Beagle 4

Brian L. Browning

Department of Medicine

Division of Medical Genetics

University of Washington

Nov 22, 2013

Contents

1 Introduction

Beagle Version 4 performs genotype calling, haplotype estimation, imputation of ungenotyped markers, homozygosity-by-descent (HBD) detection and identity-by-descent (IBD) detection. Version 4 does not test for association with a binary trait, but version 3 can be used for this purpose.

Beagle requires a Java interpreter (version 1.6 or later). Type "java –version" at the command line prompt to check if a java interpreter is installed on your system. A Java interpreter can be downloaded from **www.java.com**.

Version 4 is under active development. Release r1185 eliminates some unnecessary command line arguments for the Refined IBD method, and it adds an **nthreads** parameter.

Beagle 4 is freely available and can be downloaded from:

http://faculty.washington.edu/browning/beagle/b4.html

1.1 Citing Beagle

If you use Beagle version 4 and publish your analysis, please cite the version of the program used and the following publication:

> 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

Beagle's Refined IBD algorithm uses the GERMLINE algorithm to detect candidate IBD segments and then evaluates these segments using a probabilistic IBD model. The GERMLINE algorithm is described in:

> A Gusev, J K Lowe, M Stoffel, M J Daly, D Altshuler, J L Breslow, J M Friedman, I Pe'er (2009) Whole population, genome-wide mapping of hidden relatedness. Genome Res 19(2):318-26. doi:10.1101/gr.081398.108

1.2 Variant Call Format

Beagle uses Variant Call Format (VCF) 4.1 for input and output file. VCF files can be manipulated and analysed with VCFtools, PLINK/SEQ, and the Beagle Utilities.

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

X chromosome: At present, version 4 requires haploid male X-chromosome genotypes to be coded as homozygous diploid genotypes. In the current version, the only parent-offspring relationships with a male offspring that can be included in a pedigree file are mother-offspring duos having a male offspring.

2 Command line arguments

To run Beagle version 4, enter the following command at the computer prompt:

java –Xmx[Mb]m –jar b4.jar [arguments]

where [Mb] is the number of megabytes of memory to use for the analysis (e.g. –Xmx2000m) 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 need several gigabytes of memory.

There are only two required command line arguments: a **gt, gl** or **gtgl** argument to specify the input file and an **out** argument to specify the output file prefix. All other command line arguments are optional.

Parent-offspring relationships can be specified and modelled by use of the **ped** parameter.

A reference panel can be specified with the **ref** parameter. Use of a population-matched reference panel can increase analysis accuracy. Corresponding variants 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 variants in your data to match the reference panel.

When a reference panel is used, it determines the variants included in the analysis. Variants absent from the reference panel are excluded. Variants in the reference panel that are absent in your data will be imputed. Use the **impute=false** argument if you do not wish to impute variants that are absent in your data. If you are imputing into a pre-phased target data which have no missing alleles then use the parameters **usephase=true burnin-its=0 phase-its=0**, and do not use the **ped=** parameter.

For IBD detection, use the **ibd=true** option. For best results, you may also 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.
- **gl=**[file] specifies a VCF file containing a GL or PL (genotype likelihood) format field for each marker. If both GL and PL format fields are present for a sample, the GL format will be used. See also the **maxlr** parameter.
- **gtgl=**[file] specifies a VCF file containing a GT, GL or PL (genotype likelihood) format field for each marker. The GT field is used if the GT field is present and the genotype is non-missing; otherwise, the GL or PL field is used. If both GL and PL format fields are present for a sample, the GL format will be used. See also the **maxlr** parameter.
- **ref=**[file] specifies a reference VCF file containing additional samples and phased genotypes for each marker. See also the **impute** parameter.
- **ped=**[file] specifies a Linkage-format pedigree file for specifying family relationships. The pedigree file has one line per individual. The first 4 white-space delimited fields of each line are 1) pedigree ID, 2) individual ID, 3) father's ID, and 4) mother's ID. A "0" is used in column 3 or 4 if the father or mother is unknown. The individual IDs are required to be unique. Beagle uses the data in columns 2-4 to identify parent-offspring duos and trios in the input data. Any or all columns of the pedigree file after column 4 may be omitted. See also the **duoscale** and **trioscale** parameters.
- **out=**[prefix] specifies the output filename prefix. The prefix may be an absolute or relative filename, but it cannot be a directory name.

