BEAGLE 3.3

Brian L. Browning

Department of Medicine

Division of Medical Genetics

University of Washington

26 December 2010

Contents

Contents i
1 Introduction 1
1.1 Citing BEAGLE 1
1.2 Files in the BEAGLE software distribution 3
2 BEAGLE file formats 3
2.1 Genotypes file format 3
2.2 Genotype likelihoods file format 7
2.3 Genotype probabilities file format 7
2.4 Markers file format 7
3 Inferring haplotype phase and missing data with BEAGLE 8
3.1 Quick start guide 8
3.2 Command line arguments 9
3.2.1 Arguments for specifying files 10
3.2.2 Other phasing arguments 10
3.2.3 fastIBD arguments 12
3.2.4 IBD arguments 12
3.2.5 HBD arguments 13
3.2.6 Advanced options not intended for general use 14
3.3 Output files 14
3.3.1 log file [.log] 14
3.3.2 Phased file [.phased.gz] 14
3.3.3 Genotype probabilities file [.gprobs.gz] 14
3.3.4 Genotype dosage file [.dose.gz] 15
3.3.5 Allelic R^2 file [.r2] 15
3.3.6 fastIBD file [.fibd.gz] 15
3.3.7 IBD file [.ibd] 16
3.3.8 HBD file [.hbd.gz] 16
3.3.9 Sampled haplotype file [.*k*.sample.gz] 16
4 Association testing with BEAGLE 17
4.1 Quick start guide 17
4.2 Command line arguments 18
4.2.1 Argument for specifying input data 18

4.2.2 Arguments for building the model 18
4.2.3 Arguments for association testing 19
4.2.4 Advanced options not intended for general use 20
4.3 Output files 20
4.3.1 The log file [.log] 20
4.3.2 The P-value file [.pval] 20
4.3.3 The null distribution file [.null] 22
4.3.4 The model file [.dag.gz] 22
5 Using BEAGLE with large data sets 23
5.1 Memory management 23
5.2 Genomewide association studies 24
6 Example BEAGLE analyses 24
6.1 Inferring haplotype phase and missing data 24
6.2 Association testing 25
7 Utility programs 26
7.1 BEAGLE Utilities web site 26
7.2 pseudomarker 26
7.3 cluster2haps 27
8 References 30

1 Introduction

BEAGLE is a software program for imputing genotypes, inferring haplotype phase, and performing genetic association analysis. BEAGLE is designed to analyze large-scale data sets with hundreds of thousands of markers genotyped on thousands of samples. BEAGLE can

- phase genotype data (i.e. infer haplotypes) for unrelated individuals, parent-offspring pairs, and parent-offspring trios.
- infer sporadic missing genotype data.
- impute ungenotyped markers that have been genotyped in a reference panel.
- perform single marker and haplotypic association analysis.
- detect genetic regions that are shared homozygous-by-descent (HBD) or identical-by-descent (IBD).

You can mix-and-match four different kinds of data: phase-unknown genotype data, phase-known haplotype data, parent-offspring trio data, and parent-offspring pair data. BEAGLE will infer haplotypes and impute missing genotypes and ungenotyped markers for each kind of data. For unrelated data, BEAGLE can also accept genotype likelihoods, instead of called genotypes, when performing genotype phasing and imputation (see Section 2.2).

BEAGLE has options for estimating identity-by-descent (IBD) probabilities and homozygosity-by-descent (HBD) probabilities from called genotypes. BEAGLE version 3.3 includes a new fast algorithm for IBD detection called fastIBD.

Beagle can be used in tandem with the PRESTO software package. PRESTO can compute empirical distributions of order statistics, analyze stratified data, and determine significance levels for one and two-stage genetic association studies. PRESTO is freely available from http://faculty.washington.edu/browning/presto/presto.html.

BEAGLE is written in Java and runs on most computing platforms (e.g. Windows, Unix, Linux, Solaris, and Mac). A Java interpreter is probably already installed on your computer (type `java -version` at the command line prompt to check). However, if it is not installed or if it is not version 1.6 or later, the current Java interpreter can be downloaded free of charge from the **java.sun.com** web site. The Java interpreter is called the Java Standard Edition (SE) Runtime Environment (JRE).

1.1 Citing BEAGLE

If you use BEAGLE and publish your analysis, please report the version of the program used and please cite the appropriate publication that describes the BEAGLE methods that were used in your analysis:

1. BEAGLE's fastIBD method for fast identity-by-descent detection is not yet published. A reference describing the fast identity-by-descent detection method will be posted on the Beagle web site as soon as it is available.
2. BEAGLE's methods for estimating identity-by-descent and homozygosity-by-descent probabilities are described in

S R Browning and B L Browning (2010) High resolution detection of identity by descent in unrelated individuals. Am J Hum Genet. In Press. doi:10.1016/j.ajhg.2010.02.021.

3. BEAGLE's methods for calling genotypes from genotype likelihood data are described in

B L Browning and Z Yu (2009) Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. Am J Hum Genet 85(6):847-61. doi:10.1016/j.ajhg.2009.11.004.

4. BEAGLE's methods for imputing ungenotyped markers and phasing parent-offspring data are described in

B L Browning and S R Browning (2009) A unified approach to genotype imputation and haplotype phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet 84:210-223. doi:10.1016/j.ajhg.2009.01.005.

5. BEAGLE's methods for inferring haplotype phase and sporadic missing data in unrelated individuals are described in

S R Browning and B L Browning (2007) Rapid and accurate haplotype phasing and missing data inference for whole genome association studies using localized haplotype clustering. Am J Hum Genet 81:1084-1097. doi:10.1086/521987

6. BEAGLE's methods for association testing are described in

Browning and Browning (2007) Efficient multilocus association testing for whole genome association studies using localized haplotype clustering. Genet Epidemiol 31:365-375. doi:10.1002/gepi.20216.

7. BEAGLE's haplotype frequency model was first described in

S R Browning (2006) Multilocus association mapping using variable-length Markov chains. Am J Hum Genet 78:903-13.

1.2 Files in the BEAGLE software distribution

BEAGLE is freely available and can be downloaded from the BEAGLE web site:

http://faculty.washington.edu/browning/beagle/beagle.html

The following files are available for download:

1. beagle.jar - the BEAGLE executable file.
2. beagle_3.3_[date].pdf - the BEAGLE documentation.
3. beagle_example.zip - a folder containing input and output files for two BEAGLE analyses described in Section 6.

Utility programs that can be used to prepare input files and process output files are available at the BEAGLE Utilities web site (see Section 7.1):

http://faculty.washington.edu/browning/beagle_utilities/utilities.html

The BEAGLE web site also has three specialized utility programs available for download:

1. divide.sample - a unix shell script for dividing a sample and performing imputation in each subsample separately.
2. pseudomarker.jar - a utility program for creating a phased Beagle file of pseudomarkers from a Beagle output model (.dag) file (see Section 7.2 for more details).
3. cluster2haps.jar - a utility program for identifying the allele sequences that define a haplotype cluster that is associated with a trait (see Section 7.3 for more details).

2 BEAGLE file formats

Input text files can be compressed with the gzip algorithm. BEAGLE assumes that any file that has a name ending in ".gz" is compressed with gzip.

2.1 Genotypes file format

A Beagle genotypes file has a simple format: rows are variables and columns are individuals. Here is an example of a Beagle genotypes file with three individuals and three genotyped markers:

Example 1 - Sample Beagle file

I	id	1001	1001	1002	1002	1003	1003
A	diabetes	1	1	2	2	2	2
M	rs2289311	A	G	G	G	A	G
M	rs1248628	T	T	T	C	T	T
M	rs10762764	G	T	T	T	G	T

In a Beagle genotypes file, the first column describes the data on each line. The fields in the first column are typically single characters, but this is not required. The second column contains the name of the variable whose data is given on each line. Variable names should be unique. In Example 1 there are two columns for each individual: columns 3-4 give data for the first individual, columns 5-6 give data for the second individual, and so on.

In the Beagle genotypes file in Example 1, the first line (I ....) is called a sample identifier line and gives an identifier for each column of data. A sample identifier line is not required, but is highly recommended and will be required in later BEAGLE versions. The second line (A ...) is called an affection status line and gives an affection status (1 = unaffected, 2 = affected) for each individual. An affection status line is not required unless you are performing association testing. The last three lines (M ...) are marker lines that give marker alleles for the three markers (rs2289311, rs1248628, and rs10762764). Note that an identifier and an affection status are given for each allele (column). For diploid data, the identifier and affection status will typically be the same for both alleles.

In the Beagle genotypes file in Example 1, the three individuals have identifiers 1001 (columns 3-4), 1002 (columns 5-6), and 1003 (columns 7-8). The first individual (columns 3-4) is unaffected, and the second and third individuals (columns 5-6 and 7-8) are affected.

