IBDLD Software Documentation

Version 3.37- fast and parallelized program

Lide Han and Mark Abney

Departments of Human Genetics
The University of Chicago

IBDLD
A C++ fast and parallelized program for multipoint IBD sharing estimation based on a given pair of individuals' multilocus SNP genotypes with or without pedigree information.

Homepage: http://sourceforge.net/projects/ibdld/
Release 3.37, January 6, 2017

===

Contents

1 Getting started with IBDLD

This page contains some important information on learning to use IBDLD and how to handle any problems you encounter.

1.1 Citing IBDLD

If you use IBDLD in any published work, please cite the manuscript describing the method.

Han L. , Abney M. (2011). Identity by descent estimation with dense genome-wide genotype data. *Genetic Epidemiology*, 35 : 557-567.

Han L., Abney M. (2013). Using identity by descent estimation with dense genotype data to detect positive selection. *European Journal of Human Genetics*, 21: 205-211.

1.2 Reporting problems, bugs and questions

If you have any problems with IBDLD or would like to report a bug, or an analysis does not report the results you expect, etc, please contact the first author:

Lide Han
Department of Human Genetics
University of Chicago
Cummings Life Science Center, Room 406A,
920 E. 58th St. Chicago, IL 60637
Email: lidehanc827@gmail.com

But also please consider the following before doing so:

1) Please check the `Log` file, it often contains important information.

2) Please check the format of your data, make sure that each file has the correct number of rows, etc.,

If the above steps do not resolve your problem, then please email me. The more specific your email, the easier it will be for me to diagnose any problem or error. Please include:

1) The whole `Log` file.

2) Ideally, please try to make some reduced dataset that replicates the problem that you are able to send to me in a zip or gzip file, so that I will be able to recreate the problem; any data sent to me for these purposes will be immediately deleted after I have resolved the problem.

2 Overview of IBDLD

2.1 Basic Concepts

Identity-by-descent (IBD): two homogeneous alleles at a single locus are identical by descent if they are identical copies of the same allele in some earlier generation, i.e., both are copies that arose by DNA replication from the same ancestral sequence without any intervening mutation. For individuals in a pedigree, their IBDs are with respect to a documental common ancestry, we call them family IBD.

Homozygous by descent (HBD): two identical alleles at a given SNP in an individual are descended from a single source, as may occur in consanguineous mating.

IBD sharing at a SNP: is also named as proportion of alleles shared IBD, i.e. the proportion of alleles that are IBD at a SNP in two individuals. For the SNP k, its estimated IBD sharing can be expressed as $\hat{\pi}_k = \hat{\Delta}_{1k} + \frac{1}{2}(\hat{\Delta}_{3k} + \hat{\Delta}_{5k} + \hat{\Delta}_{7k}) + \frac{1}{4}\hat{\Delta}_{8k}$, $\hat{\Delta}_{ik}$=Pr(s_{ik}|genotype data of SNP k) $(i = 1, \cdots, 9)$, $\hat{\Delta}_{ik}$ is estimated condensed identity coefficients at SNP k, s_{ik} is condensed identity state at SNP k.

Kinship coefficient: is also called as coefficient of coancestry, a measure of relatedness between two individuals. It represents the probability that two alleles at a given genetic locus, sampled at random from each individual are IBD. It is calculated from pedigree records, we can name it as pedigree-based kinship coefficient. The kinship coefficient between individual i and j is expressed as $\phi_{ij} = E(\hat{\pi}_k)$.

Inbreeding coefficient: is the kinship coefficient between an individual‘s parents, and measures the probability that two alleles of a given locus are IBD in the individual, i.e. that they are copies of a single allele of ancestor common to both parents. It is zero if the individual is not inbred. It is calculated from pedigree records.

Genome-wide condensed identity coefficients: is the average condensed identity coefficients across K loci along the whole genome between two individuals, which is expressed as $\bar{\hat{\Delta}}_i = \frac{1}{K}\sum_{k=1}^{K} \hat{\Delta}_{ik} = \frac{1}{\sum_{h=1}^{nChr} W_h} \sum_{h=1}^{nChr} W_h \bar{\hat{\Delta}}_{hi}$ $(i = 1, \cdots, 9)$, W_h is the number of SNPs on the hth chromosome, $\bar{\hat{\Delta}}_{h.}$ is the hth chromosome's chromosome-wide condensed identity coefficients, $nChr$ is a total chromosome number.

Empirical kinship coefficient: is the average estimated IBD sharing across K loci along the whole genome between two individuals. The estimated empirical kinship coefficient between individual i and j is expressed as $\hat{\pi}_{ij} = \frac{1}{K}\sum_{k=1}^{K} \hat{\pi}_k = \frac{1}{\sum_{h=1}^{nChr} W_h} \sum_{h=1}^{nChr} W_h \hat{\pi}_h$, W_h is the number of SNPs on the hth chromosome, $\hat{\pi}_h$ is the hth chromosome's chromosome-wide empirical kinship coefficient.

Empirical inbreeding coefficient: is the average probability of autozygosity across K loci along the whole genome. The estimated empirical inbreeding coefficient of individual i is expressed as $\hat{f}_i = \frac{2}{K}\sum_{k=1}^{K} \hat{\pi}_k - 1.0 = \frac{1}{\sum_{h=1}^{nChr} W_h} \sum_{h=1}^{nChr} W_h \hat{\pi}_h$, W_h is the number of SNPs on the hth chromosome, $\hat{\pi}_h$ is the hth chromosome's chromosome-wide empirical inbreeding coefficient.

IBD segment: is a segment of DNA that is found to be identical in two related individuals, and is passed down to both of them from a common ancestor. The definition of the boundary of the estimated IBD segment is given as the first locus for which the IBD probability exceed some threshold through to the last such continuous locus. It consists of two types: IBD1 segment (which usually is named as IBD segment) and IBD2 segment. For IBD segment, its IBD probability (of sharing at least 1 allele IBD) at locus k can be expressed as $\hat{\Delta}_{1k} + \hat{\Delta}_{3k} + \hat{\Delta}_{5k} + \hat{\Delta}_{7k} + \hat{\Delta}_{8k}$ for pairs of different individuals and $\hat{\Delta}_{1k}$ for pairs of same individuals (HBD sharing segment for an individual). For IBD2 segment, its IBD probability (of sharing 2 alleles IBD) at locus k can be expressed as $\hat{\Delta}_{1k} + \hat{\Delta}_{7k}$ for only pairs of different individuals.

Empirical kinship coefficient, genome-wide condensed identity coefficient and empirical inbreeding coefficient, are also called marked-based kinship coefficient, marked-based condensed identity coefficient and marked-based inbreeding coefficient, respectively, which are estimated based on all markers or markers and pedigree joint information. The genome-wide IBD sharing consists of empirical kinship coefficient and empirical inbreeding coefficients. In contrast to pedigree-based kinship, marker-based kinship estimates is more accurate measure of true IBD sharing between pairs of individuals, and may account for the effects of deviations from expected parental contributions to progeny due to selection or genetic drift (Bernardo et al. 1993). Therefore, marker-based kinship estimates are recommended to use in mixed-model association-mapping approaches instead of pedigree-based kinship.

Note in this software, under some situations the expected value of the IBD sharing is equal to the kinship coefficient, and its range is between 0 and 1. Under the other commonly used situations the expected value of IBD sharing is twice the kinship coefficient, and its range is between 0 and 2.