- **impute=**[true/false] specifies whether variants 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** parameter is not used.
- **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 genomic coordinates in the format: CHROM:POS.
- **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. If enforcement of the maximum likelihood ratio would causes Mendelian inconsistency in a parent-offspring duo or trio, the maximum likelihood is not enforced for that marker in the duo or trio.

2.2 Other arguments

- **nthreads=**[positive integer] specifies the number of threads of execution to use during haplotype sampling (default: **nthreads=1**).
- **window=**[positive integer] specifies the number of markers to include in each sliding window (default: **window=24000**). 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.
- **overlap=**[positive integer] specifies the number of markers of overlap between sliding windows (default: **overlap=3000**). For human data, I suggest that the overlap be set to the typical number of markers in 0.5 cM (when **ibd=false**) or 1.5 cM (when **ibd=true**).
- **gprobs=**[true/false] specifies whether a GP (genotype probability) format field will be included in the output VCF file (default: **gprobs=true**).
- **usephase=**[true/false] specifies whether to use phase information in GT format fields for individuals in the input file specified with the **gt** or **gtgl** argument (default: **usephase=false**). If **usephase=false**, the allele order at heterozygous genotypes is randomized at the start of the analysis. Input phase is used only for individuals who are not part of a genotyped parent-offspring duo or trio (see the **ped** parameter).
- **seed=**[integer] specifies the random number generator seed (default: **seed=-99999**).
- **singlescale=**[positive number] specifies the model scale parameter when sampling haplotypes for unrelated individuals (default: **singlescale=1.0**). Increasing the **singlescale** parameter trades reduced single phasing accuracy for reduced run-time.

- **duoscale=**[positive number] specifies the model scale parameter when sampling haplotypes for parent-offspring duos (default: **duoscale=1.0**). Increasing the **duoscale** parameter trades reduced duo phasing accuracy for reduced run-time.
- **trioscale=**[positive number] specifies the model scale parameter when sampling haplotypes for parent-offspring trios (default: **trioscale=1.0**). Increasing the **trioscale** parameter trades reduced trio phasing accuracy for reduced run-time.

  Note regarding **singlescale, duoscale, and trioscale** parameters. The model scale parameters control the model complexity and are normally left at their default values to achieve highest accuracy. However, if the sample size is extremely large or if genotype likelihoods are used, it may be necessary to increase one or more scale parameters to obtain reasonable computation times. For example if the output log file shows that computation time is excessively long when sampling haplotypes for trios, the **trioscale** parameter can be increased to reduce this computation time. Increasing a scale factor from 1 to r will typically decrease computation time by a factor of approximately r^2 when sampling haplotypes. Computation time scales approximately linearly in the number of markers, and total computation time can be estimated from a pilot study of ~5000 markers.
- **burnin-its=**[non-negative integer] specifies the number of initial burn-in iterations (default: **burnin-its=5**).
- **phase-its=**[non-negative integer] specifies the number of iterations for estimating genotype phase (default: **phase-its=5**). Increasing this parameter (up to ~40 iterations) will typically increase genotype phase accuracy.
- **impute-its=**[non-negative integer] specifies the number of iterations for estimating genotypes at ungenotyped markers (default: **impute-its=5**). Increasing this parameter (up to ~10 iterations) will typically increase genotype imputation accuracy.

2.3 Identity by descent detection arguments

- **ibd=**[true/false] specifies whether IBD analysis will be performed (default: **ibd=false**).
- **ibdlod=**[non-negative integer] specifies the minimum LOD score for reported IBD (default: **ibdlod=3.0**).
- **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 cM). 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 to maximize IBD detection.

2.4 Advanced options not recommended for general use

- **dump=**[file] specifies a file containing sample identifiers (one identifier per line). For each marker window, all the sampled haplotypes for these individuals which are sampled after the burn-in iterations are printed to an output VCF files (dump.[window #].gz).
- **nsamples=**[positive integer] specifies the number of haplotype pairs to sample for each individual during each iteration of the algorithm (default: **nsamples=4**).
- **buildwindow=**[positive integer] specifies the number of markers used to build the haplotype frequency model at each locus (default: **buildwindow=500**).

3 Output files

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

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

The **warnings** file is created if there are any warnings generated during the analysis. For example, Mendelian inconsistent genotypes in parent-offspring duos and trios are reported in the warnings file.

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.

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)