A BEAGLE genotypes file can have either unphased data or phased data for unrelated individuals, parent-offspring trios, or parent-offspring pairs. Here is how each type of data is represented in a Beagle file:

- **Unphased unrelated data**. Each pair of columns (beginning with columns 3-4) gives the genotype for each unrelated individual. If the Beagle file in Example 1 contains unphased data for unrelated individuals, the first individual has genotypes A/G, T/T, G/T, the second individual has genotypes G/G, C/T, T/T, and the third individual has genotypes A/G, T/T, G/T for markers rs2289311, rs1248628, and rs10762764 respectively. Input files with unphased unrelated data are specified with the unphased command line argument (see Section 3.2.1).
- **Phased unrelated data**. Each column (beginning with column 3) gives a phased haplotype, and for diploid data, each pair of columns gives the pair of phased haplotypes for each diploid individual. If the Beagle file in Example 1 contains phased data for unrelated individuals, then the first individual has haplotypes ATG and GTT, the second individual has haplotypes equal to GTT and GCT, and the third individual has haplotypes ATG and GTT for markers rs2289311, rs1248628, and rs10762764 respectively. When Beagle is used to phase genotype data for unrelated individuals, the input Beagle file contains *unphased* unrelated data, and the output Beagle file contains *phased* unrelated data. Input files with phased unrelated data are specified with the phased command line argument (see Section 3.2.1).
- **Unphased trio data**. Each set of six consecutive columns (beginning with columns 3-8) gives the genotype data for one parent-offspring trio. In each set of six columns, the first two columns give the genotypes for the first parent, the middle two columns give the genotypes for the second parent, and the last two columns give the genotypes for the offspring. If the Beagle genotypes file in Example 1 contains unphased trio data, the first parent has identifier 1001 (columns 3-4), the second parent has identifier 1002 (columns 5-6), and the child has identifier 1003 (columns 7-8). Input files with unphased trio data are specified with the trios command line argument (see Section 3.2.1). Unphased trio data is not permitted to have any Mendelian inconsistencies. Genotypes causing Mendelian inconsistencies for a trio must be replaced with missing genotypes.
- **Phased trio data**. Each set of four consecutive columns (beginning with columns 3-6) gives the transmitted and untransmitted haplotypes for one parent-offspring trio. In each

set of four columns, the first column is the first parent's transmitted haplotype, the second column is the first parent's untransmitted haplotype, the third column is the second parent's transmitted haplotype, and the fourth column is the second parent's untransmitted haplotype. If the Beagle genotypes file in Example 1 contains unphased trio data, then one can tell by inspection that the first parent transmits the ATG haplotype and the second parent transmits the GTT haplotype. Thus in the preceding example, the corresponding phased trio file is obtained by deleting the offspring data (sample id 1003) in columns 7-8. Phased offspring genotypes can be reconstructed from the first column of each parent in the phased trio data. When Beagle is used to phase genotype data for parent-offpspring trios, the input Beagle file contains *unphased* trio data, and the output Beagle file contains *phased* trio data, unless the redundant=true command line option is specified (see Section 3.2.2). If the redundant=true is specified the phased child haplotypes are included in the phased output file.

- **Unphased pair data**. Each set of four consecutive columns (beginning with columns 3-6) gives the genotype data for one parent-offspring pair. In each set of four columns, the first two columns give the genotypes for the genotyped parent, and the last two columns give the genotypes for the offspring. If the Beagle file in Example 2 below contains unphased pair data, the genotyped parent has identifier 1001 (columns 3-4), and the offspring has identifier 1002 (columns 5-6). Input files with unphased pair data are specified with the pairs command line argument (see Section 3.2.1). Unphased pair data is not permitted to have any Mendelian inconsistencies. Genotypes causing Mendelian inconsistencies for a pair must be replaced with missing genotypes.
- **Phased pair data**. Each set of three consecutive columns (beginning with columns 3-5) gives the transmitted and untransmitted haplotypes for one parent-offspring pair. In each set of three columns, the first column is the genotyped parent's transmitted haplotype, the second column is the genotyped parent's untransmitted haplotype, and the third column is the ungenotyped parent's transmitted haplotype. If the Beagle file in Example 2 below contains unphased pair data, then one can tell by inspection that the genotyped parent transmits the GTT haplotype. Thus the corresponding phased pair file is obtained by deleting the fifth column in Example 2 (1002-G-T-T). Phased offspring genotypes can be reconstructed from the columns containing the transmitted haplotypes from the genotyped and ungenotyped parents. When Beagle is used to phase genotype data for parent-offpspring pairs, the input Beagle file contains *unphased* pair data, and the output Beagle file contains *phased* pair data, unless the redundant=true command line option is specified (see Section 3.2.2). If the redundant=true is specified, both the transmitted and untransmitted offspring haplotypes are included in the phased output file (in that order).

Example 2 - Sample Beagle file with parent-offspring pair data

I	id	1001	1001	1002	1002
A	diabetes	1	1	2	2
M	rs2289311	G	A	G	G
M	rs1248628	T	T	T	C
M	rs10762764	T	G	T	T

BEAGLE imposes very few constraints on your data files:

- Input data for haplotype phase inference can have missing alleles or genotypes. After phasing data, all missing data are imputed.
- Alleles and marker identifiers can be any sequence of characters that does not contain white space and that does not equal the user-specified missing allele code. Marker alleles are not restricted to A/C/G/T or to 1/2/3/4.
- Data fields (e.g. marker alleles) on each line can be separated by a space, a tab, or any combination of spaces and tabs.
- Markers can have up to 128 different alleles. In particular, triallelic SNPs and microsatellite markers can be used. However, at present, genotype probabilities are only computed when all markers are diallelic.

Beagle files contain two sections: **header lines** and **marker lines**. **Header lines** are all lines that precede the first marker line (M ...). Header lines contain non-marker data. In the preceding two example Beagle files, the header lines are the sample identifier line (I ...) and the affection status line (A ...). The header lines are optional, but it is recommended that you include a sample identifier line in your input BEAGLE genotypes files. **Marker lines** are the lines beginning with the first marker line (M ...) and ending with the last line of the file (inclusive). BEAGLE ignores any line in the markers line section whose first field is not 'M'.

Each line in a Beagle file must have the same number of fields, unless the first field is the hash character '#'. A line whose first field is the hash character, '#', is called a comment line. Comment lines (# ...) are ignored. All comment lines in the header line section are copied to the output Beagle files.

If multiple sample identifier lines are included in the header lines, only the first one will be used by BEAGLE to identify columns. You can also include a pedigree identifier (P ...) line, a father identifier (FID ...) line, a mother identifier (MID ...) line, and a population stratum line (S ...) in the header line section.

The header line section also contains all phenotype variables, including binary traits (A ...), quantitative traits (T ...), and categorical covariates (C ...). Affection status data are used for association testing, but are not used for phasing or for building the haplotype frequency model. Quantitative trait and covariate data are not currently used by BEAGLE, but can be used by other programs, such as R.[1]

BEAGLE can use transmitted and untransmitted haplotypes from parent-offspring data (affected individuals and their parents) to test for association with a binary affection status. With transmitted and untransmitted haplotypes, you will need to add an affection status line with the affection status of transmitted haplotypes coded as 2 and the affection status of untransmitted haplotypes coded as 1, and you will need to use the diplotypes=false option (see Section 4.2.3). When diplotypes=false, the analysis assumes haplotypes are independent, and haplotypes in adjacent columns (e.g. columns 3-4, 5-6, etc.) are not required to have the same affection status.

When inferring haplotype phase and missing data, you will typically create a separate Beagle file for each chromosome. The markers in the Beagle file must be in chromosomal order.

2.2 Genotype likelihoods file format

BEAGLE can also accept input files that contain genotype likelihoods for unphased, unrelated data when performing genotype phasing and imputation and fastIBD analysis. A genotype likelihood for genotype G in a sample is the value of the probability density function for the observed genotype data for when the true genotype is G ($G = AA$, AB, or BB). Genotype likelihoods convey the relative evidence for each possible genotype call. The HBD and IBD detection methods cannot accept genotype likelihood data.

Genotype likelihoods files have a simple format: rows are markers and columns are individuals. The following example genotype likelihoods file has two individuals and three markers:

Example 3 - Sample genotype likelihoods file

marker	alleleA	alleleB	1001	1001	1001	1002	1002	1002
rs2289311	A	G	0.0012	0.9858	0.0130	0.9601	0.0398	0.0001
rs1248628	C	T	0.1410	0.8555	0.0035	0.0469	0.9531	0.0000
rs10762764	G	T	0.0000	0.0005	0.9995	1.0000	0.0000	0.0000

In a genotype likelihoods file, the first line is a header line that describes the data in each column. The first column (**marker**) is the marker identifier. Markers must be listed in chromosomal order. The second and third columns (**alleleA** and **alleleB**) give the two marker alleles. The remaining columns give the estimated genotype likelihoods for the AA, AB, and BB genotypes (in that order) for each individual: columns 4-6 give the genotype likelihoods for the first individual, columns 7-9 give the genotype likelihoods for the second individual, and so on. The three genotype likelihoods must be non-negative, but do not need to sum to 1 since it is the ratio of likelihoods rather than their absolute value that is important. The sample identifier in the header line must be the same for all three columns for each individual.

In Example 3, the two sample identifiers are 1001 and 1002, the genotype of the heterozygous genotype (AG) for marker rs2289311 and sample 1001 is 0.9858, and the likelihood of the homozygous T-allele genotype (TT) for marker rs1248628 and sample 1002 is 0.0000.

2.3 Genotype probabilities file format

A genotype probability for genotype G in a sample is the probability that the true genotype is G conditional on the observed genotype data for when the true genotype is G ($G = AA$, AB, or BB). BEAGLE generates output genotype probabilities files for unphased, unrelated data. Genotype probabilities files have exactly the same format as the genotype likelihoods files. The only difference is that the genotype likelihoods are replaced by genotype probabilities. Thus for a genotype probabilities file, the three columns for each individual give the estimated probability that true genotype is AA, AB, and BB (in that order).

2.4 Markers file format

If the data for a chromosome is divided among two or more Beagle files, you must use a markers file. A markers file is required to reconstruct the marker order because BEAGLE does not require an input file to contain all the markers. For example, when imputing

ungenotyped markers in a sample using a reference panel, the ungenotyped markers may be omitted from the Beagle file containing the genotype data for the to-be-imputed sample.