2.2 About IBDLD

IBDLD is a program, written in C++, that estimates IBD sharing probabilities based on a given pair of individuals' multilocus single nucleotide polymorphism (SNP) genotypes with or without the known arbitrary pedigree. For the known pedigree, it takes each pedigree as a whole and make full use of pedigree information. It accounts for linkage disequilibrium (LD) between SNPs for the high-density SNP data.

The program executes the following tasks:

1) To compute 3 IBD states (IBD= 0, 1, 2), 4 IBD states (IBD= 0, 1, 2, 4) probabilities, 9 condensed identity coefficients, IBD sharing at each locus.

2) To estimate genome-wide condensed identity coefficients and empirical kinship coefficients for all pairs of individuals, which can be used to detect sample contaminations and duplications as well as pedigree errors and undocumented relationships.

3) To detect extended chromosomal segmental IBD sharing between pairs of related individuals, i.e. IBD segment and IBD2 segment.

4) To calculate empirical inbreeding coefficients and HBD probabilities at each locus for an individual.

These tasks are broken into two separate computational steps:

Step 1: is for estimating the background LD parameters, either the ridge regression coefficients or the two-locus haplotypes frequencies;

Step 2: is for estimating IBD sharing probabilities of many SNPs for a pair of individuals.

These tasks can be done with the five following methods.

1)`NoLD`. A standard Hidden Markov model (HMM), which considers each marker independent, i.e. no LD between markers;

2)`NoLD-S`. The method is identical to the first method, but uses a sparse set of markers, a random SNP is chosen every k centiMorgan (cM) along the chromosome;

3)`LD`. The method extends the HMM to include LD by conditioning on a single SNP with the highest correlation to the current SNP from among the previous n markers within k cM along the chromosome;

4)`LD-RR`. The method extends the HMM to include LD by conditioning on the n previous markers with ridge regression.

The above four methods need detail pedigree information, we can obtain family IBD.

5)`GIBDLD`. The method is similar to method 4, but it doesn't need pedigree information.

The program uses different algorithms to obtain the corresponding estimates in two steps for all methods.

At step 1, the background LD parameters are estimated based on a training sample (a sample representative of the study population).

At step 2, the program uses Baum's forward-backward algorithm to estimate IBD sharing probabilities at each locus.

This program computes the IBD sharing probabilities at each locus based on chromosome wide SNP genotype data, so the final result is output by chromosome and whole genome. The program allows for missing genotypes and genotyping error, so the user does not need to impute missing values. Missing rate is suggested to be less than 20% for each individual and the minimum call rate is 95% for each SNP, it is better with accurate imputation.

3 Running ibdld

Installation instructions:

GNU make is at least the version of 3.81 (`http://ftp.gnu.org/gnu/make/`). For IBDLD later Version 2.10, Eigen of version 3.1.2, a C++ template library, is used to speed up the linear algebra computations. For IBDLD later Version 3.00, the compiler implements version 2.5 of the OpenMP standard library is used, which result in the program parallel execution at run time. If the openMp is not in your computer, you must install the openMp library (version ≥ 2.5) before you compile the program. For IBDLD later Version 3.1, we improved Method 1-4 by reduce one step (The transition rates estimates are estimated by Maximum likelihood with Brent method instead of using a Monte Carlo approach for previous version). For IBDLD later Version 3.13, the g++ version is at least above 4.4. For IBDLD later Version 3.3, we improved all methods by estimating the parameters with penalized expectation maximum algorithm. For MAC, the package management system of `MacPorts` will simplify the installation of g++.

1. Download the ibdld package. This package contains documentation, source code, example input and output files, and precompiled executables for Linux and Mac platforms.

2. Read the entire documentation (this document) carefully to understand the purpose of this program and how it works.

3. Type `make`. This will build an executable programs called "ibdld" .

4. To run the executable program “`ibdld`”:

First, prepare the input files, e.g., pedigree file(*pedigree-file*), map file (*map-file*), identity coefficient file (*identity-coefficient-file*), study individual identifications or pair individuals IDs list file (*study-file*), training sample file(*training-sample-file*), marker name list file (*marker-file*) etc. Then, according to input filenames or settings, one type “`./ibdld`” and use the flags and filenames in the command line. In following command lines, k is a real number, n is a integer, both of which are nonnegative.

Below is the command and all possible command line arguments.

Primary command line argument are the following style, items within the <> are optional.

```
./ibdld \
 [-o file-prefix ]\
 [[-p|-plink pedigree-file  -m|-m_int map-file] |-plinkbf binary-data-file-prefix |-plinkbf_int binary-data-file-prefix] \
 [-method method-choice -ploci n −dist k] \
 [-step n] \
<-bgld LD-parameter-file-prefix > \
<-unphased |-phased training-sample-file > \
<-i identity-coefficient-file > \
<-s study-file > \
<-marker marker-file > \
<-rars > \
<-mind k > \
<-unsort > \
<-mincallrate k > \
<-hiddenstates 3|9 > \
<-r n > \
<-phcol n > \
<-morgan > \
<-MAF k > \
<-error k > \
<-segment −−min k −−max k −−SNP n −−length k > \
<-ibd2segment −−tmin k −−tmax k −−tSNP n −−tlength k > \
<-ibc > \
<-ibd n −−ibdtxt > \
<-hbd −−hbdtxt > \
<-chr schr > \
<-nthreads n > \
<-noprint > \
<-make-grm> \
<-v > \
<-h >
```

The primary command line arguments:

`-o` ***`file-prefix`*** Allow the user to specify a prefix for the filename for all output files. The filename prefix must be an absolute or relative filename, but it cannot be a directory. For instances, if the prefix is set as "/users/a" (`-o /users/a`), the output filenames are "/users/a.kinship" (`Empirical Kinship Coefficient File`), "/users/a.ibdbin" (`IBD File in binary format`), "/users/a.ibdtxt" (`IBD File in text format`). If the flags of "`-p`", "`-m`" and "`-i`" are not used, the pedigree-file, map-file and identity-coefficient-file are assumed to be "./prefix.ped", "./prefix.map" and "./prefix.idcoeff", respectively. Defaults to "prefix".

`-p` ***pedigree-file*** or `-plink` ***pedigree-file*** Allow the user to specify the name of pedigree file. The two files' format are different, the former is MERLIN format pedigree file, the latter is PLINK format pedigree file. Defaults to prefix.ped;

`-m` ***map-file*** or `-m_int` ***map-file*** Allow the user to specify map file name. The two files format are same except the 3rd column, the former describes each SNP's genetic position, the latter describes the genetic distance between adjoining SNPs. Defaults to prefix.map;

`-plinkbf` ***binary-data-file-prefix*** Allow the user to specify plink binary data file prefix. The program load a binary file *binary-data-file-prefix.bed* and two text files of *binary-data-file-prefix.fam* and *binary-data-file-prefix.bim.* Defaults: prefix.bed, prefix.fam and prefix.bim three files, and the 3rd column of the map file describes each SNP's genetic position;

`-plinkbf_int` ***binary-data-file-prefix*** Allow the user to specify plink binary data file prefix. The program load a binary file *binary-data-file-prefix.bed* and two text files of *binary-data-file-prefix.fam* and *binary-data-file-prefix.bim.* Defaults: prefix.bed, prefix.fam and prefix.bim three files, and the 3rd column of map file describes the genetic distance between adjacent SNPs;

`-method` ***method-choice*** Allows the user to select an analysis method from the four methods of NoLD, NoLD-S, LD, LD-RR and GIBDLD. The default value is GIBDLD. Table 1 summarizes the options that are compatible with the methods in the method-choice;

Table 1: Command line argument choices in the method-choice.

method	ploci	dist	phased	unphased
NoLD	No need	No need	Optional	Optional
NoLD-S	No need	Required	Optional	Optional
LD	Required	Required	Required	
LD-RR	Required	Required	Optional	Optional
GIBDLD	Required	Required	Optional	Optional

For method GIBDLD and LD-RR, two options `-ploci` n and `-dist` k are jointly considered for LD pattern, which means using n previous loci within k cM along the chromosome. If the n or k is not given, it defaults as above (-ploci 10 -dist 2).

