Beagle 5.5

Brian Browning

Department of Medicine

Division of Medical Genetics

University of Washington

December 17, 2024

Contents

1 Introduction

Beagle is a program for phasing and imputing missing genotypes. Sporadic missing genotypes are imputed during phasing. If a reference panel of phased genotypes is specified with the **ref** argument, ungenotyped markers that are present in the reference panel can also be imputed. Beagle version 5 does not infer genotypes from genotype likelihood input data, but Beagle versions 4.0 and 4.1 have this capability.

2 Citing Beagle

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

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

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

The Beagle 5.5 phasing algorithm is described in:

> B L Browning, X Tian, Y Zhou, and S R Browning (2021). Fast two-stage phasing of large-scale sequence data. Am J Hum Genet 108(10):1880-1890. doi:10.1016/j.ajhg.2021.08.005

The Beagle 5.5 imputation algorithm is described in:

> B L Browning, Y Zhou, and S R Browning (2018). A one-penny imputed genome from next generation reference panels. Am J Hum Genet 103(3):338-348. doi:10.1016/j.ajhg.2018.07.015

3 Command line arguments

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

To run Beagle, enter the following command at the command prompt:

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

where [*GB*] is an upper bound on the memory pool in gigabytes (e.g. –Xmx50g), and [*arguments*] is a space separated list of parameter values, each having the format **parameter=value**.

There are only two required command line arguments: a **gt** argument to specify the input file with genotype data and an **out** argument to specify the output file prefix. However, the **map** argument is recommended. It may be beneficial to specify an appropriate effective populations size with the **ne** argument if you are imputing ungenotyped markers in a small or inbred population.

The **window** and **window-markers** parameter controls the amount of computer memory that is used. Shorter windows require less memory. The **iterations** parameter controls the trade-off between compute time and phase accuracy. The default values of the **window** and **iterations** parameters perform well in most cases.

A reference panel with phased, non-missing genotypes can be specified with the **ref** parameter. Corresponding markers in the reference and target VCF files must have identical CHROM, POS, REF, and ALT fields, and must have the same order in both files. 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.

The recommended file format for the **ref** parameter is bref3 format (.bref3) because this format gives the fastest computation time. Tools for converting between VCF and bref3 are available on the Beagle web page. Beagle will also accept reference files in Variant Call Format (.vcf) and gzip-compressed VCF format (.vcf.gz).

Sporadic missing genotypes are imputed during haplotype phasing. If a reference panel is used, markers that are not present in the study sample but are present in the reference panel will be imputed after haplotype phasing. If you do not wish to impute ungenotyped markers, use the **impute=false** argument. If a reference panel is used, any markers that are not in the reference panel are excluded from the output VCF file.

3.1 Data parameters

- **gt=**[file] specifies a VCF file containing genotypes for the study samples. Each VCF record must contain a GT (genotype) format field. If any heterozygote genotype is unphased (with '/' allele separator) in a marker window, Beagle 5.4 will consider all heterozygote genotypes to be unphased, regardless of the allele separator used ('|' or '/').
- **ref=**[file] specifies a reference panel in bref3 or VCF format. Each genotype must have two phased, non-missing alleles. If a VCF file is specified, the phased allele separator must be used '|'.
- **out=**[string] specifies the output filename prefix. The prefix may be an absolute or relative filename, but it cannot be a directory name.
- **map=**[file] specifies a PLINK format genetic map with cM units. HapMap genetic maps in PLINK format for GRCh36, GRCh37, and GRCh38 are available for download. Beagle uses linear interpolation to estimate genetic positions between map positions. If no genetic map is specified, Beagle assumes a constant recombination rate of 1 cM per Mb.
- **chrom=**[chrom]**:**[start]-[end] specifies a chromosome interval: [chrom] is the CHROM field in the input VCF file and [start] and [end] are the starting and ending positions. The entire chromosome, the beginning, or the end may be specified by **chrom=**[chrom], **chrom=**[chrom]**:-**[end], and **chrom=**[chrom]**:**[start]**-** respectively.
- **excludesamples=**[file] specifies a file containing samples (one sample identifier per line) to be excluded from the analysis.

- **excludemarkers=**[file] specifies a file containing markers (one marker per line) to be excluded from the analysis. Each line of the file can be either an identifier from a VCF record's ID field or a genomic coordinate in the format: CHROM:POS.

3.2 Phasing parameters

- **burnin=**[positive number] is the maximum number of burnin iterations used to estimate an initial haplotype frequency model for inferring genotype phase (default: **burnin=3**).
- **iterations=**[positive number] is the number of iterations used to estimate genotype phase (default: **iterations=12**). Increasing this parameter will trade increased computation time for increased phasing accuracy.
- **phase-states=**[positive number] is the number of model states used to estimate genotype phase (default: **phase-states=280**).