Each line of the marker file will contain four or more white-space delimited fields. The first field is the marker identifier. The second field is the marker position. The remaining fields are the marker alleles. The markers must be given in chromosomal order. If you are performing HBD/IBD detection, the position of each marker must be given using the centiMorgan scale. If you are not performing HBD/IBD detection, the marker positions on a chromosome can be base positions, genetic positions, or sequential integers. The fastIBD algorithm does not require cM marker positions

In the Beagle file given in Example 1 above, the corresponding markers file when marker positions are given in NCBI Build 35 coordinates is

```
rs2289311    79235661    A    G
rs1248628    79236165    C    T
rs10762764   79236371    G    T
```

If you include data for the triallelic SNP, rs2032582, the corresponding line of the marker file would be:

```
rs2032582    86205269    G    T    A
```

3 Inferring haplotype phase and missing data with BEAGLE

Beagle can perform haplotype phase inference and missing data imputation using data from unrelated individuals, parent-offspring trios, parent-offspring pairs, and phase-known haplotypes.

3.1 Quick start guide

To run BEAGLE, enter the following command at the computer prompt:

```
java -Xmx<Mb>m -jar beagle.jar <arguments>
```

where <Mb> is the number of Megabytes of memory available (e.g. -Xmx1000m) and <arguments> is a space separated list of arguments. Each argument has the format **parameter**=**value**. There is no white-space between the **parameter** and **=** or between **=** and the **value**. Large data sets with thousands of samples may require several gigabytes of memory (see Section 5.1).

New BEAGLE users should use the arguments for specifying input files, the output prefix, and the missing allele code, but do not need to specify any other arguments. Other BEAGLE arguments are optional and have sensible default values. The format for Beagle input files is described in Section 2.

The commands for inferring haplotype phase and imputing missing data are very simple. Here are three example command lines:

1. java -Xmx1000m -jar beagle.jar unphased=fileA.bgl missing=? out=example
2. java -Xmx1000m -jar beagle.jar unphased=fileA.bgl phased=fileB.bgl markers=markers.txt missing=? out=example

3. java -Xmx1000m -jar beagle.jar unphased=fileA.bgl unphased=fileB.bgl trios=fileC.bgl pairs=fileD.bgl markers=markers.txt missing=? out=example

Line 1 shows how to infer haplotype phase and impute sporadic missing data. In line 1, fileA.bgl contains unphased unrelated data with missing alleles coded as "?". If fileA.bgl contains unphased parent-offspring trios, replace "unphased=fileA.bgl" with "trios=fileA.bgl". If fileA.bgl contains unphased parent-offspring pairs, replace "unphased=fileA.bgl" with "pairs=fileA.bgl".

Line 2 is a typical command for imputing ungenotyped markers in a file called fileA.bgl that have been genotyped in a reference panel called fileB.bgl. The markers file lists the markers in fileB.bgl in chromosomal order. Missing alleles in the input file are coded as "?". In line 2, fileA.bgl contains unphased, unrelated data, and the reference panel (fileB.bgl) is phased, unrelated data. If the reference panel is unphased, unrelated data, replace "phased=fileB.bgl" with "unphased=fileB.bgl". If the reference panel is parent-offspring trio data, replace "phased=fileB.bgl" with "trios=fileB.bgl". BEAGLE will assume any input file with < 7% missing alleles represents a reference panel.

Line 3 illustrates how to infer haplotype phase and impute sporadic missing data simultaneously in multiple cohorts. Using multiple cohorts permits large sample sizes, which increases the accuracy of the inferred haplotypes and missing data. In line three there are four cohorts. Two cohorts contain unphased, unrelated data (fileA.bgl, fileB.bgl), one cohort contains parent-offspring trio data (fileC.bgl), and one cohort contains parent-offspring pair data (fileD.bgl). The markers file (markers.txt) contains the markers included in the analysis, and missing alleles in the input files are coded as "?". If any markers in the markers file are not present in an input file, genotypes for the missing markers will be imputed.

A markers file is required when there is more than one input data file. If an input file contains a marker that is not present in the markers file, the marker will be ignored. If an input Beagle file is missing any markers present in the markers file, data for the missing markers will be imputed and included in the output files. The format of the markers file is described in Section 2.4. When imputing ungenotyped markers using a reference panel, the markers file should contain only the markers genotyped in the reference panel. For example, if you are imputing ungenotyped markers in an unphased sample using phased HapMap data as a reference panel, it is recommended that the markers file omit any markers that genotyped in the unphased sample, but not in the HapMap data. If you have two disjoint or nested reference panels genotyped on two different marker sets, you should run imputation twice: once with each reference panel. There is a utility program called "updategprobs.jar" available at the BEAGLE Utilities web site (see Section 7.1) that can combine the two sets of imputed data from the two BEAGLE runs.

3.2 Command line arguments

This section describes the BEAGLE command line arguments for inferring haplotype phase and missing data.

During analysis BEAGLE will create several temporary files in your system's default temporary-file directory. If your system's default temporary-file directory has insufficient space, you can specify the temporary-file directory by replacing the initial "java" in the command line with the "java -Djava.io.tmpdir=<directory>" argument where "<directory>" is the name of an alternate directory for storing temporary files.

3.2.1 Arguments for specifying files

- unphased=<unphased unrelated file> where <unphased unrelated file> is the name of a Beagle file containing **unphased unrelated** genotype data (see Section 2.1). You may use multiple unphased arguments if data from different cohorts are in different files.
- phased=<phased unrelated file> where <phased unrelated file> is the name of a Beagle file containing **phased unrelated** data (see Section 2.1). You may use multiple phased arguments if data from different cohorts are in different files.
- trios=<unphased trio file> where <unphased trio file> is the name of a Beagle file containing unphased parent-offspring trio data (see Section 2.1). You may use multiple trios arguments if data from different cohorts are in different files.
- pairs=<unphased pairs file> where <unphased pairs file> is the name of a Beagle file containing unphased parent-offspring pair data (see Section 2.1). You may use multiple pairs arguments if data from different cohorts are in different files.
- like=<unphased likelihood data file> where <unphased likelihood data file> is the name of a genotype likelihoods file for unphased, unrelated data (see Section 2.2). You may use multiple like arguments if data from different cohorts are in different files.
- markers=<markers file> where <markers file> is the name of the markers file containing marker identifiers, positions, and alleles described in Section 2.4. The markers argument is optional if you specify only one Beagle file, and is required if you specify more than one Beagle file.
- missing=<missing code> where <missing code> is the character or sequence of characters used to represent a missing allele (e.g. missing=-1 or missing=?). The missing argument is required.
- out=<output file prefix> where <output file prefix> is the prefix for the output filename. The output prefix must be an absolute or relative filename, but it cannot be a directory. Output files corresponding to an input file called Z will be [output file prefix].Z.[ext] where [ext] describes the data in the output file. The different output files are described in Section 3.3. The out argument is required.
- omitprefix=<true/false> where <true/false> is true if the output prefix should be omitted from all output filename except for the log file, and false if all files should begin with the output prefix. All output files will be written to the same directory as the log file. The omitprefix argument is optional. The default value is omitprefix=false.
- maxlr=<max likelihood ratio> where <max likelihood ratio> is the maximum permitted likelihood ratio. For input genotype likelihoods files (see Section 2.2), if the ratio of any two non-zero likelihoods (corresponding to two different genotype calls for sample) is greater than <max likelihood ratio>, the smaller likelihood will be set equal to 0.0. The maxlr argument is optional. The default value is maxlr=5000.

3.2.2 Other phasing arguments

- niterations=<number of iterations> where <number of iterations> is a positive even integer giving the number of iterations of the phasing algorithm. If an odd integer is specified, the next even integer is used. The niterations argument is optional. The default value is niterations=10. The default value typically gives good accuracy.

- nsamples=<number of samples> where <number of samples> is positive integer giving the number of haplotype pairs to sample for each individual during each iteration of the phasing algorithm. The nsamples argument is optional. The default value is nsamples=4. If you are phasing an extremely large sample (say > 4000 individuals), you may want to use a smaller nsamples parameter (e.g. 1 or 2) to reduce computation time. If you are phasing a small sample (say < 200 individuals), you may want to use a larger nsamples parameter (say 10 or 20) to increase accuracy.
- gprobs=<true/false> where <true/false> is false if genotype probability files should not be produced for files specified with the unphased parameter (see Sections 2.1 and 2.3). The gprobs argument is optional. The default value is gprobs=true. If the input Beagle files have sporadic missing data and you are not imputing data for ungenotyped markers, you may want to set gprobs=false.
- seed=<random seed> where <random seed> is an integer seed for the random number generator. The seed argument is optional. The default value is seed=-99999.
- lowmem=<true/false> where <true/false> is true if a memory-efficient, but slower, implementation of the sampling algorithm should be used. If lowmem=true the running time will increase by a factor ≤ 2, and the memory usage will be essentially independent of the number of markers. The lowmem argument is optional. The default value is lowmem=false.
- excludecolumns=<excluded columns file> where <excluded columns file> is the name of file containing column identifiers (one identifier per line) that will be excluded from the analysis and output files. Beagle will check the column identifiers in the first sample identifier line (I ...) of each Beagle input file and exclude any columns whose identifier is in the excluded samples file.
- excludemarkers=<excluded markers file> where <excluded markers file> is the name of file containing marker identifiers (one identifier per line) that will be excluded from the analysis and output files.
- verbose=<true/false> where <true/false> is true if running time and graphical model statistics are printed to the log file for each iteration of the algorithm. The verbose argument is optional. The default value is verbose=false.
- nimputations=<number of imputations> where <number of imputations> is a nonnegative integer giving the number of output files with imputed phased data to create for each input Beagle file. Each imputed data file is in Beagle genotypes file format and contains phased haplotypes that are randomly sampled conditional on the observed data and the estimated haplotype frequency model (see Section 3.3.9). The nimputations argument is optional. The default value is nimputations=0. If multiple runs of BEAGLE are used to sample haplotypes, a different random seed should be specified for each run (see the seed argument described earlier in this section).
- redundant=<true/false> where <true/false> is true if transmitted haplotypes will be printed twice in the output Beagle files for parent-offspring pairs and trios (once for the parent and once for the offspring), and false if transmitted haplotypes should be printed

only once in output files. The redundant argument is optional. The default value is redundant=false.