`-ploci` n Allow the user to set the number of previous loci n for a method. For method LD, LD-RR and GIBDLD, this flag is required, but setting n is optional. It defaults to 10 if it is not set. The flag is not need for the method NoLD and NoLD-S.

`-dist` k Allow the user to set a distance for choosing a SNP for NoLD-S, and for using all previous SNPs with k (cM) distance for LD, LDRR and GIBDLD. This flag must be used when method NoLD-S, LD, LDRR and GIBDLD are chosen, however, setting k is optional. It defaults to k= 2 (cM) if it is not set.

`-step` *n* Allow the user to set n for executing the different task, such as "-step 1" for estimating the LD parameters, "-step 2" for estimating IBD sharing. The user can use -step 0 to execute the two tasks in order. The default value is n=0.

The user sets different options and flags for the two steps.

Step 1:

```
./ibdld \
[-o prefix ] \
[[-p|-plink pedigree-file  -m|-m_int map-file] |-plinkbf binary-data-file-prefix |-plinkbf_int
binary-data-file-prefix] \
[-method method-choice -ploci n -dist k] \
[-step 1] \
<-bgld LD-parameter-file-prefix > \
<-unsort > \
<-rars > \
<-mincallrate k > \
<-unphased |-phased training-sample-file > \
<-r n > \
<-mind k > \
<-morgan > \
<-MAF k > \
<-phcol n > \
<-chr schr > \
<-noprint > \
<-nthreads n >
```

Step 2:

```
./ibdld \
[-o prefix] \
[-method method-choice -ploci n -dist k] \
[-step 2] \
[ -i identity-coefficient-file] \
<-bgld LD-parameter-file-prefix > \
<-hiddenstates 3|9 > \
<-marker marker-file > \
<-s study-file> \
<-noprint > \
<-r n > \
<-error k >\
<-ibc > \
<-rars > \
<-ibd n −−ibdtxt >\
<-hbd −−hbdtxt >\
```

<-segment −−min k −−max k −−SNP n −−length k > \
<-ibd2segment −−tmin k −−tmax k −−tSNP n −−tlength k > \
<-make-grm-chr> \
<-chr *schr* > \
<-nthreads > n

The available options are as follows:

`-bgld` *`LD-parameter-file-prefix`* Allow the user to set the filename prefix of the intermediate output files (consist of `Method Statement File`, `Background LD Parameter File`, `Ridge Regression Coefficient File, available individuals with enough genotype informations ID File`) of `step 1` when the user wants to create or use the filename prefix that is different from the one defined by `-o prefix`. The option `-bgld` are useful when you want the filename prefixes of the intermediate output files to be different from the filename prefix for the main output files. If the option was used at `step 1`, it must be used at `step 2`, which requires the *LD-parameter-file-prefix* as input. Defaults to prefix.

`-i` *`identity-coefficient-file`* Allow the user to specify the name of the identity coefficients file. This flag must be used when the user uses the NoLD, NoLD-S and LD-RR method. Defaults to prefix.idcoeff.

`-s` *`study-file`* Allow the user to specify the name of a study sample file for steps 0, 2. This file is used to define which pairs' IBD probabilities will be computed. If this flag is not set, the program will compute the IBD probabilities for each pair with condensed identity coefficients (LD-RR, NoLD, NoLD-S, LD) or all possible pairs of individuals has enough non-missing rate genotypes (GIBDLD). Defaults to prefix.study.

`-marker` *`marker-file`* Allow the user to specify the name of a file that contains SNP identifiers (or rs#), defining which SNP's IBD probabilities will be output. If this flag is not used, all SNPs IBD probabilities will be output. The analysis will always be done using all SNPs in the pedigree file (except NoLD-S). This option only affects what is output. Defaults to prefix.SNP.

`-phased` *`training-sample-file`* Allow the user to specify the name of a phased training genotype data file. This flag may be used for any method. Defaults to prefix.training.

`-unphased` *`training-sample-file`* Allow the user to specify the name of a unphased training genotype data file. This flag may be used when the user uses the NoLD, NoLD-S, LD-RR and GIBDLD method. Defaults to prefix.training.

`-morgan` Allow the user to specify a genetic (Morgan) map.

`-unsort` Allow the user not to sort SNPs from small to large according to genetic position along the chromosome. If the SNP has been sorted by genetic position from small to large along each chromosome, he/she can use the choice and quicken the program. The default is the SNPs will be sorted.

`-phcol` n Allow the user to specify n columns of phenotype traits. The phenotype traits are ignored. Defaults to $n = 0$. This is to allow convenient usage of MERLIN formatted pedigree files.

`-MAF` k Choose SNPs whose minor allele frequency is at least k to be included in the analysis, or else the SNP is excluded. Defaults to $k = 0$.

`-error` k Allow the user to set the genotype error rate. Defaults to $k = 0.005$.

`-rars` Allow the user to remove adjoining redundant SNPs (rars) when they do the computation. If the user didn't use the flag, the IBD estimation are based on all SNPs in the chromosome.

`-mind` k Allow the user to set individual genotype non-missing rate threshold. If a person's genotype non-missing rate is at least reach the threshold, the individual will be analyzed, or else he/she is excluded. Defaults to $k = 0.8$.

`-mincallrate` k Allow the user to set each SNP call rate threshold. If a SNP's call rate is smaller than the threshold, the SNP is excluded and not be analyzed. Defaults to $k = 0.99$.

`-hiddenstates 3|9` Allow the user to set IBD hidden state. There are two choices: 3, 9. Using 3 will quicken the estimation speed and is more appropriate for unrelated pairs or outbred pedigree pairs. Using 9 is appropriate for any kind of relatedness estimation, especially for large pedigree pairs. Default setting is 9;

`-ibc` Allow the user to specify whether to output empirical inbreeding coefficient for the individuals. If you only need the empirical inbreeding coefficient, you can specify pairs of individuals by themselves in the study sample file (`-s` ***study-file***).

`-ibd` n Allow the user to specify which kind of IBD probabilities at each locus, or at specific loci (when `-marker` option is used), to output. If this flag is not used, the IBD sharing probability at each locus is not output. There are eight choices of `0, 2, 3, 4, 9, 30, 40, 90` for the setting of `-hiddenstates 9` and four choices of `0, 2, 3, 30` for the setting `-hiddenstates 3`.

