Beagle 4.1

Brian L. Browning

Department of Medicine

Division of Medical Genetics

University of Washington

January 21, 2017

Contents

1 Introduction

Beagle version 4.1 has two major improvements: a more accurate haplotype phasing algorithm, and a very fast and accurate genotype imputation algorithm. Beagle version 4.1 also implements the Refined IBD algorithm for detecting homozygosity-by-descent (HBD) and identity-by-descent (IBD) segments. Version 4.1 does not model parent-offpspring relationships (this will be added later), but version 4.0 is available if you need this capability.

Beagle 4.1 requires a version 1.8 (or later) Java interpreter. Java will print an "Unsupported Class Version" error, if you use an earlier version of Java to run Beagle. Enter "java –version" at the unix command line prompt to check if a java interpreter is installed on your system. The most recent Java interpreter can be downloaded from **www.java.com**.

The Beagle software program is freely available and can be downloaded from the Beagle web site:

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

1.1 Citing Beagle

If you use Beagle and publish your analysis, please report the program version and cite the appropriate article:

The citation for Beagle's phasing algorithm is:

> S R Browning and B L Browning (2007). Rapid and accurate haplotype phasing and missing data inference for whole genome association studies by use of localized haplotype clustering. Am J Hum Genet 81:1084-97. doi:10.1086/521987

The citation for Beagle's imputation algorithm is:

> S R Browning and B L Browning (2016). Genotype imputation with millions of reference samples. Am J Hum Genet 98:116-126. doi:10.1086/j.ajhg.2015.11.020

The citation for Beagle's IBD detection algorithm is:

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

1.2 Variant Call Format

Beagle uses Variant Call Format (VCF) 4.2 for input and output file. VCF files can be manipulated and analysed with VCFtools, PLINK/SEQ, and the Beagle Utilities.

Beagle assumes that any input VCF file that has a name ending in ".gz" is compressed with gzip or bgzip, and that any reference VCF file that has a name ending in "bref" is compressed with the bref program. Output VCF, IBD, and HBD files are compressed with bgzip and can be uncompressed with the unix gunzip program.

X chromosome: version 4.1 requires male non-pseudoautosomal X-chromosome genotypes to be coded as homozygous diploid genotypes.

2 Command line arguments

To run Beagle version 4.1, enter the following command at the computer prompt:

java –Xss5m –Xmx[GB]g –jar beagle.*version*.jar *arguments*

where [GB] is the maximum permitted size of the memory pool in gigabytes (e.g. –Xmx8g), *version* is the Beagle version code (eg. "01Oct15.6a3"), and *arguments* is a space separated list of arguments. Each argument has the format **parameter=value**. There is no white-space between the parameter, equal sign, and parameter value. The java –Xss5m option sets the default stack size to 5 Mb. The –Xss option can be omitted unless you encounter a stack overflow error for your data set.

There are only two required command line arguments: a **gt, gl** or **gtgl** argument to specify the input file and type of input data, and an **out** argument to specify the output file prefix. If you are imputing ungenotyped markers, the **map** parameter is recommended.

Use the **gl** or **gtgl** argument if you want to estimate posterior genotype probabilities. The input data can be genotype likelihood data (**gl**) or a combination of genotype likelihood and genotype data (**gtgl**). The estimated genotypes in the output VCF file will be unphased. You can phase the genotypes in the output VCF file by running Beagle again and setting the **gt** argument equal to the output VCF file.

A reference panel can be specified with the **ref** parameter. All genotypes in the reference panel must be non-missing and phased. Corresponding markers in the reference and target VCF files must have identical CHROM, POS, REF, and ALT fields. Before using a reference panel, you may need to run the conform-gt program to adjust the genomic position, allele order and chromosome strand of the markers in your data to match the reference panel.

Markers that are in the reference panel but not in the target data will be imputed when the **gt** argument is used. If you do not wish to impute ungenotyped markers in the target data, use the **impute=false** argument.

For IBD segment detection, use the **gt** argument with the **ibd=true** and **impute=false** options. For best results, you may need to use the **ibdtrim** argument.

2.1 Arguments for specifying data

- **gt=**[file] specifies a VCF file containing a GT (genotype) format field for each marker. If a genotype contains the phased allele separator, '|', then Beagle will preserve the phase of the genotype during the analysis. If you use the **gt** argument, all genotypes in the output file will be phased and non-missing.
- **gl=**[file] specifies a VCF file containing a GL or PL (genotype likelihood) format field for each marker. Any data in the GT format field will be ignored. If both GL and PL format fields are present for a marker, the GL format will be used.
- **gtgl=**[file] specifies a VCF file containing a GT, GL or PL format field for each marker. If a genotype is non-missing, Beagle will ignore the genotype likelihood. If both GL and PL format fields are present for a marker, the GL field will be used.
- **ref=**[file] specifies a VCF file containing phased reference genotypes. See the **impute** parameter.
- **out=**[prefix] specifies the output filename prefix. The prefix may be an absolute or relative filename, but it cannot be a directory name.