3.3 Imputation parameters

- **impute=**[true/false] specifies whether markers that are present in the reference panel but absent in that target will be imputed (default: **impute=true**). This option has no effect if no reference panel is specified (see **ref** argument).
- **imp-states=**[positive number] is the number of model states used to impute ungenotyped markers (default: **imp-states=1600**).
- **imp-segment=**[positive number] is the minimum cM length of haplotype segments that will be incorporated in the HMM state space for a target haplotype (default: **imp-segment=6.0**).
- **imp-step=**[positive number] is the length in cM of the step in centiMorgans used for detecting short IBS segments (default: **step=0.1**).
- **imp-nsteps=**[integer > 1] is the number of consecutive steps (see **imp-step** argument) that will be considered when detecting long IBS segments (default: **imp-nsteps=7**).
- **cluster=**[nonnegative number] specifies the maximum cM distance between individual markers that are combined into an aggregate marker when imputing ungenotyped markers (default: **cluster=0.005**).
- **ap=**[true/false] specifies whether AP1 and AP2 (allele probability) fields will be included in the output VCF file when imputing ungenotyped markers (default: **ap=false**). By default, allele probabilities are not printed because the the sum of the two allele probabilities (the allele dose) is always printed in the DS format field.
- **gp=**[true/false] specifies whether a GP (genotype probability) format field will be included in the output VCF file when imputing ungenotyped markers (default: **gp=false**). Genotype probabilities are calculated from allele probabilities assuming Hardy-Weinberg Equilibrium. Consequently, the alleles in the genotype with highest genotype probability may occasionally be different than the genotype obtained by taking the allele with highest probability on each haplotype, which is the genotype reported in the GT format field.

3.4 General parameters

- **ne=**[integer] specifies the effective population size (default: **ne=100000**). If the input genotypes are unphased, Beagle will automatically estimate the **ne** parameter prior to haplotype phasing unless **em=false**.
- **err=**[nonnegative number] specifies the allele mismatch probability for the hidden Markov model. If the input genotypes are unphased, Beagle will automatically estimate the **err** parameter prior to haplotype phasing unless **em=false**. If no **err** parameter is specified, the **err** parameter will be set equal $\theta/\big(2(\theta+H)\big)$ where H is the number of haplotypes and $\theta = 1/(0.5+\ln H)$.
- **em=**[true/false] specifies whether the initial **ne** and **err** parameters will be replaced with estimated values prior to haplotype phasing (default: **em=true**). If **em=true**, the initial parameter values will be updated using an expectation maximization (EM) algorithm. If **em=false**, the initial **ne** and **err** parameters will be used for phasing and imputation. The **em** has no effect if the input genotypes are phased.
- **window=**[positive number] specifies the maximum cM length of each sliding window (default: **window=40.0**). The **window** parameter must be at least 1.1 times as large as the **overlap** parameter. Reducing the **window** parameter can reduce the amount of memory required for the analysis.
- **window-markers=**[positive integer>=100000] specifies the maximum number of markers in each sliding window (default: **window-markers=4000000**). Reducing the window-markers parameter can reduce the amount of memory required for an analysis.
- **overlap=**[positive number] specifies the cM length of overlap between adjacent sliding windows (default: **overlap=2.0**).
- **seed=**[integer] specifies the seed for the random number generation (default: **seed=-99999**). Repeating an analysis with the same **seed** and **nthreads** parameters will produce the same phased and imputed genotypes.
- **nthreads=**[positive integer] specifies the number of computational threads that will be used. If the **nthreads** parameter is not specified, the **nthreads** parameter will be set equal to the number of CPU cores on the host machine. The **nthreads** parameter value is printed in the output **log** file.

4 Input files

Beagle uses Variant Call Format (VCF) 4.3 for input and output genotype data. Pseudoautosomal and non-pseudoautosomal X-chromosome genotypes must be in separate input files and analysed separately unless male haploid genotypes are coded as homozygous diploid genotypes.

Beagle assumes that an input VCF file that has a name ending in ".gz" is compressed with gzip or bgzip, and that a reference VCF file that has a name ending in ".bref3" is compressed with bref version 3.

5 Output files

There are two output files. The **log** file gives a summary of the analysis that includes the Beagle version, the command line arguments, and compute time.

The **vcf.gz** file is a bgzip-compressed VCF file that contains phased, non-missing genotypes for all non-reference samples. The output **vcf.gz** file can be uncompressed with the unix gunzip utility.

If a reference panel is specified and ungenotyped markers are imputed, the VCF INFO field will contain:

- A “DR2” subfield with the estimated squared correlation between the estimated allele dose and the true allele dose
- An “AF” subfield with the estimated alternate allele frequencies in the target samples
- The “IMP” flag if the marker is imputed