`0` means output IBD sharing at each locus;

`2` means output IBD sharing at each locus, and $\Delta 7$ from 9 condensed identity coefficients ($d7$ coefficient noticed in software SOLAR-formatted MIBD (or IBD) files under the setting of `-hiddenstates 9`) or IBD=2 state probability under the setting of `-hiddenstates 3`;

`3` means output 3 IBD states probabilities (IBD=0, 1, 2);

`4` means output 4 IBD states probabilities (IBD=0, 1, 2, 4), this is the default setting, $n = 4$;

`9` means output 9 condensed identity coefficients;

`30` means the first 3 columns of output are 3 IBD states probabilities, the last column is IBD sharing at each locus;

`40` means the first 4 columns of output are 4 IBD states probabilities, the last column is IBD sharing at each locus;

`90` means the first 9 columns of output are 9 condensed identity coefficients, the last column is IBD sharing at each locus.

`--ibdtxt` Allow the user to specify whether the IBD probability at all loci, or at specific loci is written to a text file with the filename extension of ".ibdtxt" or not. If this flag is not used, the IBD sharing probability is written to a binary file with the filename extension of ".ibdbin".

`-hbd` Allow the user to specify whether to output HBD probability at each locus or at specific loci (when `-marker` option is used).

`--hbdtxt` Allow the user to specify whether HBD probability at each locus is written to a text file with the filename extension of ".hbdtxt" or not. If this flag is not used, the probability is written to a binary file with the filename extension of ".hbdbin'.

`-segment` Allow the user to specify whether to output IBD segments. If this flag is not used, the IBD segment are not output.

`−−min` k Only output IBD segments where each locus's IBD probability is no less than k. Defaults to k= 0.80.

`−−max` k Only output IBD segments where each locus's IBD probability is no larger than k. Defaults to k=1.01.

`−−SNP` n Only output IBD segments which contain at least n continuous SNPs. Defaults to n=1.

`−−length` k Only output IBD segments with at least k kb in length. Defaults to k=500.

`-ibd2segment` Allow the user to specify whether to output IBD2 segments. If this flag is not used, the IBD2 segments are not output.

`−−tmin` k Only output IBD segments where each locus's IBD2 probability is no less than k. Defaults to k= 0.80.

`−−tmax` k Only output IBD segments where each locus's IBD2 probability is no larger than k . Defaults to k=1.01.

`−−tSNP` n Only output IBD2 segments which contain at least n continuous SNPs. Defaults to n=1.

`−−tlength` k Only output IBD2 segments with at least k kb in length. Defaults to k=500.

`-chr` *schr* Allow the user to specify the chromosome *schr* to analyze. Defaults to all available chromosomes in map file.

`-make-grm` Allow the user to set the estimated chromosome-wide kinship coefficients and genome-wide ones output format ("file-prefix_Chr*ChrName*.grm" and "file-prefix_genome.grm"). The lower triangle elements of the coefficients matrix will be saved in text file. The individual identification(ID) in the file can be found in the file of file-prefix.id (columns are family ID and individual ID). The estimated chromosome-wide kinship coefficients and genome-wide ones matrix are not positive definite. If the lower triangle elements contain no missing values, the matrices are transformed into the positive definite ones and the transformed lower triangle elements are also output into the file of "file-prefix_Chr*ChrName*.PositiveDefiniteTransform.grm" and "file-prefix_genome.PositiveDefiniteTransform.grm".

`-nthreads` n Allow the user to specify the number of threads of execution. Usually n is equivalent to the number of processors available on the system. By default, the flag is omitted, the actual number of available processors is used.

`-r` n Allow the user to allocate approximately n megabytes(Mb) of random-access memory (RAM) for the program. Defaults to $n = 1000$.

`-noprint` Allow the user to suppress the display of all output. Defaults to display all output.

Other options:

`-v` print version;

`-h` print help.

5. The users can test the executable program ibdld by running it with the sample input files in these example file folders: Example1, $\cdots$, Example4. In each folder, there are flags and options setting in the file "CommandLineSetting". For example, in Example2, if

you want to estimate the empirical kinship coefficients for the pairs in the file of scan.study with GIBDLD method (based on 10 previous SNPs), you can use the following command line:

```
ibdld -p scan.ped -m scan.map -o scan -unsort -method GIBDLD -ploci 10 -phcol 1 -s scan.study -step 0 -nthreads 10
```

For the command line, you can subdivide it into two consecutive steps,

```
1) ibdld -p scan.ped -m scan.map -o scan -method GIBDLD -ploci 10 -phcol 1 -unsort -step 1 -s scan.study -nthreads 10
2) ibdld -o scan -method GIBDLD -ploci 10 -s scan.study -step 2 -nthreads 10
```

In the above example directory, there are a pedigree file (scan.ped), an identity-coefficient-file (scan.idcoeff), a map file (scan.map, sorted map) and a study pair file (scan.study), the output kinship coefficient file will generate with the prefix of ”scan” automately.

6. If the program of IBDLD succeeds in executing the steps, some sentences will be shown, for example, “Step 1 Finished!” means that the program finished executing the step of estimating background LD parameters. If the sentence “Step 2 Finished!” come out, that means that you have finished all the steps, or else you have to perform the next step. These programs stop if any errors are detected in the running.

4 Input Files

4.1 Required Input Files

Notes on input file conventions:

1. Input files should be saved as plain text files.

2. The input files can be comma-delimited, space-delimited, tab-delimited, and semicolon-delimited, slash-delimited, or mixed use of those, i.e. the entries can be separated by commas, spaces, semicolons, tabs or slash.

3. All input files can contain empty lines, and comment lines: lines starting with # are ignored by ibdld. The following sections describe the format of each input file in more detail.

4.1.1 Map File

The default filename is “prefix.map”. To analyze these SNPs, the software requires information on their chromosomal location, which is usually provided in a map file. It need *mapFile_Array*, each line in the map file describes a single marker and contains exactly 4 columns.

column 1: chromosome
column 2: marker name (rs# or SNP identifier)
column 3: genetic position (in cM)
column 4: base-pair position (bp)

The SNP identifiers can contain any characters except spaces or tabs. The physical base-pair position may be 0 when their physical position is unknown. If there are known physical position, but no genetic position, you can approximately transform the physical position into

the genetic position according to the equation 1Mb=1cM. **The `Map File` must contain as many markers as those in the `Pedigree File`.**
For example:

chr22	rs9605923	0.845470443	14550436
chr22	rs5747999	0.888157955	14560203
chr22	rs11089263	0.907050668	14715506
chr22	rs16981741	1.260310733	15272858
chr22	rs175148	1.260583644	15284080

The above format is appropriate for the flag `-m`. When the third column is **genetic distance** between the adjoining SNPs, not **genetic position**, the user should be use the flag `-m_int`. Their SNPs should be sorted by Physical position from small to large. If the unit is Morgan, you can use the flag ”-morgan”.
For example:

1	rs28705211	0.0151734	890368
1	rs9777703	0.00779571	918699
1	rs3121567	0.00402622	933331
1	rs3934834	0.0171594	995669
1	rs3737728	0.00972966	1011278

4.1.2 Pedigree File

The pedigree file follows Linkage or MERLIN similar format (with the -p flag) or PLINK similar format (with -plink). The default filename is “prefix.ped”. The first five columns are mandatory:
column 1: family ID
column 2: individual ID
column 3: paternal ID
column 4: maternal ID
column 5: sex

The IDs are alphanumeric, the combination of family and individual ID should uniquely identify a person. **The individuals in the same family must be grouped together and make sure parents must come before their children in the file.** If the paternal (or maternal ID) is unknown, they are denoted as “0”. The sex is coded as 2 (female) and 1 (male) or 0 (unknown).

