User Manual for *BestHet*

Alexandre M. Harris and Michael DeGiorgio

September 20, 2016

Contents

1 Introduction

BestHet is a script written in R that computes the unbiased estimator of expected heterozygosity, $\widetilde{H}_{\mathrm{BLUE}}$, described in Harris and DeGiorgio [2016], at a given locus. In addition, *BestHet* provides two applications of $\widetilde{H}_{\mathrm{BLUE}}$, an estimator of F_{ST} that incorporates $\widetilde{H}_{\mathrm{BLUE}}$ [Hudson et al., 1992], and an estimator of the locus-specific branch length (LSBL) statistic [Shriver et al., 2004].

If in your use of this program you identify any bugs or experience any issues, please contact Alex Harris at `amh522@psu.edu` to report the issue.

If you use this software, please cite it as

> A M Harris and M DeGiorgio. An unbiased estimator of gene diversity with improved variance for samples containing related and inbred individuals of any ploidy. *Submitted*, 2016

2 Operation

This program is meant for use on a UNIX system. We distribute *BestHet* in compressed (.tgz) format. In this file, we include the script, manual, and a directory containing example data. To unpack *BestHet* from the command line, enter

```
tar -xzvf BestHet_program.tgz
cd ./BestHet_program
```

These commands will create the directory `BestHet` in the current directory, and switch the user to the `BestHet` directory. `BestHet` contains the script, manual, and `example_data` directory.

It is recommended that all input files be in the same directory as *BestHet*, and output files are by default written to the local directory if the user does not explicitly provide a path. *BestHet* can run from the command line as

```
Rscript BestHet.R
```

This command is followed by 5-11 user-specific command line arguments (see Section 4).

3 Input file format

The user must provide a kinship matrix and a locus-specific data matrix. The order of individuals in these matrices must be identical. These matrices must be prepared as files wherein each line contains space-separated values with no further formatting.

3.1 Kinship matrix

For a sample of n individuals, the kinship matrix is a symmetric, $n \times n$ matrix whose elements represent the kinship (Φ) between two sampled individuals. Each element in the matrix is the kinship of an individual from row j with an individual from column k (Φ_{jk} where $\Phi_{jk} \in [0, 1]$). The diagonal represents the kinship of each individual with itself (the values of diagonal elements must be between 0.5 and 1). An example 5×5 kinship matrix suitable for use by *BestHet* is presented below:

$$\begin{matrix} 0.5 & 0 & 0 & 0 & 0.25 \\ 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0 \\ 0.25 & 0 & 0 & 0 & 0.5 \end{matrix}$$

Here, five outbred diploid individuals were sampled, and individuals 1 and 5 are first-degree relatives. The style of this matrix is as in chapter 5 of Lange [2002].

3.2 Data matrix

For a sample of n individuals for which I distinct alleles were sampled at a locus, the data matrix is an $n \times I$ matrix whose elements must add up to the ploidy of the individual along a row, and to the count of the allele in the sample along a column. The elements in the data matrix are integer values. For example, consider a sample of $n = 5$ individuals, at a locus with $I = 3$ distinct alleles, and the following data matrix.

$$\begin{matrix} 1 & 0 & 1 \\ 0 & 2 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 2 \\ 0 & 1 & 1 \end{matrix}$$

Individual 1 is represented in the first row of the data matrix and is a heterozygous diploid for alleles 1 and 3 (which have a total count of 2 and 4 copies in the sample, respectively), the first and third entries in row 1 will be 1, and the second entry is 0. Similarly, individual 2 is represented in the second row and is homozygous for allele 2 (which has a total allele count of 4 in the sample), such that the first and third entries are 0 while the second entry is 2.

BestHet can receive data for individuals of any ploidy, and can also process data matrices containing empty rows. For individuals not genotyped at the submitted locus, the user may submit a data matrix containing rows with only 0 values. The processing steps incorporated into *BestHet* will adjust the kinship matrix accordingly to remove individuals not genotyped for a particular locus.

4 Output options

BestHet provides three output types, selected with the first command line argument, each of which can be adjusted following user specification. These are the computation of $\widetilde{H}_{\mathrm{BLUE}}$ for a population at a single locus, of $\widetilde{F}_{\mathrm{ST,BLUE}}$ (derived from $\widetilde{H}_{\mathrm{BLUE}}$) for two populations at a single locus, and of LSBL (derived from $\widetilde{F}_{\mathrm{ST,BLUE}}$) for three populations at a single locus. For each of these options, the user must define the output file by name, and can choose to include an identifying index for the locus. Additionally, not all kinship matrices are computationally invertible. In such situations, $\widetilde{H}_{\mathrm{BLUE}}$ cannot be computed, and the user must specify the manner in which *BestHet* should proceed.