3.2.3 fastIBD arguments

- fastibd=<true/false> where <true/false> is true if a fastIBD analysis is to be performed and false otherwise. The fastibd argument is optional. The default value is fastibd=false. For best results, we recommend performing a fastIBD analysis 10 times using different seed parameters (see Section 3.2.2), and combining the fastIBD output from the 10 runs.
- fastibdthreshold=<score threshold> where <score threshold> is the fastIBD score threshold that controls fastIBD output. Pairs of individuals whose fastIBD score for the shared haplotype is less than the fastibdthreshold parameter will have the fastIBD score and starting and ending marker indices of the shared haplotype printed to the fastIBD output file (see Section 3.3.6). BEAGLE can perform fastIBD estimation for unphased, unrelated data, unphased trio data, and genotype likelihood data (specified with the "unphased=", "trio=" or "like=" parameters). The fastIBD analysis is performed for all pairs of individuals within each file of unrelated data and for all parents in each file of trio data. The fastibdthreshold parameter must satisfy $0.0 \leq$ fastibdthreshold ≤ 1.0. The default value is fastibdthreshold=1.0e-6 (i.e. 10^{-6}) which is suitable for an outbred populations.
- ibdscale=<IBD, HBD, and fastIBD tuning parameter> where <IBD, HBD, and fastIBD tuning parameter> is a tuning parameter that controls the complexity of the haplotype frequency model when performing IBD, HBD, and fastIBD analysis.. Higher values of the tuning parameter correspond to reduced model complexity. The ibdscale parameter must be a positive real number. The default value is ibdscale=2.0. The optimal value may depend on the population, marker density, and sample size. The default value was selected based on analysis of 1500 individuals of European ancestry from the 1958 UK Birth Cohort, genotyped on the Illumina 550K array. We expect that the optimal ibdscale parameter will be between 1.0 and 4.0 for most data sets.

3.2.4 IBD arguments

- ibdpairs=<IBD pairs file> where <IBD pairs file> is the name of a file containing pairs of sample identifiers. If you use the ibdpairs argument, the Beagle genotypes file must contain a sample identifier line (see Section 2.1). Each line of the IBD pairs file must have two white-space delimited fields containing two non-identical sample identifiers. If an IBD pairs file is specified, BEAGLE will estimate IBD for the pair of individuals on each line, and write the output to a file with an ".ibd" extension. Currently, BEAGLE will only estimate IBD for pairs of unphased, unrelated individuals whose genotype data are in the same file (specified with the "unphased=" parameter). If the ibdpairs argument is used, a markers file must be used and the marker positions must be given in the cM scale. For best results, we recommend performing IBD estimation 10 times using different seed parameters (see Section 3.2.2), and using the maximum IBD probability at each locus (maximized over the 10 runs). The ibdpairs argument is optional.
- nonibd2ibd=<non-IBD to IBD transition rate> where <non-IBD to IBD transition rate> is the transition rate from non-IBD to IBD per cM for each sample. The markers file must

give positions for each marker using the cM scale (see Section 2.4). The nonibd2ibd argument must be non-negative and is optional. The default value is nonibd2ibd=0.0001.

- ibd2nonibd=<IBD to non-IBD transition rate> where <IBD to non-IBD transition rate> is the transition rate from IBD to non-IBD per cM for each sample. The markers file must give positions for each marker using the cM scale (see Section 2.4). The ibd2nonibd argument must be non-negative and is optional. The default value is ibd2nonibd=1.0
- ibderror=<genotype error rate> where <genotype error rate> is the estimated genotype error rate when estimating IBD. The ibderror argument is optional. The default value ibderror=0.005.
- ibdscale=<IBD, HBD, and fastIBD tuning parameter> where <IBD, HBD, and fastIBD tuning parameter> is a tuning parameter that controls the complexity of the haplotype frequency model when performing IBD, HBD, and fastIBD analysis.. Higher values of the tuning parameter correspond to reduced model complexity. The ibdscale parameter must be a positive real number. The default value is ibdscale=2.0. The optimal value may depend on the population, marker density, and sample size. The default value was selected based on analysis of 1500 individuals of European ancestry from the 1958 UK Birth Cohort, genotyped on the Illumina 550K array. We expect that the optimal ibdscale parameter will be between 1.0 and 4.0 for most data sets.

3.2.5 HBD arguments

- estimatehbd=<true/false> where <true/false> is true if HBD estimation will be performed, and false if HBD estimation will not be performed. If the estimatehbd=true argument is used, a markers file must be used and the marker positions must be given in the cM scale. The estimatehbd argument is optional. The default value is estimatehbd=false. For best results, we recommend performing an HBD analysis 10 times using different seed parameters (see Section 3.2.2), and using the maximum HBD probability (maximized over the 10 runs) at each locus.
- nonhbd2hbd=<non-HBD to HBD transition rate> where <non-HBD to HBD transition rate> is the transition rate from non-HBD to HBD per cM for each sample. The markers file must give positions for each marker using the cM scale (see Section 2.4). The nonhbd2hbd argument must be non-negative and is optional. The default value is nonhbd2hbd=0.0001.
- hbd2nonhbd=<HBD to non-HBD transition rate> where <HBD to non-HBD transition rate> is the transition rate from HBD to non-HBD per cM for each sample. The markers file must give positions for each marker using the cM scale (see Section 2.4). The hbd2nonhbd argument must be non-negative and is optional. The default value is hbd2nonhbd=1.0.
- hbderror=<genotype error rate> where <genotype error rate> is the estimated genotype error rate when estimating HBD. The hbderror argument is optional. The default value hbderror=0.005.
- ibdscale=<IBD, HBD, and fastIBD tuning parameter> where <IBD, HBD, and fastIBD tuning parameter> is a tuning parameter that controls the complexity of the haplotype frequency

model when performing IBD, HBD, and fastIBD analysis.. Higher values of the tuning parameter correspond to reduced model complexity. The ibdscale parameter must be a positive real number. The default value is ibdscale=2.0. The optimal value may depend on the population, marker density, and sample size. The default value was selected based on analysis of 1500 individuals of European ancestry from the 1958 UK Birth Cohort, genotyped on the Illumina 550K array. We expect that the optimal ibdscale parameter will be between 1.0 and 4.0 for most data sets.

3.2.6 Advanced options not intended for general use

- maxwindow=<maximum window size> where <maximum window size> is a positive integer giving the maximum number of consecutive markers that will be considered when building the haplotype frequency model. The maxwindow argument is optional. The default value is maxwindow=500. The default value should be sufficiently large for most data sets. For example, if your marker density is one marker per kilobase, BEAGLE will consider markers in a 500 kilobase window when using the default maxwindow parameter.

3.3 Output files

After inferring haplotype phase and missing data, a log file (.log) is created that summarizes the analysis, and a phased file (.phased) is created for each input data file. The phased file gives the most likely pair of phased haplotypes for each sample. Depending upon the command line arguments, additional output files (.gprobs, .r2, and .sample) may also be created.

3.3.1 log file [.log]

The log file gives a summary of the analysis that includes the BEAGLE version, a list of the command line arguments for the analysis, and the running time for the analysis. If verbose=true is included as a command line argument, the log file will also include the running time for each iteration of the phasing algorithm and descriptive statistics for the graphical haplotype frequency model at each iteration of the phasing algorithm. The filename for the log file is set with the out command line argument (see Section 3.2.1).

3.3.2 Phased file [.phased.gz]

Each input Beagle file will have a corresponding phased output Beagle genotypes file that gives imputed missing data and inferred haplotypes. The phased file gives the most likely haplotype pair for each individual conditional upon the genotypes for the individual and the haplotype frequency model. If the input file contains data for unrelated individuals, the corresponding phased output file will contain phased unrelated data. If the input file contains unphased parent-offspring trio data, the corresponding phased output file will contain phased trio data. If the input file contains unphased parent-offspring pair data, the corresponding phased output file will contain phased pair data (see Section 2.1). A filename prefix for the phased output files can be specified with the out command line argument (see Section 3.2.1).

3.3.3 Genotype probabilities file [.gprobs.gz]

Each input Beagle file that contains only diallelic markers (e.g. SNPs) and unphased unrelated data will have a corresponding genotype probabilities output file unless the gprobs=false argument is specified (see Section 3.2.2). The format of the genotype

probabilities file is described in Section 2.3. If a genotype is non-missing in an input Beagle genotypes file for unphased, unrelated data, the corresponding genotype in the output genotype probabilities file will have probability 1.0. A filename prefix for the genotype probabilities output files can be specified with the out command line argument (see Section 3.2.1).

3.3.4 Genotype dosage file [.dose.gz]

If a sample has genotype probabilities ($P(AA)$, $P(AB)$, $P(BB)$) for a marker, then the estimated B-allele dosage is $0 \times P(AA) + 1 \times P(AB) + 2 \times P(BB)$. Each input Beagle file that contains only diallelic markers (e.g. SNPs) and unphased unrelated data will have a corresponding genotype dosage output file unless the gprobs=false argument is specified (see Section 3.2.2). The header line of the genotype dosage file is similar to the header line of the genotypes probabilities output file (see Section 3.3.3), except that the sample identifiers are listed only once (one column per sample). The remaining lines give data for each marker (one marker per line). The first three columns of the genotype dosage file are identical to the first 3 columns in the genotype probabilities output file (see Section 3.3.3). The remaining columns give the estimated B-allele dosages for each marker and each sample (one sample per column). A filename prefix for the genotype dosage output files can be specified with the out command line argument (see Section 3.2.1).