Alleles can be any character (e.g. 1, 2, 3, 4 or A, C, G, T or anything else) except 0 or ‘N’ (‘n’) which denotes the missing genotype character. The characters are case insensitive, for example, ‘a’ and ‘A’ are the same. **All markers should be biallelic.** The following are all valid genotype entries 1/1 (homozygote for allele 1), 0/0 (missing genotype), and 3/4 (heterozygote for alleles 3 and 4). The A/A, A/C and C/C are also valid genotypes. Following the genotype columns are optional trait column which may have any alphanumeric value(such as “x”). The trait data is ignored, such as Format 1 and Format 2:

Format 1

A4	1	0	0	1	0	0	0	0	0	0	1
A4	2	0	0	2	0	0	0	0	0	0	x
A4	3	1	2	2	2	2	1	2	1	2	2.5
A4	4	1	2	2	1	2	1	2	1	2	2.3
A4	5	1	2	2	2	2	1	1	1	1	3.5
A4	6	1	2	1	1	2	1	2	1	2	4.2
B5	1	0	0	1	0	0	0	0	0	0	2.1
B5	2	0	0	2	0	0	0	0	0	0	-1.5

Format 2

1	11	0	0	0	T/T	T/T	A/C
1	12	0	0	0	T/T	T/T	A/C
1	13	0	0	0	A/A	T/T	A/A
1	14	0	0	0	A/A	T/T	A/A
1	15	11	13	0	n/n	n/n	n/n
1	16	11	12	0	A/T	T/T	n/n
1	17	11	12	0	A/T	T/T	A/C
1	18	14	12	0	A/A	T/T	A/C
1	19	18	16	0	A A	T T	A C

The following pedigree file is in PLINK format. For PLINK binary PED files, e.g. test.fam, test.bim and test.bed, the user can read PLINK user manual for details.

Format 3

A4	1	0	0	1	1	0	0	0	0	0	0
A4	2	0	0	2	x	0	0	0	0	0	0
A4	3	1	2	2	2.5	2	2	1	2	1	2
A4	4	1	2	2	2.3	1	2	1	2	1	2
A4	5	1	2	2	3.5	2	2	1	1	1	1
A4	6	1	2	1	4.2	1	2	1	2	1	2
B5	1	0	0	1	2.1	0	0	0	0	0	0
B5	2	0	0	2	-1.5	0	0	0	0	0	0

4.2 Optional Input Files

4.2.1 Identity Coefficient File

The file is required for the method NoLD, NoLD-S, LD, LD-RR, but not needed for GIB-DLD. The default filename is “prefix.idcoeff”. This file contains identity coefficients for every pair of individuals within each family (including pairs of same individuals) for whom the user wishes to compute IBD sharing probabilities.

The identity coefficient file has the following columns:
column 1: family ID
column 2: individual 1 ID
column 3: individual 2 ID
column 4: Δ_1
⋮ ⋮
column 12: Δ_9
where Δ_x is Jacquard's xth condensed identity coefficient. (Ken Lange's book, Mathematical and Statistical Methods for Genetic Analysis, has a good description of condensed identity coefficients). For example,

1	3	3	0	0	0	0	0	0	1	0	0
1	3	4	0	0	0	0	0	0	0.25	0.5	0.25
1	3	5	0	0	0	0	0	0	0.25	0.5	0.25
1	3	6	0	0	0	0	0	0	0.25	0.5	0.25
1	4	4	0	0	0	0	0	0	1	0	0
1	4	5	0	0	0	0	0	0	0.25	0.5	0.25
1	4	6	0	0	0	0	0	0	0.25	0.5	0.25

The software program that can be used to compute identity coefficients based on pedigree information is the "IdCoefs 2.1" software by Abney M. (2009), which can be found at `http://home.uchicago.edu/$\sim$abney/abney_web/Software.html`.

Notice, if the output file has no family ID column, you have to insert the family ID into the first column for each pair of individuals on each line.

4.2.2 Training Sample File

The training sample file will be used to estimate background LD when the option `-unphased(or -phased) TSFile` is used. The use of a training sample is optional. If the study sample size is small, it is better to use the additional training sample. If a training sample isn't denoted, the genotype data in the pedigree file will be considered as a training sample to estimate background LD. The default filename is "prefix.training". The file is organized in SNPs by individuals format, so every row represents a SNP and every two columns denote a genotype for an individual. The 1st column indicates chromosome, the 2nd column is the marker name. Entries with '0' or 'N' ('n') represent missing genotype character at a SNP. Alleles in a genotype may be separated by a /, | or space character. In the example below, the line that begins with a # will be ignored by the program.
The training genotype data may be either unphased or phased.

`Unphased data`

Each pair of columns (beginning with column 3) gives the unphased genotype for each individual. If the file in the above example contains unphased data for 4 individuals, the 1st individual has genotypes A/T,C/G, T/A, the 2nd individual has genotypes T/A, G/C, A/A for markers SNP1, SNP2 and SNP3, respectively.

#chromosome	Marker	Ind1	Ind2	Ind3	Ind4
chr1	SNP1	A/T	T/A	T/A	A/T
chr1	SNP2	C/G	G/C	C/C	G/C
chr1	SNP3	T/A	A/A	T/A	A/T

`Phased data`

#chromosome	Marker	Ind1	Ind2	Ind3	Ind4
chr1	SNP1	A|T	T|A	T|A	A|T
chr1	SNP2	C|G	G|C	C|C	G|C
chr1	SNP3	T|A	A|A	T|A	A|T

Each pair of column (beginning with column 3) gives the phased genotype for each individual. If the file in the above example contains phased data for 4 individuals, the 1st individual has haplotypes ACT and TGA, the 2nd individual has haplotypes equal to TGA and ACA for markers SNP1, SNP2, and SNP3, respectively.

`LD` requires haplotype frequencies, which are computed from phased genotype data (included in training sample file, `-phased TSFile`). If the genotype data is unphased, you can obtain phased haplotypes by using appropriate software such as fastPHASE (Scheet and Stephens, 2006) or by directly downloading the phased data of same population from International HapMap Project website (The International HapMap Consortium, 2007) (http://hapmap.ncbi.nlm.ni

`NoLD` and `NoLD-S` need allele frequencies, `LD-RR` and `GIBDLD` require ridge regression coefficients, all of them are inferred from either the above phased genotype data (`-phased TSFile`) or unphased genotype data (included in the pedigree file or training sample file, `-unphased TSFile`). If the training sample file option is specified, the allele frequencies or the ridge regression coefficients are estimated based on the training sample. If the option is not specified, the genotypes in the pedigree file are used to estimate the allele frequencies or the ridge regression coefficients. It is not a good choice to use a separate training file unless the size of study sample is small. By combining a small study sample with a large training sample, there will increase the result estimate accuracy.

Notice: The allele types at a marker in the training sample file must be the same as the allele types at that marker in the pedigree file, or else the SNP is removed from the analysis.

