Flexible Monte Carlo Identity-By-Descent Matrix Estimation with Given Base Generation Structures in F2 Intercross Designs

Xia Shen*[1,3] Carl Nettelblad[2] , Lars Rönnegård[3,4] and Örjan Carlborg[1,4]

[1]The Linnaeus Centre for Bioinformatics, Uppsala University, Uppsala, Sweden
[2]Department of Information Technology, Uppsala University, Uppsala, Sweden
[3]Statistics Group, Dalarna University, Borlänge, Sweden
[4]Department of Animal Breeding & Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden

Email: Xia Shen*- xia.shen@lcb.uu.se; Carl Nettelblad - carl.nettelblad@it.uu.se; Lars Rönnegård - lrn@du.se; Örjan Carlborg - orjan.carlborg@hgen.slu.se;

*Corresponding author

Abstract

Summary: We present a new stochastic identity-by-descent (IBD) matrix estimation program for large F_2 intercrosses using an R package interface. Here, arbitrary segregation structures for the founder alleles are allowed. Genotypic and gametic IBD matrices can be estimated for single-locus or for two-loci with epistatic effects. To enhance the program performance, parallelized computing using the **snowfall** package is implemented. Additional functions allow calculating locus-specific IBD matrices along entire chromosomes or epistatic IBD matrices for pairs of chromosomes. The output matrix has several format options, where we propose a format of principle-component incidence matrix that improves the estimation efficiency of variance component models in quantitative trait loci (QTL) analyses. This package is useful for our previously developed flexible intercross analysis (FIA) method, which models within-line segregation in analyses of outbred line-cross QTL mapping experiments.

Availability: The flexible Monte Carlo IBD estimation program has been implemented in the R package **MCIBD**, which is open source and freely available from http://r-forge.r-project.org/projects/mcibd. The website describes the package with a tutorial of installation in R. Users can find a full documentation after installing the package.

Contact: xia.shen@lcb.uu.se

1 INTRODUCTION

In both animal breeding and human genetics, variance component methods have been widely used for detecting quantitative trait loci (QTL). The variance-covariance matrix of the random QTL effect, *i.e.* the *identity-by-descent (IBD) matrix*, is required to conduct such QTL analyses. Either deterministic [1, 2] or stochastic [3, 4, 5, 6] approaches have been used for IBD estimation in different population structures, using molecular markers and pedigree data. We have recently proposed an improved method for QTL detection using variance component models [7], where within-line segregation is included in the covariance structure. For estimating the IBD matrices for *FIA (flexible intercross analysis)*, we previously used a deterministic method. However, to develop the use of FIA, new IBD estimation softwares are needed, which have a flexible capacity for estimating different types of IBD matrices with arbitrary segregation structure of founders, high efficiency with better use of partially informative markers, and a user-friendly interface. In detecting epistasis [8], IBD matrices with defined segregation structure of founders are also strongly needed where the epistatic IBD matrix for arbitrary two test loci are involved in variance component models.

The aim of this application note is to introduce a new stochastic IBD estimation program, which satisfies the above requirements. The method is based on Monte Carlo sampling and supports any F_2 intercross pedigree with arbitrarily large size. The implementation relies on the `cnF2freq` routine [9] and is referred to as *Monte Carlo identity-by-descent (MCIBD)* matrix estimation.

2 PACKAGE IMPLEMENTATION

The **MCIBD** package is implemented as an integration of the C++ based software `cnF2freq` and the R sources for Monte Carlo sampling and relative calculation. `cnF2freq`, using a *hidden Markov model*, computes the inheritance probabilities for each individual at each locus. These probabilities are thereafter used by the Monte Carlo sampling routine which constructs the incidence matrices and calculates the IBD matrices (Figure 1).

The current version of **MCIBD** contains six objects, in which four functions are defined (Table 1). `cnF2freq` calculates all the inheritance probabilities for a given pedigree structure and marker genotypes along a particular chromosome. The output probabilities are necessary for all the other functions. `MCIBD` calculates a single IBD matrix for a given locus or a single epistatic IBD matrix for two given loci. `MCIBD.chro` calculates locus-specific IBD matrices for the entire chromosome. `MCIBD.epi2chro` calculates epistatic IBD matrices for two chromosomes and is capable of computing epistatic IBD matrices for linked loci.

