# User Manual for *BALLET* v1.0

Michael DeGiorgio, Kirk E. Lohmueller, Rasmus Nielsen

August 27, 2014

# 1. Introduction

*BALLET* is a program to perform genome-wide scans of ancient balancing selection using the spatial distribution of polymorphisms and substitutions in the genome. The current version of the software implements the $T_1$ and $T_2$ test statistics of DeGiorgio *et al*. (2014).

This software is still in its early stages. If you identify any bugs or issues with the software, then please contact Michael DeGiorgio at `mxd60@psu.edu` to report the issue.

If you use this software, then please cite it as

> M DeGiorgio, KE Lohmueller, R Nielsen (2014) A model-based approach for identifying signatures of ancient balancing selection in genetic data. *PLoS Genet* 10:e1004561.

# 2. Installation

*BALLET* should run on any UNIX system. It requires that the GNU Scientific Library is installed (http://www.gnu.org/s/gsl/). The GNU Scientific Libraries must be installed prior to running the following compilation steps. On the command line enter:

```
>tar -xzvf ballet_v1.0.tar.gz
>cd ballet_v1.0
>gcc ballet_v1_0.c –o BALLET –lm –lgsl -lgslcblas
```

# 3. Input file format

Note: If have used a Windows operating system to generate the following input files, then you will likely need to run the UNIX command `dos2unix` on each of the files before using them as input in *BALLET*.

## 3.1. Polymorphism and substitution file

The polymorphism and substitution input file is tab-delimited and contains a header. Each row represents the polymorphism or substitution status for a position in the genome, and the rows are ordered by increasing position along the chromosome. There should be a separate polymorphism and substitution file for each chromosome when performing a scan for balancing selection. At each row, the first column is the position on the chromosome, the second column is ancestral allele count ($x$), and the third column is the sample size ($n$). Note that alleles must be polarized using an outgroup, and the ancestral allele count can take on values $x = 0, 1, \ldots, n-1$. If $x = 0$, then the site is a substitution. Otherwise, the site is a polymorphism. Note that the methods of DeGiorgio *et al*. (2014) do not consider sites with $x = n$, and so these sites should not be included in the input file. An example input file is (say for chromosome 6):

| position | x | n |
|---|---|---|
| 460000 | 9 | 100 |
| 460010 | 0 | 100 |
| 460210 | 30 | 78 |
| 463000 | 0 | 94 |
| … | … | … |

The first line of the example input file displays the header, which must be identical to this example. The next four rows display polymorphism and substitution data for four positions on a chromosome (positions 460000, 460010, 460210, and 463000). Each row indicates the number of ancestral alleles observed (and the total number of alleles observed) at the given chromosomal position. At position 460000, 9 ancestral alleles were

observed out of 100 total alleles (50 diploid individuals) leading to an observed polymorphism in the sample. At position 460010, no ancestral alleles were observed out of 100 total alleles (50 diploid individuals) leading to an observed substitution in the sample. At position 460210, 30 ancestral alleles were observed out of 78 total alleles (39 diploid individuals), leading to another observed polymorphism in the sample. At position 463000, no ancestral alleles were observed out of 94 total alleles (47 diploid individuals), leading to another observed substitution in the sample. If the true sample size was 50 diploid individuals, then positions 460210 and 463000 would be genomic positions with missing data in the sample.

### 3.2. Recombination file

The recombination rate input file is tab-delimited and contains a header. Each row represents the population-scaled recombination rate between the position in the genome and the previous position in the file, and the rows are ordered by increasing position along the chromosome. Every position, and only those positions, in the polymorphism and substitution input file should be included in the respective recombination input file. There should be a separate recombination rate file for each chromosome when performing a scan for balancing selection. At each row, the first column is the position on the chromosome and the second column is the population-scaled recombination rate ($\rho$). For the first position in the input file $\rho = 0$. An example recombination input file matching the above example polymorphism and substitution input file:

```
position    rate
460000      0.0
460010      0.01
460210      0.2
463000      2.79
…           …
```

The first line of the example input file displays the header, which must be identical to this example for every recombination input file. The next four rows display population-scaled recombination rates for four positions on a chromosome (positions 460000, 460010, 460210, and 463000). At position 460000, the rate is 0, because it is the first position in the file. The rate between positions 460000 and 460010 is 0.01, between positions 460010 and 460210 is 0.2, and between positions 460210 and 463000 is 2.79.

## 4. Helper files (required before using –T1 and –T2 options)

It is necessary to generate these files prior to using the `-T1` and `-T2` options to scan for balancing selection. For all helper files, it is required that the user first combined their polymorphism and substitution files into a single polymorphism and substitution file with exactly the same format as the example in section 3.1, and we will refer to this file as `CombinedSNPFile`. The reason to create this combined polymorphism and substitution file is to generate genome-wide estimates of the inter-species coalescent time, the proportion of polymorphisms and substitutions, and the empirical frequency spectrum. As an example, suppose we have data from each of the 22 human autosomes, and each chromosome $k$ has its own polymorphism and substitution file called `SNPFile_k`, $k = 1, 2, \ldots, 22$. The `CombinedSNPFile` would have all of the data contained in `SNPFile_k`, $k = 1,2, \ldots, 22$, in one file. There would be a line in `CombinedSNPFile` for each of the data lines contained in `SNPFile_k`, $k = 1,2, \ldots, 22$.