4.2.3 Study Sample File

The default filename is "prefix.study". The study sample file is used (`-s studyFile`) when the user is interested in specific pairs' IBD sharing probabilities. The file has two formats. One is that each line contains three or four columns. The program will limit these pairs analysis. The other format is to have two columns, in which case IBD sharing probabilities will be computed for every possible pair of IDs in each family (including pairs of same individuals). For the method of NoLD, NoLD-S, LD and LD-RR, the above pairs must have the pedigree-based condensed identity coefficients. The pairs having no the coefficients will be discarded. If a study sample file option is not specified, the program will compute the

IBD sharing probabilities for each pair with pedigree based condensed identity coefficients. For the method GIBDLD, if the option is not specified, the program will compute the IBD sharing probabilities for each possible pairs with the enough SNP genotypes.

For example,

Format 1 in 3 columns (Common family ID between individual 1 and individual 2, Individual 1 ID, Individual 2 ID)

```
1  3  4
1  5  6
1  5  5
1  3  5
1  4  5
```

Format 2 in 4 columns (Family ID of individual 1, Individual 1 ID, Family ID of individual 2 , Individual 2 ID).

```
1  3  1  4
1  5  1  6
1  5  1  5
1  3  1  5
1  4  1  5
```

In the above file, IBD sharing probabilities will be computed for the same five pairs as those in format 1.

Format 3 in 2 columns (Family ID, Individual ID)

```
1  3
1  4
1  5
1  6
```

In the above file IBD sharing probabilities will be computed for 10 pairs, which consist of 1 3 1 3, 1 3 1 4, 1 3 1 5, 1 3 1 6, 1 4 1 4, 1 4 1 5, 1 4 1 6, 1 5 1 5, 1 5 1 6, 1 6 1 6 pair (including pairs of same individuals).

If the users want to only estimate inbreeding coefficients (`-ibc`) or HBD probabilities (`-hbd`), they just specify pairs of same individuals.

For example,

Format 1 in 3 columns (Shared same family ID between individual 1 and individual 2, Individual 1 ID, Individual 2 ID).

```
1  3  3
1  4  4
1  5  5
1  6  6
```

Format 2 in 4 columns (Family ID of individual 1, Individual 1 ID, Family ID of individual 2, Individual 2 ID).

```
1  3  1  3
1  4  1  4
1  5  1  5
1  6  1  6
```

In the above file, HBD probabilities or inbreeding coefficients will be computed for 4 individuals.

4.2.4 Marker List File

The default filename is “prefix.SNP”. This file is used (`-marker markerFile`) when the user wants to output IBD sharing probabilities for certain SNPs. The IBD sharing probabilities will still be computed using all the genotypes in the pedigree file, this option only reduces the number of IBD sharing probabilities output, and output those SNPs listed in this file. It can be used when the user selects one of NoLD, LD, LD-RR or GIBDLD. In this file, each line describes the chromosome and a single marker in two columns:
column 1: chromosome
column 2: rs# or SNP identifier
For example,

```
chr22  SNP1
chr22  SNP3
chr22  SNP5
chr22  SNP8
```

5 Output Files

5.1 Intermediate Output Files

Intermediate output files are created by IBDLD as output from one step to be used as input to a subsequent step. User can typically ignore these files. They are listed here for reference.

5.1.1 Method Statement File [.mthd]

The default filename is named “prefix.mthd”. The output file, which is used as an input file in `step 2`, is output when the user executes any method in `step 1`. The method statement file is used as a method checking so as to make sure that the user uses the same method in `step 2` as the one in `step 1`. If it fails, the user must recompute from `step 1`, or execute the method shown in the file for step 2. In the file, the first row denotes the method used and the parameter setting in `step 1`. The following rows show the available chromosome for analyzing IBD sharing probabilities. For example,

```
LD    20  2
chr8
chr10
```

The above example shows that the user has used LD method, which conditions on a single SNP with the highest correlation to the current SNP from among the previous 20 markers within 2.0 cM along the chromosome(`-ploci 20 -dist 2`) in `step 2`, he/she can analyze IBD sharing probabilities for chromosome chr8 and chr10.

Background LD Parameter File [.gtype] and Ridge Regression Coefficient File [.reg] are in binary format, here we don't describe them. The file of "prefix.aInd" denotes these available individuals with enough non-missing rate genotype data who can be used to further be analyzed.

5.2 Final Output Files

If any following final output file size is larger than 100MB, it will be compressed and named as "file_name.gz".

5.2.1 Log File [.log]

A log file includes details of the run parameters used and any warnings generated. When sending in a bug report, it is important to include the `Log File` as an attachment. The file is named "prefix.log".

These following result files are written when `step 2` is executed, and will overwrite any existing related ibd, kinship and shared segment file if you set same output filename prefix (`-o prefix`). If you want to avoid this, you have to rename these filenames or set different main output filename prefix (`-o prefix`) before you run `step 2`.

5.2.2 Chromosome-wide Identity Coefficient File of PRIMAL [.primal]

PRIMAL (PedigRee-based IMputation Algorithm, http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004139) is an accurate phasing and imputation algorithm for related individuals. The software was originally written to impute the genomes of 1,317 members of the South Dakota Hutterites population from 98 whole-genome Hutterite sequences. The IBDLD can provide the Identity Coefficient File input of PRIMAL(https://github.com/orenlivne/ober/wiki/File-Formats). The 13 columns are listed as :

column 1: family ID of individual 1

column 2: individual 1 ID

column 3: family ID of individual 2

column 4: individual 2 ID

column 5: *lam*,
column 6: Δ_1
$\vdots$ $\vdots$
column 13: Δ_9
lam is the estimated recombination transition rate. Example,

```
1  2592  1  2592  5.2885  0.0181     0        0        0        0        0       0.9819    0        0
1  2592  1  2882  0.0863  0.0001  0.00063  0.0026   0.0148   0.0054   0.0330   0.0019   0.1434   0.7982
1  2592  1  2912  0.0765  0.0001  0.0006   0.0026   0.0148   0.0054   0.0330   0.0019   0.1434   0.7982
1  2592  1  3332  0.0482  0.0001  0.0004   0.0016   0.0161   0.0013   0.0246   0.0012   0.0905   0.8644
```

5.2.3 Chromosome-wide Empirical Kinship Coefficient File [.kinship]

The file describe the information about the estimated chromosome-wide empirical kinship coefficient for each pair of individuals. The file is named as "prefix_*chromosome*.kinship" by default. For the two "-hiddenstates" flag choices of "-hiddenstates 3" and "-hiddenstates 9", two output files with 9 and 15 columns are produced, respectively. The first 5 columns are same between them. The five columns are listed as :
column 1: family ID of individual 1
column 2: individual 1 ID
column 3: family ID of individual 2
column 4: individual 2 ID
column 5: number of SNPs on the chromosome
For the flag "-hiddenstates 9", column 6-14: 9 chromosome-wide condensed identity coefficients (Δ_1, Δ_2, Δ_3, Δ_4, Δ_5, Δ_6, Δ_7, Δ_8, Δ_9 in order); column 15: chromosome-wide empirical kinship coefficient.