4.1 Global options for output

4.1.1 Output file

By default, *BestHet* will output to a user-defined file within the local directory, and is set to append to this file if it has been previously created. This option eases implementation of *BestHet* within a user-programmed loop, and allows the results for all loci to be conveniently located in a single file. Regardless of option, the output of *BestHet* is a single line.

4.1.2 Reporting locus by identifier

We provide the option to include a locus identifier. The identifier may be entered into the command line as a number or non-whitespace character string. When this option is chosen, the output line will contain both the value of the statistic and the index value corresponding to the locus. If this option is not chosen, then only the statistic will be printed.

4.1.3 Options for kinship matrices that cannot be inverted

For cases in which *BestHet* is unable to compute $\widetilde{H}_{\text{BLUE}}$, we provide three options for data handling. First, the locus can be skipped (enter `skip` for command line argument `backup_option`; see following sections). In this case, the entry in the output file will be the character string `Skipped` rather than a numeric value corresponding to a statistic. Second, the value of the statistic at the locus can be reevaluated using the unbiased estimator of expected heterozygosity $\widetilde{H}$ [DeGiorgio et al., 2010] (enter `old` for command line argument `backup_option`; see following sections). This estimator has a larger variance than $\widetilde{H}_{\text{BLUE}}$, but is nonetheless a viable backup option because it is unbiased. The output for this option includes the indicator tag `old`. Finally, we provide the option to jitter the kinship matrix by randomly and symmetrically adding or subtracting values on the order of 10^{-10} to its nonzero, off-diagonal entries (enter `jitter` for command line argument `backup_option`; see following sections). This procedure is repeated 1000 times and the mean of these 1000 jittered replicates is taken as the value of $\widetilde{H}_{\text{BLUE}}$. In most cases, this option should allow for computation of $\widetilde{H}_{\text{BLUE}}$. The output for this option includes the indicator tag `jitter`. **We recommend users select `skip` for most datasets, as the number of sites incompatible with $\widetilde{H}_{\text{BLUE}}$ is expected to be small.**

4.2 Computing $\widetilde{H}_{\text{BLUE}}$

The computation of $\widetilde{H}_{\text{BLUE}}$ for a population at a single locus requires a kinship matrix and a data matrix. The command line argument to select this output is `H`. Six command line arguments are required, including the output type. The command line arguments are: `statistic` (select `H`), `kinship_matrix`, `data_matrix`, `output_filename`, `backup_option` (select `skip`, `old`, or `jitter`), `locus_index` (either a locus identifier or `no_loc`). The implementation of this option is

```
Rscript BestHet.R H kinship_matrix data_matrix output_filename backup_option
locus_index
```

4.3 Computing $\widetilde{F}_{\text{ST,BLUE}}$

The computation of $\widetilde{F}_{\text{ST,BLUE}}$ for two populations at a single locus requires a kinship matrix and a data matrix for each sampled population. Nine command line arguments are required. The command line arguments are: `statistic` (select `F`), `kinship_matrix_1`, `kinship_matrix_2`, `data_matrix_1`, `data_matrix_2`, `output_filename`, `backup_option` (select `skip`, `old`, or `jitter`; will be applied to both samples), `Fst_option` (`nd` or `calc`), `locus_index` (either a locus identifier or `no_loc`). We include the option to output the numerator and denominator of $\widetilde{F}_{\text{ST,BLUE}}$ separately (option `nd` of `Fst_option`) so that the user may compute the weighted mean across all loci, as in Reynolds et al. [1983]; the option to output a single value for locus $\widetilde{F}_{\text{ST,BLUE}}$ is selected with option `calc`. The implementation of this option is

```
Rscript BestHet.R F kinship_matrix_1 kinship_matrix_2 data_matrix_1 data_matrix_2
output_filename backup_option Fst_option locus_index
```

4.4 Computing LSBL

The computation of LSBL for three populations at a single locus requires a kinship matrix and a data matrix for all sampled populations. Eleven command line arguments are required. The command line arguments are: `statistic` (select `B`), `kinship_matrix_1`, `kinship_matrix_2`, `kinship_matrix_3`, `data_matrix_1`, `data_matrix_2`, `data_matrix_3`, `output_filename`, `backup_option` (select `skip`, `old`, or `jitter`; will be applied to all samples), `LSBL_option` (select `all` or `first`), `locus_index` (either a locus identifier or `no_loc`). LSBL is calculated for each population in order of input if `all` is chosen for `LSBL_option`. If `first` is chosen instead, then only the first LSBL statistic is computed, for the first population. The implementation of this option is

```
Rscript BestHet.R B kinship_matrix_1 kinship_matrix_2 kinship_matrix_3 data_matrix_1
data_matrix_2 data_matrix_3 output_filename backup_option LSBL_option locus_index
```

5 Examples

The example data used here derive from the MS5795 composite human microsatellite dataset of Pemberton et al. [2013]. We use the D5S817 locus (index 117) in our examples containing invertible kinship matrices for all populations, and the D2S1391 locus for examples containing a non-invertible kinship matrix (index 116, requiring implementation of `backup_option`). These examples are found in the `example_data` directory of `BestHet_program.tgz`.

