Refined IBD documentation

Brian L. Browning
Department of Medicine
Division of Medical Genetics
University of Washington

December 23, 2017

Contents

Contents......i

1 Introduction......1

2 New features......1

3 Command line arguments......1

3.1 Arguments for specifying data......1

3.2 Analysis parameters......2

4 Output files......3

1 Introduction

Refined IBD is a software program for detecting identity-by-descent (IBD) segments within and between individuals. The program is freely available and can be downloaded from the Refined IBD web site:

http://faculty.washington.edu/browning/refined-ibd.html

If you use the Refined IBD program and publish your analysis, please report the program version and cite the article describing the Refined IBD algorithm:

> B L Browning and S R Browning (2013). Improving the accuracy and efficiency of identity by descent detection in population data. Genetics 194(2):459-71. doi:10.1534/genetics.113.150029

2 New features

This version of Refined IBD is a stand-alone program with four improvements to the version of Refined IBD in Beagle 4.1:

1) Parameter have been simplified (see Section 3).
2) All length parameters (**window**, **ibd**, and **trim**) are in cM units.
3) The overlap between adjacent sliding windows is set automatically.
4) The .ibd and .hbd output files have an additional column containing IBD segment lengths.

3 Command line arguments

The Refined IBD program is run using Java version 1.8 (or a later version). Enter "java -version" at the unix command prompt to check the version of java installed on your computer. The most recent Java interpreter can be downloaded from **www.java.com**. Attempting to run Refined IBD with an earlier version of Java will produce an "Unsupported Class Version" error.

To run Refined IBD, enter the following command at the command prompt:

java –Xss5m –Xmx[*GB*]g –jar refined-ibd.[*version*].jar [*arguments*]

where [*GB*] is the size of the memory pool in gigabytes (e.g. –Xmx50g), [*version*] is the Refined IBD version code (eg. "23Dec17.c32"), and [*arguments*] is a space separated list of parameter values, each having the format **parameter=value**. The –Xss5m argument can be omitted unless there is a stack overflow error. The –Xmx[*GB*]g argument can be omitted unless there is an out-of-memory error.

There are only two required command line arguments: the **gt** argument to specify the input VCF file and the **out** argument to specify the output file prefix. A third parameter, the **map** parameter, is recommended, but not required. Other parameters have sensible default values.

3.1 Arguments for specifying data

❖ **gt=**[*file*] specifies a Variant Call Format (VCF) file with a GT (genotype) format field for each record. Each genotype must have two phased, non-missing alleles separated by '|'. If your genotype data has any missing or unphased alleles, you can run Beagle, and use the

Beagle output VCF file as the input VCF file. Refined IBD assumes that any input file that has a name ending in “.gz” is compressed with gzip or bgzip. Male non-pseudoautosomal X-chromosome genotypes must be coded as homozygous diploid genotypes.

- **out=**[*string*] specifies the prefix for the output filenames (see Section 4). The prefix may be an absolute or relative filename. It cannot be a directory name.
- **map=**[*file*] specifies a PLINK format genetic map with cM units. HapMap GrCh36 and GrCh37 genetic maps in PLINK format are available for download from the Refined IBD web site. Refined IBD will use linear interpolation to estimate genetic positions between map positions. If no genetic map is specified, Refined IBD will assume a constant recombination rate of 1 cM per Mb.
- **chrom=**[*chrom*]:[*start*]-[*end*] specifies a chromosome interval: [*chrom*] is the chromosome identifier in the input VCF file and [*start*] and [*end*] are the starting and ending positions. The entire chromosome, the beginning of the chromosome, and the end of the chromosome can be specified by **chrom=**[*chrom*], **chrom=**[*chrom*]:-[*end*], and **chrom=**[*chrom*]:[*start*]-respectively.
- **excludesamples=**[*file*] specifies a file containing samples (one sample identifier per line) to be excluded from the analysis.
- **excludemarkers=**[*file*] specifies a file containing markers (one marker per line) to be excluded from the analysis. Each line of the file can be either an identifier from a VCF record’s ID field or a genomic coordinate written as CHROM:POS.

3.2 Analysis parameters

- **nthreads=**[*positive integer*] specifies the number of threads of execution. If no **nthreads** parameter is specified, the **nthreads** parameter will be set equal to the number of CPU cores on the host machine.
- **window=**[*positive number*] specifies the cM length of the sliding marker window (default: **window=40.0**). Genetic length is determined by the genetic map (see the **map** argument). The overlap between adjacent windows will be twice the length of the **ibd** parameter. Increasing/decreasing the **window** parameter will increase/decrease the memory required for the analysis.
- **lod=**[*positive number*] specifies the minimum LOD score for reported IBD segments (default: **lod=3.0**).
- **length=**[*positive number*] specifies the minimum cM length in cM for reported IBD segments (default: **ibd=1.5**). Genetic length is determined by the genetic map (see the **map** argument).
- **trim=**[*non-negative number*] specifies the cM trimmed from the end of a shared haplotype when calculating the IBD LOD score (default: **trim=0.15**). Genetic length is determined by the genetic map (see the **map** argument).
- **scale=**[*non-negative number*] specifies the scale parameter used to build the haplotype frequency model for IBD analysis. If no **scale** parameter is specified or if **scale=0**, the scale parameter for the IBD analysis will be set to $\max\left\{2, \sqrt{[\text{sample size}]/100}\right\}$, which we have found to work well for outbred populations.

4 Output files

There are three output files. The **log** file gives a summary of the analysis that includes the Refined IBD version, the command line arguments, and the running time. The **hbd** file contains detected homozygous-by-descent segments within each individual, and the **ibd** file contains identity-by-descent segments between individuals. The **hbd** and **ibd** files are gzip-compressed and can be uncompressed with the unix gunzip utility.

Each line of an **hbd** and **ibd** output file represents one segment and contains nine tab-delimited fields:

1)	First sample identifier	
2)	First sample haplotype index	(1 or 2)
3)	Second sample identifier	
4)	Second sample haplotype index	(1 or 2)
5)	Chromosome	
6)	Starting genomic position	(inclusive)
7)	Ending genomic position	(inclusive)
8)	LOD score	(larger values indicate greater evidence for IBD)
9)	cM length of IBD segment	