For example, in the file of "prefix_chr22.kinship", you can read the following lines,

```
11  204  1  11  1001  0.001  0    0      0      0.001  0.001  0.950  0.003  0.043  0.478
11  204  2  21  1001    0    0    0    0.017    0    0.013  0.181  0.476  0.313  0.210
11  204  3  31  1001    0    0    0      0      0      0    0.350  0.030  0.619  0.183
11  204  4  41  1001    0    0  0.004    0      0      0    0.015  0.897  0.084  0.234
11  204  5  51  1001    0    0    0      0    0.003  0.001    0    0.989  0.007  0.249
11  204  6  61  1001    0    0    0      0      0      0    0.377  0.265  0.358  0.255
11  204  7  71  1001    0    0    0      0      0      0    0.002  0.857  0.141  0.215
```

In the file, it describes chromosome-wide condensed identity coefficients and empirical kinship coefficients estimates between the individual 204 (family 11) and other individuals based on 1001 SNPs information along Chromosome chr22.

For the flag "-hiddenstates 3", column 6-8: 3 chromosome-wide IBD states probabilities ($IBD = 0$, $IBD = 1$, $IBD = 2$ in order), column 9: chromosome-wide empirical kinship coefficients.

For example, in the file of "prefix_chr10.kinship", you can read the following lines,

11	208	1	11	1001	0.6293	0.3707	0	0.09268
11	208	2	21	1001	0.9504	0.04945	0.0001637	0.01244
11	208	3	31	1001	0.6078	0.392	0.0002098	0.0981
11	208	4	41	1001	0.9732	0.02677	0	0.006695
11	208	5	51	1001	0.7548	0.2452	0	0.06131
11	208	6	61	1001	1	0	0	0
11	208	7	71	1001	0.7386	0.2237	0.03776	0.0748

In the file, it describes 3 chromosome-wide IBD states probabilities and empirical kinship coefficients estimates between the individual 208 (family 11) and other individuals based on 1001 SNPs information along Chromosome chr10.

5.2.4 Empirical Inbreeding Coefficient File [.ibc]

This file is created only when `-ibc` option is used. This file gives the estimated empirical inbreeding coefficient by chromosome for each individual. The file is named as "prefix_*chromosome*.ibc" by default, which has the following columns:
column 1: family ID of individual 1
column 2: individual 1 ID
column 3: number of SNPs on the chromosome
column 4: empirical inbreeding coefficient
For example,

1	A1	1001	0.0007
1	A2	1001	0
1	A3	1001	0.0362
1	A4	1001	0.0562

5.2.5 SNP IBD File [.ibdbin or .ibdtxt]

The SNP IBD file contains the information about 9 condensed identity coefficients, 3 (or 4) IBD states probabilities, IBD sharing of each SNP or specified SNPs (i.e. when `-marker` option is used). The output files is named as "prefix_*chromosome*.ibdbin" (in binary format) or "prefix_*chromosome*.ibdtxt" (in text format) . For the flag "-hiddenstates 9", there are eight choices of `0, 2, 3, 4, 9, 30, 40, 90` in total, which correspond to proportion of alleles shared IBD at a SNP, IBD sharing at each SNP and $\Delta 7$ from the 9 condensed identity coefficients, 3 IBD states probabilities (IBD=0, 1, 2), 4 IBD states probabilities (IBD=0, 1, 2, 4), 9 condensed identity coefficients, and the combination of 3, 4, or 9 with 0. For the flag "-hiddenstates 3", there are four choices of `0, 2, 3, 30` in total, which correspond to proportion of alleles shared IBD at a SNP, IBD sharing at each SNP and $IBD = 2$ state probabilities, 3 IBD states probabilities (IBD=0, 1, 2), and the combination of 3 with 0. All defaults to 0. Among them, only one can be selected. These choices correspond to eight or three kinds of files, they share a common format for the first row, it describes all SNPs which had been analyzed. It begins with IBD output choice, pairs of number, chromosome and number of SNPs, and the following columns are SNPs names (rs# or SNP identifier).

column 1: choice type

column 2: number of pairs

column 3: indicating chromosome

column 4: number of SNPs on the chromosome

column 5: marker1

column 6: marker2

column 7: marker3

The following rows describe all pairs's locus IBD information. The following lines begins with ID (include family ID, individual ID) information of a pair of individuals, i.e. the first four columns are

column 1: family ID of individual 1

column 2: individual 1 ID

column 3: family ID of individual 2

column 4: individual 2 ID

The following columns are IBD information for each SNP, for example, for the choice of "9", the following columns are the condensed identity coefficients (Δ_1, Δ_2, Δ_3, Δ_4, Δ_5, Δ_6, Δ_7, Δ_8, Δ_9) of each SNP in order of the SNP on the first line (such as SNP1, SNP2, SNP3). This is the format for the first pair. The other pairs follow the lines are described with same format.

```
9   3   chr22   2   rs1048   rs7831
1   A1    1     A2    0        0      0  0.002  0  0.004  0  0  0.994  0  0  0  0.002  0  0  0  0  0.998
1   A2    1     A3    0        0      0  0.002  0  0.004  0  0  0.994  0  0  0  0.002  0  0  0  0  0.998
1   A3    1     A4    0        0      0  0.002  0  0.004  0  0  0.994  0  0  0  0.002  0  0  0  0  0.998
```

For the choice of "4", the following rows describe four IBD state probabilities (IBD= 0, 1, 2, 4) in the order of SNPs (SNP1, SNP3) for all pairs.

```
4  3  chr22  2  SNP1   SNP3
1  7    1    8  0.049  0.270  0.681  0  0.047  0.269  0.684  0
1  8    1    9  0.052  0.268  0.680  0  0.055  0.271  0.674  0
1  5    1    7  0.052  0.268  0.680  0  0.055  0.271  0.674  0
```

For the choice of "0", the following rows describe IBD sharing at each locus in the order of SNPs (SNP1, SNP3) for all pairs.

0	3	chr22	2	SNP1	SNP3
1	7	1	8	0.408	0.409
1	8	1	9	0.407	0.405
1	5	1	9	0.407	0.405

For the combination, such as the choice of “40”, the first 4 columns are the same as the choice “4”, the last column is proportion of alleles shared IBD at a SNP, the format is same as above for each row.

40	3	chr22	2	SNP1	SNP3								
1	7	1	8	0.049	0.270	0.681	0	0.408	0.047	0.269	0.684	0	0.409
1	5	1	8	0.052	0.268	0.680	0	0.407	0.055	0.271	0.674	0	0.405
1	5	1	9	0.052	0.268	0.680	0	0.407	0.055	0.271	0.674	0	0.405

As the user set the flag ”–ibdtxt” for IBD file output setting, the IBD file will output in text format and it will use a lot of space. In order to save space, the files can be saved in much more compress binary format.

5.2.6 HBD File [.hbdbin or .hbdtxt]

The HBD file is created only when the `-hbd` option is specified. It is named as“prefix_*chromosome*.hbdbin” (in binary format) or “prefix_*chromosome*.hbdtxt” (in text format). The file contains the information about estimated HBD probabilities of each locus or specified loci (i.e. when `-marker` option is used) for a given individual. They have the following columns for the first line, the markers are named as rs# or SNP identifier.

column 1: number of individuals
column 2: indicating chromosome
column 3: number of markers on the chromosome
column 4: SNP1
column 5: SNP2
column 6: SNP3

The following rows describe HBD information for all individuals. The following lines begin with an individual's ID (include family ID, individual ID) information, i.e. the first two columns are

column 1: family ID of individual
column 2: individual ID

The following columns are HBD information of each marker for the individual,

column 3: HBD probability on SNP1
column 4: HBD probability on SNP2
column 5: HBD probability on SNP3

This is the format for the first individual. The other individuals follow the lines are described with same format. For example,

2	chr22	3	rs10488368	rs7831661	rs4045956
1	A2	0.30	0.50	0.20	
1	A5	0.30	0.50	0.20	

As the user set the flag ”–hbdtxt” for HBD file output setting, the HBD file will output in text format and it will use a lot of space. In order to save space, the files can be saved in much more compress binary format.