3.3.5 Allelic R^2 file [.r2]

The allelic R^2 file is produced whenever a genotype probabilities file is produced (see Section 3.3.3). The allelic R^2 file contains two columns, the first column gives the marker identifier, and the second column gives the estimated squared correlation ($0 \leq R^2 \leq 1$) between the allele dosage with highest posterior probability in the genotype probabilities file and the true allele dosage for the marker. Larger values of allelic R^2 indicate more accurate genotype imputation. The estimated allelic R^2 for a marker is "NaN" (Not-a-Number) when the estimated allelic R^2 cannot be computed. The allelic R^2 cannot be computed for a marker when the most likely genotype is the same for each individual (i.e. only one genotype is observed), or when there is no data for the marker in any Beagle input file. The estimated squared correlation can be used when deciding whether to retain or discard a marker in downstream analyses. The allelic R^2 file can also be used to detect inter-cohort differences in imputation accuracy. A directory or a filename prefix for the allelic R^2 output files can be set with the out command line argument (see Section 3.2.1).

The allelic R^2 measure reported by BEAGLE is closely related to the ratio-of-variances R^2 measure reported by the MACH software package.[2] BEAGLE's allelic R^2 measure estimates the squared correlation between the most likely allele dosage and the true allele dosage, and MACH's ratio-of-variances R^2 measure estimates the squared correlation between the estimated allele dosage ($0 \times P(AA) + 1 \times P(AB) + 2 \times P(BB)$) and the true allele dosage. When we first described the BEAGLE allelic R^2 measure,[3] we had not realized that the MACH ratio-of-variances measure (which predates BEAGLE's measure), was a direct estimate of a squared correlation.

3.3.6 fastIBD file [.fibd.gz]

A fastIBD file is produced when the fastibd=true parameter is specified (see Section 3.2.3). A separate estimated fastibd output file is produced for each input file of unphased, unrelated, genotype likelihood, or trio data. Lines of the fastIBD output file report

haplotypes shared by pairs of samples within the corresponding input file that have fastIBD score less than the threshold specified by the fastibdthreshold parameter. The fastIBD output file has five columns. The first two columns list the two sample identifiers for the shared haplotype described on each line. The next two columns list the starting (inclusive) and ending (exclusive) marker indices for the shared haplotype. The first marker has index 0. The last column gives the fastIBD score for the shared haplotype. A fastIBD score $< 10^{-10}$ provides strong evidence that the shared haplotype is IBD if the length of the shared haplotype length is ≥ 1 cM.

3.3.7 IBD file [.ibd]

An IBD output file is produced when an IBD pairs input file is specified with the ibdpairs parameter (see Section 3.2.4). An IBD ouput file is produced for each file of unphased, unrelated data. Each IBD output file contains estimated IBD probabilities for pairs of samples that are present in both the corresponding input file of unphased, unrelated data and in the IBD pairs input file. Lines correspond to markers, and the first column lists the marker identifiers. The first two lines specify the pair of sample identifiers corresponding to each column of data. There are four columns of output data for each pair of individuals:

1) Estimated probability that the two samples are IBD at the marker (the marker identifier is specified in the first field of each line).
2) Estimated allele that is shared IBD if the two samples are IBD at the marker. The missing allele symbol is reported if the shared allele could not be estimated.
3) The estimated probability that the pair of individuals are IBD at the marker AND that the shared allele in the preceding column is correctly identified.
4) The genotype for the pair of individuals, separated by the "/" symbol. Homozygote genotypes are specified by the homozygous allele, heterozygote genotypes are specified by "H", and missing genotypes are specified by the missing data symbol. For example, "H/G" indicates the first individual is heterozygous and the second individual is homozygous for the G-allele. A trailing "*" indicates the two genotypes do not share an allele and are incompatible with IBD status (e.g. "A/G*").

3.3.8 HBD file [.hbd.gz]

An HBD file is produced when the estimatehbd=true parameter is specified (see Section 3.2.5). A separate HBD file is produced for each input file of unphased, unrelated data. Lines correspond to samples and columns correspond to markers. The first line lists the marker identifiers in chromosomal order, and the first column lists the sample identifiers. The estimated probability that a sample is HBD at a marker is given in the field in the row corresponding to the sample and in the column corresponding to the marker.

3.3.9 Sampled haplotype file [*k*.sample.gz]

Sampled haplotype files are generated only when the nimputations command line argument is used (see Section 3.2.2). If nimputations=K is specified (K = 1, 2, …), each input Beagle file will have a K phased output files with names ending in "*k*.sample.gz" for $k = 1, 2, \ldots, K$. Each sampled haplotype file is a phased Beagle genotypes file (see Section 2.1). A sampled haplotype file will have phased unrelated data, phased parent-offspring trio data, or phased parent-offspring pair data, according to the data in the input Beagle file. In contrast to the phased output file which gives the most likely haplotypes for each individual, parent-

offspring pair, or parent-offspring trios (see Section 3.3.2), each sampled haplotype output file gives randomly sampled haplotypes. The randomly sampled haplotypes are sampled conditional on the genotypes for the individual, parent-offspring pair, or parent-offspring trio and the haplotype frequency model. A filename prefix for the output sampled haplotype files can be set with the out command line argument (see Section 3.2.1).

4 Association testing with BEAGLE

BEAGLE can perform single marker and haplotypic tests for association for case-control studies and for parent-offspring trio studies with affected offspring. BEAGLE can also calculate multiple-testing adjusted P-values using permutation of the trait status. BEAGLE performs haplotypic association testing by building a graphical model of haplotype frequencies, clustering haplotypes that are similar near each marker, and testing the haplotype clusters for association with the trait status.

BEAGLE can be used in conjunction with the PRESTO software package (http://faculty.washington.edu/browning/presto/presto.html) and the pseudomarker utility program (see Section 7.2). PRESTO can compute empirical distributions of order statistics, analyze stratified data, and determine significance levels for one and two-stage genetic association studies.

4.1 Quick start guide

To run BEAGLE, enter the following command at the computer prompt:

```
java -Xmx<Mb>m -jar beagle.jar <arguments>
```

where <Mb> is the number of Megabytes of memory available (e.g. -Xmx1000m) and <arguments> is a space separated list of arguments. Each argument has the format **parameter**=**value**. There is no white-space between the **parameter** and **=** or between **=** and the **value**. For large data sets with thousands of samples, genetic association analysis may require several hundred megabytes of memory (see Section 5.2).

The commands for building a graphical model of haplotype frequencies and for performing single marker and haplotypic association tests are very simple. New BEAGLE users should initially use only the arguments for specifying the data file, the trait, the output file prefix, and the association tests illustrated in the three example command lines given below. Other BEAGLE arguments are optional and have sensible default values.

Here are three examples command lines for performing association tests or building a graphical model of haplotype frequencies.

1. java -Xmx800m -jar beagle.jar data=data.bgl trait=T2D out=example
2. java -Xmx800m -jar beagle.jar data=data.bgl trait=CD test=adr out=example
3. java -Xmx800m -jar beagle.jar data=data.bgl out=example

The input file for these examples is a phased Beagle file called data.bgl. The input phased Beagle file cannot have missing data. Line 1 performs allelic tests for association with an affection status variable named T2D. Line 2 performs allelic (a), dominant (d), and recessive (r) tests for association with an affection status variable named CD. Line 3 builds a graphical model of haplotype frequencies without performing association testing.

BEAGLE can also perform association testing with transmitted and untransmitted haplotypes from parent-offspring data with affected offspring. For details, please see the discussion at the end of Section 2.1.

4.2 Command line arguments

This section describes the different BEAGLE command line arguments for performing association testing.

4.2.1 Argument for specifying input data

- data=<phased Beagle file> where <phased Beagle file> is the name of a Beagle file containing phased data with no missing alleles. The phased data can be any combination of phased unrelated data, phased parent-offspring trio data, or phased parent-offspring pair data (see Section 2.1). The data argument is required.
- trait=<affection status identifier> where <trait> is the name of the affection status variable to use for association testing. The trait argument is optional. If the trait argument is omitted, then a graphical model of haplotype frequencies will be created and written to a .dag file (see Section 4.3.4), but association testing will not be performed.
- out=<output file prefix> where <output file prefix> is the prefix for the output files. The different output files are described in Section 4.3. The out argument is required. The output prefix must be an absolute or relative filename, but it cannot be a directory. The output files corresponding to an input file called z.bgl will be [output file prefix].z.bgl.[ext] where [ext] describes the data in the output file.

4.2.2 Arguments for building the model

The scale and shift parameters control the number of haplotype clusters in the graphical model for the phased data. Having too many or too few haplotype clusters can reduce the power of association testing. The default values for the scale and shift parameters have performed well in simulation studies and real data analyses.[4, 5]

The scale and shift parameters determine whether pairs of nodes will be merged during construction of the graphical model. If node A represents n_A haplotypes and node B represents n_B haplotypes, the two nodes will not merge if their similarity score [4, 6] is greater than

$$m\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}+b$$

where is m the scale parameter and b is the shift parameter.

- scale=<threshold scale> where <threshold scale> is a positive number giving the scale parameter used when deciding whether to merge a pair of nodes.[4] The scale argument is optional. The default value is scale=4.0. Decreasing the scale parameter will increase the number of haplotype clusters in the graphical model.
- shift=<threshold shift> where <threshold shift> is nonnegative number less than or equal to 1.0 giving the shift parameter used when deciding whether to merge a pair of nodes.[4] The shift argument is optional. The default value is shift=0.2. Decreasing the shift parameter will increase the number of haplotype clusters in the graphical model.