3 CHIEF ARGUMENTS

The Monte Carlo strategy estimates IBD matrices by simulating incidence matrices using allele dropping. With such incidence matrices, several important implementations are straightforward. We refer to the design of QTL random effect as the *incidence matrix*, $\mathbf{Z}$. So that if the corresponding genotypic IBD matrix is denoted by $\mathbf{\Pi}$, the equation $\mathbf{\Pi} = \frac{1}{2}\mathbf{Z}\mathbf{Z}'$ holds [10].

Defining Segregation Argument `segregation` defines the segregation structure of the founder alleles. FIA, for instance, requires two types of IBD matrices, where one assumes fixation within lines and another assumes complete segregation of founders. Segregation is an input vector in R with integer indices for founder alleles.

IBD Storage Formats The package is capable of estimating both `'gametic'` and `'genotypic'` IBD matrices. Since a full-sized IBD matrix is usually big and not easy to use in variance component estimation, we propose four different output formats, which similarly as for the argument `output.Z` can be selected from the following. `'none'` - An ordinary full-sized IBD matrix is saved for the test locus. We have shown that such a full-sized IBD matrix usually has a high rank, which will lead to unnecessary computational requirements [11]. `'all'` - Like `'none'`, an ordinary full-sized IBD matrix is saved for the test locus. Furthermore, all the incidence matrix imputes are saved for later use. This is generally not recommended since more space is needed for storage. `'av'` - An average incidence matrix is saved for the test locus. Instead of taking the mean of IBD matrices, we directly take the mean of all the incidence matrix imputes. This average incidence matrix gives the allelic probabilities for each F_2 individual, which is the likelihood that each founder allele is received by a particular F_2 offspring. It also saves storage space. `'pc'` - A *principle-component incidence (PCI) matrix* is saved for the test locus. The IBD matrix at a specific locus is, or approximately, semi-positive definite. Recently, we proposed to use a rank-reduced IBD matrix by eigenvalue decomposition, which makes estimation of the statistical models computationally more efficient [11]. Now we develop this idea without storing the rank-reduced IBD matrix by using the corresponding PCI matrix $\mathbf{Z}$ instead.

High Performance Computing To speed up the calculations for whole genome scans using high performance computing on multi-core computers or clusters, we allow a user-friendly interface for parallelization of the functions `MCIBD`, `MCIBD.chro` and `MCIBD.epi2chro`. The package calls the **snowfall** package for parallelization [12]. The argument `hpc = TRUE` turns on the high performance computing, and `n.cpus` gives the number of processors on which the computing is performed.

References

1. Wang T, Fernando R, van der Beek S, Grossman M, van Arendonk J: **Covariance between Relatives for a Marked Quantitative Trait Locus**. *Genet. Sel. Evol.* 1995, **27**:251–274.
2. Pong-Wong R, George A, Woolliams J, Haley C: **A Simple and Rapid Method for Calculating Identity-by-descent Matrices using Multiple Markers**. *Genet. Sel. Evol.* 2001, **33**:453–471.
3. Pérez-Enciso M, Varona L, Rothschild M: **Computation of Identity by Descent Probabilities Conditional on DNA Markers via a Monte Carlo Markov Chain Method**. *Genet. Sel. Evol.* 2000, **32**:467–482.
4. Thompson EA, Heath SC: **Estimation of conditional multilocus gene identity among relatives**. *Statistics in Molecular Biology and Genetics* 1999, **33**:95–113.
5. Besnier F, Carlborg Ö: **A General and Efficient Method for Estimating Continuous IBD Functions for Use in Genome Scans for QTL**. *BMC Bioinformatics* 2007, **8**(440).
6. Mao Y, Xu S: **A Monte Carlo algorithm for computing the IBD matrices using incomplete marker information.** *Heredity* 2005, **94**(3):305–315.
7. Rönnegård L, Besnier F, Carlborg Ö: **An improved method for quantitative trait loci detection and identification of within-line segregation in F2 intercross designs.** *Genetics* 2008, **178**(4):2315–2326.
8. Carlborg Ö, Haley CS: **Epistasis: too often neglected in complex trait studies?** *Nat. Rev. Genet.* 2004, **5**.
9. Nettelblad C, Holmgren S, Crooks L, Carlborg Ö: **cnF2freq: Efficient Determination of Genotype and Haplotype Probabilities in Outbred Populations Using Markov Models**. *Lecture Notes in Bioinformatics (LNBI)* 2009, **5462**:307–319.
10. Rönnegård L, Carlborg Ö: **Separation of base allele and sampling term effects gives new insights in variance component QTL analysis.** *BMC Genetics* 2007, **8**:1.
11. Rönnegård L, Mischenko K, Holmgren S, Carlborg Ö: **Increasing the efficiency of variance component quantitative trait loci analysis by using reduced-rank identity-by-descent matrices.** *Genetics* 2007, **176**(3):1935–1938.
12. Knaus J, Porzelius C, Binder H, Schwarzer G: **Easier Parallel Computing in R with snowfall and sfCluster**. *The R Journal* 2009, **1**.