5.2.7 Shared Segments File [.segment or .ibd2segment]

The segment files consist of two types of IBD segment file, i.e. IBD segment and IBD2 segment, their files are named as "prefix_*chromosome*.segment" and "prefix_*chromosome*.ibd2segment" by default, respectively. Both formats are complete same, so we just describe one IBD segment file format.

A shared segment file is output when the user selects the option `-segment` to detect extended chromosomal segmental IBD sharing in pairs of individuals. It has the following columns:

column 1: family ID of individual 1
column 2: individual 1 ID
column 3: family ID of individual 2
column 4: individual 2 ID
column 5: indicating chromosome
column 6: start physical position of segment_1 (bp)
column 7: end physical position of segment_1 (bp)
column 8: start SNP of segment_1
column 9: end SNP of segment_1
column 10: number of SNPs within this segment_1
column 11: physical length of segment_1 (kb)
column ⋮
column 24 start physical position of segment_n (bp)
column 25 end physical position of segment_n (bp)
column 26 start SNP of segment_n
column 27 end SNP of segment_n
column 28 number of SNPs within this segment_n
column 29 physical length of segment_n (kb)

For example,

1	A1	1	A1	chr8	14560203	15607600	rs10488368	rs7003378	69	1047.4	···	20	21.43
1	A1	1	A2	chr8	14560203	15460378	rs10488368	rs6984685	33	900.175	···	14	14.619
1	A2	1	A3	chr8	15476864	15639954	rs1669703	rs6983835	46	163.09			

6 Supplementary PERL Command

If the command line argument of *−chr* is used in the command line of *./ibdld*, the chromosome based kinship coefficients (or primal coefficients) will be obtained, however the genome-wide kinship coefficients (or primal coefficients) won't be estimated. If the user is still interested in them, he/she can execute `perl` script to estimate them after these chromosome based coefficients had been obtained. All these coefficients files must be put under the same file folder. Below is the command and all possible command line arguments.

```
perl ./ObtainGenomeWideKinship_Primal.pl \
 [−o file−prefix ]\
 [−−hiddenstates 3|9] \
 [−−exchr chrID] \
 [−−filetype kinship|primal]
```

The command line arguments:

`-o` *file-prefix* Allow the user to set a common filename prefix for all chromosome based kinship (or primal) coefficients input files. The genome−wide kinship (or primal) coefficient files are named as *file-prefix*_genome.kinship (or *file-prefix*_genome.primal), which are put in the same file folder of chromosome based coefficient files.

`--hiddenstates 3|9` Allow the user to set IBD hidden state. There are two choices of 3 and 9. The default setting is 9.

These two above command line arguments are same as those in the command line of *./ibdld*.

`--filetype kinship|primal` Allow the user to set file type. There are two choices of kinship and primal. The default setting is kinship. If the user want to obtain genome-wide primal coefficients, he/she can set the primal.

`--exchr` *chrID* Allow the user to set chromosome ID. The default setting is 0. By using the default setting, the whole genome-wide kinship/primal coefficients are estimated with all available chromosome based coefficient files in the setting file folder. If any chromosome ID is set, its corresponding chromosome based kinship or primal file is excluded while the remainder available chromosomes based kinship or primal files are used to estimated to genome-wide kinship(primal) coefficient files .

Example1:

perl ./ObtainGenomeWideKinship_Primal.pl −o Example2/scan −−hiddenstates 3 −−exchr 14 −−filetype kinship

Based on all available chromosome based kinship files with filename prefix of scan in the file folder of Example2/, exclude chromosome 14 based kinship coefficient files, the genome-wide kinship coefficients are estimated using the remainder of them.

Example 2:

perl ./ObtainGenomeWideKinship_Primal.pl −o Example2/scan −−hiddenstates 3 −−filetype primal

Based on all available chromosome based primal files with filename prefix of scan in the file folder of Example2/, the genome-wide primal coefficients are estimated using all of them.

7 Tips

1. The program will stop if errors are detected in the formats of any input files. Please read Section 4 carefully and make sure the input files are in the correct format and have concordant information.

2. The program is divided into two steps, you can run `step 0` for the full process. If a cluster-computing environment is available, such jobs are easily parallelized, i.e. you can run `step 1` at first, then you run `step 2`.

3. If you analyze many pairs in multiprocessing environment or in a single processing environment, such as more than 50,000, you appropriately enlarge the RAM (`-r n`) according to the equation `1M=100 pairs` to increase speed. In cluster-computing environment, the user can use the `-s studyFile` to divide total pairs into many groups, especially for the SNPs along a long chromosome, each of which constitute a study sample file, and then run `step 2` for multiple groups at the same time so as to improve the speed.

4. The program computes IBD probability based on chromosome-wide data. If data are available on multiple chromosomes, we suggest analyzing each chromosome separately, and use the flags (`-chr chromosome`) to assign the computation of different chromosomes to different processors so as to reduce running time. At the same time, the user can judge if they need split the total pairs into many smaller subgroups to do parallel computation based on the output show how long it will spend finishing computation. If the remainder time is too long , it is better to separate all the pairs to several groups to do parallel computation.

5. If the user analyzes huge whole genome data (such as more than 2,000,000 SNPs), he/she should create a separate `Pedigree File` and `Map File` for each chromosome so as to make data input easier. The user can run `step 1` for all chromosome's `Pedigree Files` and `Map Files`. After the step finished, the user runs `step 2` to estimates IBD sharing probabilities. When running `step 2`, if the user chooses a different filename prefix at `step 1` for `-prefix` in `step 2`, he/she should use `-bgld` to specify the `Method Statement File`, `Background LD Parameter File` and `Ridge Regression Coefficient File` output from `step 1`.

6. If there are no documental pedigree information, GIBDLD is the only choice. If there is a documental pedigree, it is better to use LD-RR rather than GIBDLD, the former usually give more accurate IBD sharing estimation.

8 References

Scheet P, Stephens M. (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase.

American Journal of Human Genetics, 78: 629 -644.

Bernardo R, (1993). Estimation of coefficient of coancestry using molecular markers in maize. Theor. Appl. Genet. 85:1055-1062.

Abney M (2009) A graphical algorithm for fast computation of identity coefficients and generalized kinship coefficients. *Bioinformatics*, 25:1561-1563

The International HapMap Consortium. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449: 851-861.

Han L, Abney M (2011) Identity by descent estimation with dense genome-wide genotype data. *Genetic Epidemiology*, 35 : 557-567

Han L, Abney M (2013). Using identity by descent estimation with dense genotype data to detect positive selection.*European Journal of Human Genetics*, 21, 205-211.

9 Acknowledgements

This work was supported by National Institutes of Health grant HG02899.