### 4.1. Estimate of the inter-species coalescent time (`-inter_coal_time`)

**Note:** This is a simple estimate of the inter-species coalescent time, and users are advised to calculate the inter-species coalescent time using more sophisticated methods.

To perform a simple estimate of the inter-species coalescent time between then ingroup and outgroup, use the `-inter_coal_time` option. The command is:

```
./BALLET -inter_coal_time CombinedSNPFile K 4Nu DivFile
```

where `CombinedSNPFile` is a polymorphism and substitution file combined across all chromosomes in the analysis (to get a genome-wide estimate), $K$ is the total sequence length that the polymorphism and substitution input file is based on, $4Nu$ is a user-defined estimate of the population-scaled mutation rate, and `DivFile` is the name of a file where the results will be printed. Note that the value of $K$ is likely much larger than the total number of polymorphisms and substitutions in the dataset, as it is the total sequence length rather than the number of sites in the dataset.

### 4.2. Compute proportion of polymorphisms and substitutions (`-poly_sub`)

To compute the proportion of polymorphisms and substitutions, use the `-poly_sub` option. The command is:

```
./BALLET -poly_sub CombinedSNPFile PolySubFile
```

where `CombinedSNPFile` is a polymorphism and substitution file combined across all chromosomes in the analysis (to get a genome-wide estimate) and `PolySubFile` is the name of a file where the results will be printed.

### 4.3. Compute empirical frequency spectrum (`-spect`)

To compute the empirical frequency spectrum, use the `-spect` option. The command is:

```
./BALLET -spect CombinedSNPFile SpectFile
```

where `CombinedSNPFile` is a polymorphism and substitution file combined across all chromosomes in the analysis (to get a genome-wide estimate) and `SpectFile` is the name of a file where the results will be printed.

## 5. Scanning for ancient balancing selection

### 5.1 Scan using $T_1$ test statistic (`-T1`)

To perform a scan for ancient balancing selection with the $T_1$ statistic of DeGiorgio *et al*. (2014), use the `-T1` option. The command to perform this scan is

```
./BALLET -T1 W DivFile PolySubFile SNPFile RecFile OutFile
```

where `W` is a user-defined window size ($W$ polymorphisms or substitutions directly upstream of a test site and $W$ polymorphisms or substitutions directly downstream of a test site) to compute the test statistic, `DivFile` is an input file containing an estimated inter-species coalescence time between the ingroup and outgroup, `PolySubFile` is an input file containing the ratio of polymorphisms to substitutions calculated using the `-poly_sub` option in section 4.2, `SNPFile` is the polymorphism and substitution input file, `RecFile` is the respective population-scaled recombination rate file, and `OutFile` is the name of a file where the results will be printed. Here, the `SNPFile` and `RecFile` would be for a specific chromosome (or region of the genome) rather than combined across all chromosomes.

## 5.2. Scan using $T_2$ test statistic (-T2)

To perform a scan for ancient balancing selection with the $T_2$ statistic of DeGiorgio *et al*. (2014), use the `-T2` option. The command to perform this scan is

```
./BALLET -T2 W DivFile PolySubFile SpectFile SNPFile RecFile PATH OutFile
```

where `W` is a user-defined window size ($W$ polymorphisms or substitutions directly upstream of a test site and $W$ polymorphisms or substitutions directly downstream of a test site), `DivFile` is an input file containing an estimated inter-species coalescence time between the ingroup and outgroup, `PolySubFile` is an input file containing the ratio of polymorphisms to substitutions calculated using the `-poly_sub` option in section 4.2, `SpectFile` is an input file containing the ancestral frequency spectrum calculated using the `-spect` option in section 4.3, `SNPFile` is the polymorphism and substitution input file , `RecFile` is the respective population-scaled recombination rate file, `PATH` is the path to the directory containing the simulated frequency spectra under a model of balancing selection (included in `simulated_spectra.tar.gz`), and `OutFile` is the name of a file where the results will be printed. Here, the `SNPFile` and `RecFile` would be for a specific chromosome (or region of the genome) rather than combined across all chromosomes.

To extract the simulated spectra, type the command

```
tar -xzvf simulated_spectra.tar.gz
```

The directory structure for the decompressed folder has the form

```
simulated_spectra/nX/
```

for $X = 2, 3, \ldots, 50$, where `nX` is the directory containing simulated spectra for a sample of size $X$. When using the `-T2` option, the `PATH` variable should be set to

```
PATH = /[paths]/simulated_spectra/
```

For example, suppose the simulated_spectra.tar.gz file was extracted to directory `/home/user/project/`. Then the appropriate command would look like

```
./BALLET -T2 W DivFile PolySubFile SpectFile SNPFile RecFile
/home/user/project/simulated_spectra/ OutFile
```

where in this example we set the PATH variable to

```
PATH = /home/user/project/simulated_spectra/
```

The program will now know to look for simulated frequency spectra in /home/user/project/simulated_spectra/.

**Note:** Over time I will update the `simulated_spectra.tar.gz` file with larger sample sizes. Currently, the maximum is 100 alleles (50 diploid individuals).

# 6. References

M DeGiorgio, KE Lohmueller, R Nielsen (2014) A model-based approach for identifying signatures of ancient balancing selection in genetic data. *PLoS Genet* 10:e1004561.