4.2.3 Arguments for association testing

- test=<association tests> where <association tests> is one or more characters from the set "ardo" where
  - a = allelic test.
  - r = recessive test (groups major allele homozygotes and heterozygotes).
  - d = dominant test (groups minor allele homozygotes and heterozygotes).
  - o = overdominant test (groups minor and major allele homozygotes).
- For example test=a or test=ardo. Fisher's exact test for 2 x 2 table is used for each test. If a marker has more than two alleles, each allele defines a diallelic marker by grouping the other alleles. Thus a triallelic marker will define 3 diallelic markers and these diallelic markers will be tested. The test argument is optional. The default value is test=a. Genotypic tests (r, d, and o) are not permitted when diplotypes=false. See the diplotypes argument in this section for more details.
- seed=<random seed> where <random seed> is an integer seed for the random number generator. The seed argument is optional. The default value is seed=-99999.
- nperms=<number of permutations> where <number of permutations> is a nonnegative integer giving the number of permutations of the affection status to use for permutation testing. Permutation testing determines multiple-testing adjusted P-values for all the haplotypes and single markers that are tested (see Section 4.3.2). You can skip permutation testing by setting nperms=0. The nperms argument is optional. The default value is nperms=1000. The computation time for permutation testing is linear in the number of permutations.
- edgecount=<minimum edge count> where <minimum edge count> is a positive integer giving the minimum number of haplotypes in a haplotype cluster that is required to test the cluster for association with a trait. Each haplotype cluster is defined by an edge of the graphical model.[6] The edgecount argument is optional. The default value is edgecount=20. The default value should work well, but expert users may want to adjust the edgecount parameter based on the sample size and a priori knowledge of the disease allele frequency.
- othercount=<minimum other count> where <minimum other count> is a nonnegative integer giving the minimum number of haplotypes in the set of edges that merge with an edge E that is required to test the haplotype cluster defined by E. Edges E_1 and E_2 are said to merge if E_1 and E_2 point to the same child node.[6] The othercount argument is optional. The default value is othercount=1. Increasing the othercount parameter will decrease the number of haplotype clusters (i.e. edges of the graphical model) tested for association with the affection status. A value of 0 is permitted but generally is not recommended because it will result in testing non-merging edges.
- diplotypes=<true/false> where <true/false> is true if the affection status is permuted for haplotype pairs so that both haplotypes for each individual have the same permuted trait status and <true/false> is false if the affection status is permuted for individual haplotypes (rather than for haplotype pairs). Only the allelic test can be performed when diplotypes=false (see the test argument in this section). The diplotypes argument is

optional. The default value is diplotypes=true. The diplotypes argument generally should not be used unless the phased data consist of transmitted and untransmitted haplotypes, and the affection status identifies the transmitted and untransmitted haplotypes as described at the end of Section 2.1.

4.2.4 Advanced options not intended for general use

- maxwindow=<maximum window size> where <maximum window size> is a positive integer giving the maximum number of consecutive markers that will be considered when building the graphical model of haplotype frequencies. The maxwindow argument is optional. The default value is maxwindow=500. The default value should be sufficiently large for most data sets. For example, if your marker density is one marker per kilobase, BEAGLE will consider markers in a 500 kilobase window when using the default maxwindow parameter.

4.3 Output files

Depending on the command line arguments for the analysis, BEAGLE can produce up to four output files. Most users will be interested in only two output files: the log file (.log), and the P-value file (.pval). The remaining two files, the null distribution file (.null) and the model file (.dag), will be useful to some, but not most, users.

4.3.1 The log file [.log]

A log file is generated each time BEAGLE is run. The log file gives a summary of the analysis that includes the BEAGLE version, list of the command line arguments for the analysis, and the running time for the analysis.

If the trait parameter is specified, the log file gives a list of all markers or haplotype clusters from the P-value file (see Section 4.3.2) with a permutation P-value < 0.2.

4.3.2 The P-value file [.pval]

The P-value file records the P-values from testing markers and haplotype clusters for association with the affection status. Haplotype clusters are defined by edges of a graphical model. Each haplotype defines a path between the initial and terminal node of the graph.[6] For each edge E in the graphical model we define a haplotype cluster C_E to be the set of all haplotypes whose path from initial to terminal node traverses edge E. Haplotype clusters define diallelic markers, that we call **pseudomarkers**. Given a haplotype cluster C, the diallelic pseudomarker for a haplotype is 2 if the haplotype is in C and 1 if the haplotype is not in C. Representing haplotype clusters as pseudomarkers enables us to test the haplotype cluster for association with the trait status in the same way we test other diallelic markers: using Fisher's exact test for 2 x 2 tables. The pseudomarker program (see Section 7.2) can create a phased Beagle file of pseudomarkers representing the haplotype clusters in the graphical model. The phased Beagle file of pseudomarkers can be imported into a program like R[1] for statistical analysis. The first line of the P-value file is a header line describing the columns of the file. Each line (except the header line) gives the P-values from testing one marker or pseudomarker for association with the trait status. First, the P-values for the genotyped markers are given, then the P-values for the pseudomarkers are given. The edgecount and othercount parameters described in Section 4.2.3 determine which pseudomarkers are selected for testing.

Example 4 - First and last lines of a P-value file

Marker	Allele	allelic_p	min _p	min_p_perm
m14954	0	0.4065	0.4065	1.000
m14992	1	0.1064	0.1064	1.000
m15081	1	0.7968	0.7968	1.000

...skipped lines of P-value file...

m28524	0.1	0.9121	0.9121	1.000
m28524	1.0	0.3161	0.3161	1.000
m28662	1.0	0.3266	0.3266	1.000
m28662	0.0	0.7873	0.7873	1.000
m28662	0.1	0.3983	0.3983	1.000

The first field on the line is the marker identifier. The second field is the marker allele that is tested. If the marker has more than two alleles, the allele is used to define a diallelic marker by grouping all other alleles as the second allele (see the test parameter in Section 4.2.3). For pseudomarkers defined by graph edges, the allele field has the format "parent.allele" where parent is the parent node number and allele is the marker allele for the edge (see the last 5 lines of Example 4). The marker identifier, the parent node number, and the marker allele uniquely determines the edge of the graphical model[4, 6] (see Section 4.3.4). After the marker field and marker allele field, the next columns give the P-values for allelic, recessive, dominant, and overdominant tests. Columns corresponding to tests which were not performed are omitted (see the test command line argument in Section 4.2.3). The second-to-last column gives the minimum P-value observed for the marker. If only one test is performed, as is the case in the preceding P-value file excerpt (Example 4), the minimum P-value equals the P-value for that test. The final column gives the permutation P-value.

The permutation P-value is a measure of significance that accounts for multiple testing. For example, if your significance level is $\alpha = 0.05$, and a marker allele or pseudomarker allele has a permutation P-value $p < 0.05$, the association is significant after accounting for multiple testing. More generally, for a given significance level α $(0 \leq \alpha \leq 1)$, the probability of observing one or more marker alleles or pseudomarker alleles with a permutation P-value less than α is less than or equal to α under the null hypothesis that the trait and marker data are independent.[7]

Given a set of tests (specified with the test parameter), the permutation test randomly permutes the trait status and tests the marker alleles and pseudomarker alleles for association with the permuted trait status. If diplotypes=true, which is the default setting, the trait status is permuted for the individuals so that both haplotypes for each individual have the same permuted trait status. When the data consists of transmitted and untransmitted haplotypes, the diplotypes=false argument must be used so that the trait status is permuted for the haplotypes rather than for the individuals (see Section 5.4).

For each permuted trait status, the set of tests (determined by the test parameter) is applied to all markers and the minimum P-value (minimized over all markers and pseudomarkers and all tests) is saved and written to the null distribution file (see Section 4.3.3). If a marker or pseudomarker has a minimum P-value of p_0 (minimized over all tests for that marker or pseudomarker) when tested for association with the unpermuted trait status, and if for k out

of N permutations of the trait status there exists at least one marker or pseudomarker with a minimum P-value less than or equal to p_0 when tested for association with the permuted trait status, the permutation P-value for the marker is $(k+1)/(N+1)$.[7] Under the null hypothesis, expect most alleles to have a permutation P-value of 1.000 since the P-value of a single allele is being compared to the minimum P-value from all alleles of all markers. The P-value file is designed to be imported into a spreadsheet or a statistical software package. However, if you want to quickly identify the most significant markers, look in the output log file. The log file contains a list of all markers and pseudomarkers with a permutation P-value < 0.2.

4.3.3 The null distribution file [.null]

The null distribution file lists the minimum P-values (minimized over all tests for all markers and pseudomarkers) observed when testing the marker and pseudomarker alleles for association with a randomly permuted trait status. The P-values in the null distribution file are used to determine the multiple-testing adjusted P-value given in the P-value file (see Section 4.3.2). The j-th line gives the minimum P-value from the j-th permuted trait status for $j = 1,2,\ldots,N$ where N equals the value of the nperms parameter (see Section 4.2.3). The null distribution file gives an empirical distribution of the minimum P-value under the null hypothesis. If the argument nperms=0 is used, the null distribution file will be empty. The sequence of permutations of the trait status is determined by the seed argument (see Section 4.2.3).

4.3.4 The model file [.dag.gz]

The model file gives the graphical model for the haplotype clusters. The model output file is GZIP-compressed. The graphical model is a directed acyclic graph (DAG): levels of the DAG correspond to markers and edges of the DAG correspond to haplotype clusters.[4, 6]

The first line of the file is a header line describing the columns in the file. Each line describes an edge of the graph. Edges corresponding to the same marker are grouped together and preceded and succeeded by a blank line. The first columns and lines of the model file look like this:

Level	Marker	Parent	Child	Allele	Count	Haplotype identifiers						
0	m14954	0	0	0	2964	0	1	5	6	7	8	...
0	m14954	0	1	1	1036	2	3	4	10	12	15	...
1	m14992	0	0	1	2207	0	5	6	7	8	14	...
1	m14992	0	1	0	757	1	9	11	13	27	28	...
1	m14992	1	2	1	1036	2	3	4	10	12	15	...

The first six fields in a line of the model file are:

1. The level of the parent node of the edge. If there are M markers, the level numbers are 0, 1, 2, ... , M − 1.
2. The marker identifier given in the input BEAGLE file corresponding to the level of the parent node.
3. The node number of the parent node of the edge. Node numbering begins at 0 for each level.