- **excludesamples=**[file] specifies a file containing non-reference samples (one sample per line) to be excluded from the analysis and output files.
- **excludemarkers=**[file] specifies a file containing markers (one marker per line) to be excluded from the analysis and the output files. An excluded marker identifier can either be an identifier from the VCF record's ID field or a genomic coordinate in the format: CHROM:POS.
- **map=**[file] specifies a PLINK format genetic map on the cM scale. HapMap GrCh36 and GrCh37 genetic maps in PLINK format are available for download from the Beagle website. Use of a genetic map is recommended if you are imputing ungenotyped markers. If no genetic map is specified, Beagle will assume a constant recombination rate of 1 cM / Mb.
- **chrom=**[chrom:start-end] specifies a chromosome or chromosome interval using a chromosome identifier in the VCF file and the starting and ending positions of the interval. The entire chromosome, the beginning of the chromosome, and the end of a chromosome can be specified by **chrom=**[chrom], **chrom=**[chrom:-end], and **chrom=**[chrom:start-] respectively.
- **maxlr=**[number ≥ 1] specifies the maximum likelihood ratio (default: **maxlr=5000**) at a genotype. If M is the maximum of the likelihoods of each possible genotype, any likelihood that is less than (M/**maxlr**) is set to 0.0 to improve computational efficiency.

2.2 Other arguments

- **nthreads=**[positive integer] specifies the number of threads of execution. If no **nthreads** parameter is specified, the **nthreads** parameter will be set equal to the number of CPU cores on the host machine.
- **lowmem=**[true/false] specifies whether a memory efficient algorithm should be used. The memory efficient algorithm increases run-time by a factor less than 2.0 (default: **lowmem=false**).
- **window=**[positive integer] specifies the number of markers to include in each sliding window (default: **window=50000**). The **window** parameter must be at least twice as large as the **overlap** parameter. The **window** parameter controls the amount of memory used in the analysis. For human data, I recommend that the **window** parameter be greater than or equal to the typical number of markers in 5 cM.
- **overlap=**[positive integer] specifies the number of markers of overlap between sliding windows (default: **overlap=3000**). For human data, I recommend that the overlap be set to the typical number of markers in 0.5 cM (when **ibd=false**) or two times the **ibdcm** parameter (when **ibd=true**).
- **niterations=**[nonnegative integer] specifies the number of phasing iterations (default: **niterations=5**). The phasing iterations are preceded by 10 burn-in iterations which carry out the Beagle version 4.0 phasing algorithm. If you want to phase your data with the Beagle 4.0 phasing algorithm, use **niterations=0**. Accuracy and compute time increase with the number of iterations.

- **impute=**[true/false] specifies whether markers that are present in the reference panel but absent in your data will be imputed (default: **impute=true**). This option has no effect if the **ref** and **gt** arguments are not used.
- **cluster=**[non-negative number] specifies the maximum cM distance between individual markers that are combined into an aggregate marker when imputing ungenotyped markers (default: **cluster=0.005**).
- **gprobs=**[true/false] specifies whether a GP (genotype probability) format field will be included in the output VCF file when imputing ungenotyped markers (default: **gprobs=false**). By default, a GP fields is not printed because a DS (alternate allele dose) format field is always printed when imputing ungenotyped markers.
- **ne=**[integer] specifies the effective population size when imputing ungenotyped markers. The default value is suitable for a large outbred human population (default: **ne=1000000**). Smaller values in the hundreds or thousands for the **ne** parameter are suggested for inbred human and animal populations.
- **err=**[nonnegative number] specifies the allele miscall rate. The default value should give good results for most sequence and SNP array data (default: **err=0.0001**).
- **seed=**[integer] specifies the seed for the random number generator (default: **seed=-99999**).
- **modelscale=**[positive number] specifies the model scale parameter when sampling haplotypes for unrelated individuals (default: **modelscale=0.8**). Increasing the **modelscale** parameter will trade reduced phasing accuracy for reduced run-time. However, when estimating posterior probabilities from genotype likelihood data, increasing the modelscale parameter could improve both accuracy and run-time.

2.3 Identity by descent detection arguments

- **ibd=**[true/false] specifies whether IBD analysis will be performed when the **gt** argument is used (default: **ibd=false**).
- **ibdlod=**[non-negative integer] specifies the minimum LOD score for reported IBD (default: **ibdlod=3.0**).
- **ibdcm=**[positive number] specifies the minimum length in cM of shared haplotypes that are reported in the output IBD file (default: **ibdcm=1.5**). If a genetic map is not specified with the **map** argument, the cM position of a marker will be estimated by dividing the position coordinate by 1,000,000.
- **ibdscale=**[non-negative number] specifies the scale parameter used to build the haplotype frequency model for IBD analysis. If no **ibdscale** parameter is specified the scale parameter for the IBD analysis will be set to $\max\left\{2, \sqrt{[\text{sample size}]/100}\right\}$, which we have found to work well for outbred populations.
- **ibdtrim=**[non-negative integer] specifies the number of markers trimmed from the end of a shared haplotype when testing for IBD (default: **ibdtrim=40**).

  Note: The default **ibdtrim** parameter is designed for European samples genotyped with a 1M SNP array (~ 1 marker per 3 kb). For human SNP array data, I suggest setting the

ibdtrim parameter to the typical number of markers in a 0.15 cM region. Pilot studies of randomly selected genomic regions can be used to fine-tune the values of the **ibdtrim** parameter.

3 Output files

All output filenames begin with the output file prefix specified on the command line. Output filenames end with the suffixes: **.log**, **.vcf.gz**, **.hbd.gz**, and **.ibd.gz**.

The **log** file gives a summary of the analysis that includes the Beagle version, the command line arguments, and the running time.

The output **VCF** file contains information for all non-reference samples in the analysis. Estimated haplotypes are reported in the GT format field as phased genotypes. If ungenotyped markers are imputed, the INFO field will contain an "IMP" flag, the estimated squared correlation, and the estimated alternate allele frequencies in the target samples.

An **HBD** file and an **IBD** file are produced when the **ibd=true** option is specified (see Section 2.3). Each line of an HBD/IBD output file has 8 fields and represents a detected HBD/IBD segment:

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