Object	Type	Description
`cnF2freq`	*function*	Calculating inheritance probabilities using `cnF2freq`
`MCIBD`	*function*	Estimating IBD matrices using Monte Carlo sampling
`MCIBD.chro`	*function*	Estimating IBD matrices along a whole chromosome
`MCIBD.epi2chro`	*function*	Estimating epistatic IBD matrices for two chromosomes
`pedigree`	*data*	A pig pedigree structure data set
`probabilities`	*data*	A data set of probabilities of pig chromosome 6 from `cnF2freq`

Table 1: Objects of the **MCIBD** package.

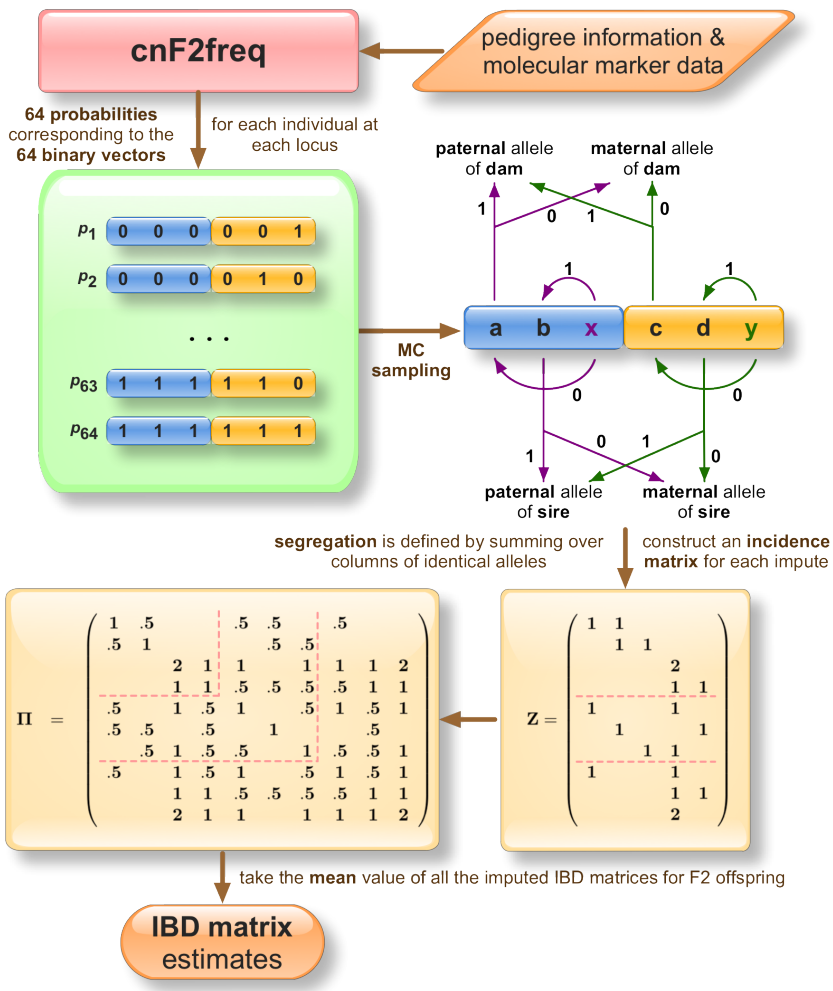

Figure 1: An interpretive flow chart of the **MCIBD** package.