4. The node number of the child node of the edge. Node numbering begins at 0 for each level, and the level of the child node is one more than the level of the parent node.
5. The marker allele that is carried by all haplotypes in the cluster defined by the edge.
6. The number of haplotypes in the cluster defined by the edge.

If sixth field is *K*, then there are *6+K* fields on the line. The final *K* fields are nonnegative integers identifying the haplotypes in the phased Beagle file specified with the data argument (see Section 4.2.1) that are in each haplotype cluster. The haplotypes are numbered 0, 1, 2, ... in the order they appear as columns in the phased Beagle genotypes file (see Section 2.1), so 0 refers to column 3 in the phased Beagle genotypes file, 1 refers to column 4 in the phased Beagle file, and so on.

5 Using BEAGLE with large data sets

5.1 Memory management

You can increase the amount of memory available to the Java interpreter using the -Xmx command line argument. If [Mb] is a positive integer, then -Xmx[Mb]m sets the maximum amount of memory that will be used by the Java interpreter to [Mb] megabytes. It is helpful to set the -Xmx parameter somewhat higher than the minimum memory required to analyze your data because having the additional memory available can result in decreased computation time.

For association testing, memory usage depends on the number of individuals in the analysis, but is independent of the number of markers when analyzing more than 1000 markers. In general, association testing will not require more than 500-2000 Mb of memory.

For haplotype phase inference and imputation of missing data with default BEAGLE options, memory usage increases with the number of markers. It you need to reduce the amount of memory BEAGLE is using, try one or more of the following techniques:

1. Use the lowmem=true command line argument (see Section 3.2.2). The lowmem option makes memory requirements essentially independent of the number of markers.
2. Divide longer chromosomes into two parts or more parts, and infer haplotype phase and perform genotype imputation on each part separately. You may want to allow some overlap between sections since "edge effects" can reduce accuracy of the haplotype frequency model at the edges of a marker set.
3. If you are imputing ungenotyped markers, you can divide the sample into subsamples, and perform imputation on each subsample separately (but do not divide the reference panel). **If you divide the sample, individuals should be randomly assigned to subsamples so that the assignment is independent of the individual's phenotype**. It is okay to have cases and controls for a subsample in separate input files as long as cases and controls are analyzed together in the same Beagle analysis. A sample can be divided into subsamples by using the cut.jar or filtercolumns.jar utility program (see Section 7.1), or by specifying individuals to exclude from the subsample using the excludecolumns argument (see Section 3.2.2). Output files can be pasted together using the paste utility program (see Section 7.1). If all your samples are in a single file, and if the order of samples in your input file is randomized, a unix shell program called

divide.sample can be downloaded from the BEAGLE web site (see Section 1.2) to facilitate imputation analysis using subsamples of the data.

Long stretches of consecutive markers with completely missing data in a parent-offspring trio can require enormous amounts of memory. If a trio is missing all its genotypes in a genomic region, the region may need to be analyzed separately with the problematic trio removed.

5.2 Genomewide association studies

If you want to perform association testing on data from a genomewide association study, first divide the data by chromosome, and then phase the data for each chromosome separately. If your input data is divided among multiple input files, use the paste utility to combine the resulting phased Beagle files for a chromosome into a single phased Beagle file (see Section 7.1). When there are multiple input Beagle files, make sure the multiple phased output files are combined in the same order for each chromosome. After phasing each chromosome, concatenate the phased Beagle files for each chromosome (e.g. with the unix cat command), and test the resulting phased Beagle file (see Section 4). It is recommended that you test all markers in a single analysis so that BEAGLE can perform permutation testing to determine statistical significance.

6 Example BEAGLE analyses

6.1 Inferring haplotype phase and missing data

The BEAGLE software distribution includes sample files to illustrate genotype imputation. The following input files are contained in the example/imputation folder:

1. hapmap.markers - A markers file listing 100 markers on chromosome 1. The four columns give the marker identifier, the base position, and the two alleles.
2. hapmap.phased.bgl - A phased Beagle file with 58 phased individuals genotyped on the 100 markers in the hapmap.markers file. The data is from the HapMap CEU panel[8].
3. hapmap.unphased.bgl - An unphased Beagle file with 2 individuals genotyped on 50 markers in the hapmap.markers file. The data is from the HapMap CEU panel.

The missing 50 markers in hapmap.unphased.bgl can be imputed using the data in hapmap.phased.bgl as a reference panel. The BEAGLE command line is

```
java -Xmx500m -jar beagle.jar markers=hapmap.markers phased=hapmap.phased.bgl
unphased=hapmap.unphased.bgl missing=? out=example
```

Five output files are created by the preceding command line:

1. **example.hapmap.phased.bgl.phased.gz** - phased Beagle output file corresponding to the hapmap.phased.bgl input file (see Section 3.3.2). Any missing data in the input Beagle file is imputed using the most likely haplotypes.
2. **example.hapmap.unphased.bgl.phased.gz** - phased Beagle file corresponding to the hapmap.unphased.bgl input file (see Section 3.3.2). The phased output file gives the most likely haplotype pair for each individual.

3. **example.hapmap.unphased.bgl.gprobs.gz** - genotype probabilities file corresponding to the individuals in the hapmap.unphased.bgl file (see Section 3.3.3). Posterior genotype probabilities are given for all markers. The posterior genotype probability will be 1.0 for all assayed (i.e. non-imputed) genotypes.
4. **example.hapmap.unphased.bgl.r2** - allelic R^2 file corresponding to the hapmap.unphased.bgl file (see section 3.3.4). For each genotyped or imputed marker, the allelic R^2 file gives the estimated correlation between the allele dosage with highest posterior probability and the true allele dosage. In this example analysis, the allelic R^2 file can be ignored because the sample size (2 individuals) is too small to obtain accurate estimates.
5. **example.log** - log file summarizing the BEAGLE analysis (see Section 3.3.1).

6.2 Association testing

The BEAGLE software distribution includes sample files to illustrate single marker and haplotypic association testing using BEAGLE. The example/testing folder in the beagle_example.zip file (see Section 0) contains an input Beagle file with phased unrelated data called data.bgl. The example data are 200 markers for 4000 haplotypes (1000 case and 1000 control individuals), generated using Cosi version 1.0.[9] The affections status variable in the input file is named T2D. The command for association testing is

```
java -Xmx500m -jar beagle.jar data=data.bgl trait=T2D out=example
```

Four output files are created by the preceding command line:

1. **example.data.bgl.log** - a log file summarizing the association analysis (see Section 4.3.1)
2. **example.data.bgl.pval** - a P-value file with association test statistics and multiple-testing adjusted P-values for markers and haplotype clusters (see Section 4.3.2)
3. **example.data.bgl.null** - a null distribution file with minimum test statistics observed under the null hypothesis (see Section 4.3.3)
4. **example.data.bgl.dag.gz** - a model file describing the graphical model of haplotype clusters (see Section 4.3.4)

The output log file (data.bgl.log) contains the following excerpt listing the 3 marker alleles and 3 pseudomarkers alleles (i.e. haplotype clusters) that have permutation P-values less than 0.2:

Alleles with permutation P-values less than 0.2:

Marker	Allele	allelic_p	min_p	min_p_perm
m23508	0	0.0004167	0.0004167	0.1179
m23612	0	0.0002015	0.0002015	0.06593
m24828	0	0.0006412	0.0006412	0.1908
m23171	7.1	0.0002056	0.0002056	0.06593
m23892	10.1	4.559e-05	4.559e-05	0.01898
m24828	2.1	0.0006412	0.0006412	0.1908

6 alleles with permutation P-value < 0.2

One test was significant at the $\alpha = 0.05$ level after accounting for multiple testing (permutation P-value 0.01898). Allele 10.1 of marker m23892 is the haplotype cluster with parent node 10 and allele 1 at marker m23892. The haplotypes in this cluster are recorded in the output model file (example.data.bgl.dag.gz).

The allele sequence that defines the haplotypes in the associated haplotype cluster can be identified using the cluster2haps utility program (see Section 7.3). The cluster2haps command line is:

```
java -jar cluster2haps.jar dag=data.example.bgl.dag.gz phased=data.bgl trait=T2D
marker=m23892 parent=10 allele=1 out=m23892_10_1.haps
```

The output data from in the file m23892_10_1.haps shows that the cluster can be defined by many different allele sequences. One of the allele sequences defining the cluster is the allele sequence 1 - 0 - 1 for markers m23490 - m23628 - m23892.

7 Utility programs

In this section, we describe a suite of utility programs available at the BEAGLE Utilities web site (see Section 7.1) and two specialized utility programs, **pseudomarker** and **cluster2haps** that are available from the BEAGLE web site (see Section 1.2), and which may be useful to some users.

7.1 BEAGLE Utilities web site

Please take full advantage of the utility programs that are available at the BEAGLE Utilities web site:

http://faculty.washington.edu/browning/beagle_utilities/utilities.html

Many of these utility programs are useful for preparing input files and for processing output files. For example, there are utilities for extracting lines or columns from a file, for pasting files together, for transposing rows and columns of a file, and for converting between linkage file format and BEAGLE genotypes file format. The BEAGLE utilities can also be combined with standard unix utilities, including cat, zcat, head, tr, tail, cut, grep, sort, uniq, and wc.