5.1 Computing $\widetilde{H}_{\mathbf{BLUE}}$

To calculate $\widetilde{H}_{\mathrm{BLUE}}$ for the Waunana population of MS5795 [Pemberton et al., 2013], where the D5S817 locus is indexed as 117 and displayed in the output line, enter

```
Rscript BestHet.R H 848_Waunana_kinship.txt 848_Waunana_loc117_BHTest_data.txt
Waunana_H_output.txt skip 117
```

The *BestHet* output for this computation is

```
0.8219372 117
```

Note that although `skip` was chosen as the `backup_option`, this command line argument has no impact on the output if the kinship matrix is invertible.

5.2 Computing $\widetilde{F}_{\mathbf{ST,BLUE}}$

To calculate $\widetilde{F}_{\mathrm{ST,BLUE}}$ for the Waunana and Karitiana populations of MS5795 [Pemberton et al., 2013], where the numerator and denominator of $\widetilde{F}_{\mathrm{ST,BLUE}}$ are individually displayed in the output line in that order, and the D2S1391 locus is indexed as 116 and displayed in the output line, enter

```
Rscript BestHet.R F 848_Waunana_kinship.txt 82_Karitiana_kinship.txt
848_Waunana_loc116_BHTest_data.txt 82_Karitiana_loc116_BHTest_data.txt
Waunana_Karitiana_Fst_output.txt jitter nd 116
```

The *BestHet* output for this computation is

```
0.04784798 0.7621704 116 jitter
```

Because the Waunana sample kinship matrix is not invertible at this locus, the `backup_option` of `jitter` adds random noise symmetrically to the nonzero, off-diagonal elements of `848_Waunana_kinship.txt`, allowing it to be inverted and incorporated into $\widetilde{H}_{\mathrm{BLUE}}$ and therefore $\widetilde{F}_{\mathrm{ST,BLUE}}$. Note that running this computation again would yield slightly different output values because no two jittered computations are identical.

5.3 Computing LSBL

To calculate LSBL for the Waunana, Karitiana, and Orcadian populations of MS5795 [Pemberton et al., 2013], where all three possible LSBL statistics are individually displayed in the output line, and the D2S1391 locus chosen but not displayed in the output line, enter

```
Rscript BestHet.R B 848_Waunana_kinship.txt 82_Karitiana_kinship.txt
20_Orcadian_kinship.txt 848_Waunana_loc116_BHTest_data.txt
82_Karitiana_loc116_BHTest_data.txt 20_Orcadian_loc116_BHTest_data.txt
Waunana_Karitiana_Orcadian_LSBL_output.txt old all no_loc
```

The *BestHet* output for this computation is

```
0.05325769 0.002410236 -0.002410236 old
```

Because the Waunana sample kinship matrix is not invertible at this locus, the `backup_option` of `old` directs *BestHet* to compute $\widetilde{H}$ [DeGiorgio et al., 2010] rather than $\widetilde{H}_{\text{BLUE}}$ for all populations, and incorporate this into the calculation of LSBL. Although $\widetilde{H}$ has higher variance than $\widetilde{H}_{\text{BLUE}}$, it is always possible to compute this value.

References

M DeGiorgio, I Jankovic, and Rosenberg N A. Unbiased estimation of gene diversity in samples containing related individuals: exact variance and arbitrary ploidy. *Genetics*, 186:1367–1387, 2010.

A M Harris and M DeGiorgio. An unbiased estimator of gene diversity with improved variance for samples containing related and inbred individuals of any ploidy. *In review*, pages XX–YY, 2016.

R R Hudson, M Slatkin, and W P Maddison. Estimation of Levels of Gene Flow From DNA Sequence Data. *Genetics*, 132:583–589, 1992.

K Lange. *Mathematical and Statistical Methods for Genetic Analysis*. Springer, New York, 2nd edition, 2002.

T J Pemberton, M DeGiorgio, and N A Rosenberg. Population structure in a comprehensive data set on human microsatellite variation. *G3 (Bethesda)*, 3:909–916, 2013.

J Reynolds, B S Weir, and C C Cockerham. Estimation of the coancestry coefficient: basis for a short-term genetic distance. *Genetics*, 105:767–779, 1983.

M D Shriver, G C Kennedy, E J Parra, H A Lawson, V Sonpar, J Huang, J M Akey, and K W Jones. The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. *Hum. Genomics*, 1:274–286, 2004.