TrueDetect"(genotype errors that are changed wrongly), "FalseNegative" (genotype errors that are not detected), "FalsePositive" (correctly observed genotype that are changed), "FalseImpute" (wrongly imputed genotypes), and the "Rest". In the resulting figure, the statuses are labelled by different colors that can be changed by user; the status of "Rest" is always labelled as white. Besides the options for MatrixPlot of Mathematica, the options include one extra option `linkageGroupSet->All` to specify the set of linkage groups.

`plotErrorPatternGUI[obsmagicsnp, estmagicsnp,options]` returns an user interface for visualizing the estimations of genotypes when true magicsnp is unknown. Genotypes are assigned one of statuses: "NonImputed" (missing genotypes are not imputed), "Imputed" (missing genotypes are imputed), "Correction" (observed genotypes are changed), and the "Rest".

4 RABBIT for map construction

4.1 Command line

The Mathematica command line used for map construction is given by

```
magicMap[magicsnp, model, popdesign, ngroup, options]
```

where the first three arguments are explained in the Introduction except that linkage groups and markers positions are set to missing ("NA"). The addition argument ngroup specifies the number of linkage group.

The magicMap consists of three consecutive steps:

```
pairwisefile=magicPairwiseSimilarity[magicsnp,model,popdesign,options]
initmapfile=magicMapConstruct[pairwisefile,ngroup,options]
refinefiles=magicMapRefine[initmapfile,magicsnp,model,popdesign,options]
```

The magicMap returns all output files `{pairwisefile,initmapfile,refinefiles}`, which will be explained later.

4.2 Options

founderAllelicError -> 0.005 to specify the allelic error probability in founders.

offspringAllelicError -> 0.005 to specify the allelic error probability in offspring.

isFounderInbred -> True to specify whether the founders are completely inbred.

sequenceDataOption -> optionvalue to specify options for sequence data with allelic depth. The default optionvalue = {isFounderAllelicDepth -> Automatic, isOffspringAllelicDepth -> Automatic, minPhredQualScore -> 30, priorFounderCallThreshold -> 0.99}.

> **isFounderAllelicDepth -> Automatic** to specify whether genetic data of founders are allelic depths or called genotypes. By default, the form is detected automatically from input `magicsnp`.
>
> **isOffspringAllelicDepth -> Automatic** to specify whether genetic data of offspring are allelic depths or called genotypes. By default, the form is detected automatically from input `magicsnp`.
>
> **minPhredQualScore -> 30** to specify the minimum of Phred quality scores among all markers.
>
> **priorFounderCallThreshold -> 0.99** to specify the threshold for prior calling of missing founder genotypes. Before founder genotype imputation, single locus calling of founder genotypes is performed if the posterior probability of the true genotype is greater than the threshold.

outputFileID -> "" to specify the stem of output filenames.

isPrintTimeElapsed -> True to specify whether to print information such as running time.

imputingThreshold -> 1 to specify an imputing threshold. A missing offspring genotype is imputed only if its posterior probability is greater than the threshold. By default, we do not impute missing offspring genotypes in each iteration.

detectingThreshold -> 0.9 to specify an imputing threshold. A missing offspring genotype is imputed only if its posterior probability is greater than the threshold.

computingLodType -> "both" to specify the type of two-locus analysis. It must be "independence", "linkage", and "both", corresponding to independence test, linkage analysis, or both. "

isRunInParallel -> True to specify whether to compute in parallel.

minLodSaving -> 1 to specify the minimum LOD score for saving results. For a given pair of markers, If the LOD of linkage analysis or independence test is smaller than the threshold, the two-locus analysis will not be saved.

miniComponentSize -> 5 to specify the minimum size of a graph component. The markers in a component are ungrouped if the component size is smaller than the threshold.

graphLaplacian -> "rwNormalized" to specify one of three graph Laplacians: "unNormalized","rwNormalized",or "symNormalized".

lodTypeClustering -> "both" to specify the LOD type for clustering. It must be one of ïndependence, l̈inkage, and b̈oth, corresponding to independence test, linkage analysis, or both.

lodTypeOrdering -> "both" to specify the LOD type for ordering. It must be one of ïndependence, l̈inkage, and b̈oth, corresponding to independence test, linkage analysis, or both.

minLodClustering -> Automatic to specify the minimum LOD score for clustering. The similarity between two markers is set to 0 if its LOD score is smaller than the threshold.

minLodOrdering -> Automatic to specify the minimum LOD score for ordering. The similarity between two markers is set to 0 if its LOD socre is smaller than the threshold.

nNeighborFunction -> $(Sqrt[\#]\&)$ to specify the pure function defining the number of neighbors used in the spectral ordering of magicMapConstruct. "

nNeighborSaving -> 10 to specify the number of strongest neighbors to be saved for magicMapRefine.

referenceMap -> None to specify the filename of a reference map, which is used only to compare with estimated map.

nReplicateAnnealing -> 1 to specify the number of times repeating simulated annealing.

initTemperature -> 2 to specify the initial annealing temperature.

coolingRate -> 0.85 to specify the cooling rate of annealing temperature.

freezingTemperature -> 0.5 to specify the freezing temperature at which cooling rate increases.

deltLoglThreshold -> 1 to specify the stopping threshold. Simulated annealing is finished if the change of log likelihood is small than the threshold in three consecutive iterations.

MaxIterations -> 50 to specify the maximum number of iterations for simulated annealing.

4.3 Output files

The magicMap returns output files `{pairwisefile,initmapfile,refinefiles}`, corresponding to the output files of `magicPairwiseSimilarity`, `magicMapConstruct`, and `magicMapRefine`, respecitively. The `refinefiles` consists of three files: `refinedmapfile`, `refinedmagicsnpfile`, and `refininghistory` file.

The most useful output files are the two map files `initmapfile` and `refinedmapfile` in CSV format, and the other output files are for re-performing some steps and possibly detailed examinations. The first columns of each map file are marker ID, linage group, and genetic position in cM, other columns showing some details of map construction.

4.4 Visualization

The function `plotMapComparison[mapfile1,mapfile2,isordering,linestyle,options` is used for map comparisons. Here mapfile1 or mapfile2 can be `initmapfile`, `refinedmapfile`, or any other map files (e.g. physical map) three columns: marker ID, linage group, and genetic position. The argument `isordering` is to specify whether comparing only marker ordering. If `isordering=True`, the third column is not required. Any extra non-required columns in map files will be neglected. The 4th argmument `linestyle` specifies the line style of chromosome boundaries. The opitons are the same as the otions of ListPlot of Mathematica.

The function `plotHeatMap[pairwisefile,mapfile,options]` returns heat map of pairwiserecombination fraction or LOD score matrix. The `pairwisefile` is the outputfile returned by magicPairwiseSimilarity, and the mapfile can be `initmapfile`, `refinedmapfile`, or any other map files (e.g. physical map) the first two columns: marker ID (ordered) and linage group. Besides the options for MatrixPlot of Mathematica, the options include two extra options: `rescaleSimilarity -> True` specifies if rescale recombination fraction or LOD score, and `linkageGroupSet->All` specifies the set of linkage groups.

The function `plotHeatMapGUI[pairwisefile,mapfile,options]` is similar to `plotHeatMap[pairwisefile,mapfile,options]`, but with an user interface for visualizing heat map.