7.2 pseudomarker

Haplotype clusters can be represented as diallelic pseudomarkers (see Section 4.3.2). The **pseudomarker** program converts the haplotype clusters in a BEAGLE output model file (see Section 4.3.4) to diallelic or multi-allelic markers and writes the markers to a Beagle phased genotypes file (see Section 2.1). The phased Beagle file of pseudomarker data can be imported into a statistical software package like "R"[1] for analysis using standard statistical methods such as logistic regression (to allow for covariates) or analysis of variance for quantitative trait data. The commands for running the **pseudomarker** program are

```
java -jar pseudomarker.jar <arguments>
```

where <arguments> is a space separated list of arguments. Each argument has the format **parameter=value**. There is no white-space between the **parameter** and **=** or between **=** and the **value**. The arguments are

- dag=<Beagle .dag file> where <Beagle .dag file> is the filename of a Beagle output model file (see Section 4.3.4). The dag argument is required. If the filename of the Beagle output model file ends in ".gz", the pseudomarker program will assume the file is compressed with the GZIP algorithm.
- out=<output file> where <output file> is the phased Beagle file that will be created. The markers of the output file will correspond to haplotype clusters in the specified Beagle model file. The out argument is required.
- edgecount=<minimum edge count> where <minimum edge count> is a positive integer giving the minimum number of haplotypes in a haplotype cluster defined by an edge E that is required to create a diallelic pseudomarker m_E. For more details, see Section 4.2.3. The edgecount argument is optional. The default value is edgecount=1.
- othercount=<minimum other count> where <minimum other count> is a nonnegative integer giving the minimum number of haplotypes on the set of edges that merge with an edge E that is required to create a diallelic pseudomarker m_E. For more details, see Section 4.2.3. Edges E_1 and E_2 are said to merge if E_1 and E_2 point to the same child node. The othercount argument is optional. The default value is othercount=0.

The edgecount and othercount arguments are similar to the arguments of the same name in Section 4.2.3, but their default values are different. With default edgecount and othercount parameters, the pseudomarker program creates **multi-allelic** pseudomarkers which correspond to **markers** in the specified model file. With a non-default value for the edgecount or othercount parameter, the pseudomarker program creates **diallelic** pseudomarkers which correspond to **edges** in the specified model file.

If the edgecount and othercount arguments are not used (or if their default values are specified), a multi-allelic pseudomarker is created for each level (i.e. marker) of the graphical model. In the output Beagle file, the multi-allelic pseudomarkers will have the same name and order as the markers in the specified model file. For a given marker, a haplotype has allele k (k = 1, 2, . . .), if it is in the k-th edge that connects the given marker with the next marker in the specified model file.

If at least one of the edgecount and othercount parameters is set to a non-default value, diallelic pseudomarkers are created for some of the edges. (This differs from the preceding paragraph, where multi-allelic psedomarkers are created for each marker). A diallelic pseudomarker is created for an edge only if the edge contains at least edgecount haplotypes and only if the edge merges with one or more edges whose haplotype clusters contain at least othercount haplotypes. The name of the diallelic pseudomarker has the format **marker_node.allele** where **marker** is the name of the marker that identifies the level of the parent node of the edge, **node** is the parent node number for the edge, and **allele** is the marker allele that labels the edge (see Section 4.3.2).[4, 6] A haplotype has allele 2 if it is in the haplotype cluster corresponding to the edge and has allele 1 if it is not in the haplotype cluster.

7.3 cluster2haps

The **cluster2haps** program is used after performing a haplotypic analysis with BEAGLE. The **cluster2haps** program can identify the sequence or sequences of alleles that are present in a haplotype cluster that is associated with a trait. The **cluster2haps** program prints out all

allele sequences that are present in a specified haplotype cluster for increasingly large marker windows, tests the allele sequences for association with the trait status, and reports the P-values.

The command line syntax for the program is similar to the syntax for Beagle:

```
java -Xmx600m -jar cluster2haps.jar <arguments>
```

where <arguments> is a space separated list of arguments. Each argument has the format **parameter=value**. There is no white-space between the **parameter** and = or between = and the **value**. The arguments are

- dag=<dag file> where <dag file> is the filename of the Beagle output model file (see Section 4.3.4) giving the graphical haplotype frequency model. The dag argument is required.
- phased=<phased BEAGLE file> where <phased BEAGLE file> is the filename of the phased Beagle file used to build the graphical haplotype frequency model of haplotype structure. The markers in the phased file must be in chromosomal order, and the phased Beagle file must contain an affection status line giving the affection status for each allele of each individual. The data argument is required.
- trait=<trait ID> where <trait ID> is the name of the affection status variable (A <trait ID> ...) used in the association analysis. The trait argument is required.
- out=<output file> where <output file> is the name of the output file for the analysis. The out argument is required.
- marker=<marker ID> where <marker ID> is the identifier of the marker locus for the haplotype cluster given in the P-value file (see Section 4.3.2). The marker argument is required.
- parent=<parent node> where <parent node> is the parent node number for the haplotype cluster. The parent node number is given before the dot (".") in the allele (second) column of the Beagle output P-value file (see Section 4.3.2). The parent argument is required.
- allele=<allele id> where <allele id> is the identifier of the marker allele for the haplotype cluster. The marker allele for the haplotype cluster is given after the dot (".") in the allele (second) column of the Beagle output P-value file (see Section 4.3.2). The allele argument is required.
- maxhaps=<maximum haplotypes> where <maximum haplotypes> controls the size of the marker windows that will be considered. The largest marker window considered will be the smallest window for which the number of distinct haplotypes in the clusters is greater than or equal to the maximum number of haplotypes. The maxhaps argument is optional. The default value is maxhaps=8.

cluster2hap has 7 required arguments: 3 arguments to specify files (dag, phased, and out), 1 argument to specify the affection status variable (trait), and 3 arguments to specify the haplotype cluster (marker, parent, and allele).

The output file gives:

1. The allele sequences that are contained in the haplotype cluster

2. For each allele sequence in 1, the number of haplotypes with the allele sequence.
3. For each allele sequence in 1, the number (and proportion) of haplotypes with the allele sequence that are in the haplotype cluster
4. For each allele sequence in 1, the P-value from Fisher's exact allelic test of association of the allele sequence with the trait.

Example 5 - A sample cluster2haps analysis

Here is an example using quality-control filtered Wellcome Trust Case Control consortium type 1 diabetes data for the IL2RA region.[10] The line in the Beagle output P-value file (see Section 4.3.2) for the most significantly haplotype cluster in the region is

Marker	Allele	allelic_p	min_p	min_p_perm
rs12722489	2.C	5.104e-09	5.104e-09	0.0009990

We can use cluster2haps to analyze this significantly associated haplotype cluster. The command line for the cluster2haps program is:

```
java -jar cluster2haps.jar dag=ex.dag data=ex.phased trait=T1D out=cluster.out
marker=rs12722489 parent=2 allele=C
```

The cluster localizes to the marker rs12722489. The "Allele" field for the haplotype cluster gives the parent node number (2), and the allele (C) for the haplotype cluster. In this example the model file was "ex.dag", the Beagle file of phased haplotypes was "ex.phased", the name of the affection status variable was "T1D", and the output file was "cluster.out".

An excerpt of the output file from this **cluster2haps** analysis is given below. The excerpt contains results for the three shortest marker windows. Each marker window ends with the marker specified with the marker command line argument (rs12722489).

Excerpt from cluster2haps output:

```
9802 haplotypes:        3926  cases / 5876 controls
                         971  in cluster / 8831 not in cluster

Markers:  rs12722489
allele sequence                 count          # in cluster          P-value
C                                8147          971  (11.9%)           0.492

Markers:  rs17149458  rs12722489
allele sequence                 count          # in cluster          P-value
T C                              8147          971  (11.9%)           0.492

Markers:  rs2104286  rs17149458 rs12722489
allele sequence                 count          # in cluster          P-value
C T C                             989          971  (98.2%)           1.02e-08
```

From the preceding output excerpt, we see that for the 3 shortest marker windows there is only one allele sequence present in the haplotype cluster. For the C-T-C allele sequence (for

markers rs2104286 rs17149458 rs12722489), 971 of the 989 of the haplotypes with this allele sequence are in the haplotype cluster, and the C-T-C allele sequence is strongly associated with type 1 diabetes ($p = 1.02 \times 10^{-8}$). Thus the C-T-C haplotype accounts for nearly the entire association signal for the haplotype cluster. The output also shows that the haplotype counts for the C and T-C allele sequences are the same. Thus the middle T allele in the C-T-C sequence is not necessary (due to linkage disequilibrium), and the C-C haplotype (rs2104286-rs12722489) gives a simpler characterization of the haplotype cluster.

8 References

1. R Development Core Team, *R: A Language and Environment for Statistical Computing*. 2006, Vienna, Austria: R Foundation for Statistical Computing.
2. Li, Y., J. Ding, and G. Abecasis, *Mach 1.0: rapid haplotype reconstruction and missing genotype inference,* in *the 56th Annual Meeting of The American Society of Human Genetics*. 2006: New Orleans, Louisiana.
3. Browning, B.L. and S.R. Browning, *A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals.* Am J Hum Genet, 2009. **84**(2): p. 210-23.
4. Browning, B.L. and S.R. Browning, *Efficient multilocus association testing for whole genome association studies using localized haplotype clustering.* Genet Epidemiol, 2007. **31**(5): p. 365-75.
5. Browning, B.L. and S.R. Browning, *Haplotypic analysis of Wellcome Trust Case Control Consortium data.* Hum Genet, 2008. **123**(3): p. 273-80.
6. Browning, S.R., *Multilocus association mapping using variable-length Markov chains.* Am J Hum Genet, 2006. **78**(6): p. 903-13.
7. Besag, J. and P. Clifford, *Sequential Monte-Carlo p-values.* Biometrika, 1991. **78**(2): p. 301-304.
8. The International HapMap Consortium, *A second generation human haplotype map of over 3.1 million SNPs.* Nature, 2007. **449**(7164): p. 851-61.
9. Schaffner, S.F., et al., *Calibrating a coalescent simulation of human genome sequence variation.* Genome Research, 2005. **15**(11): p. 1576-83.
10. The Wellcome Trust Case Control Consortium, *Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.* Nature, 2007. **447**(7145): p. 661-78.