User manual for *BalLeRMix*

Xiaoheng Cheng and Michael DeGiorgio

April 23, 2019

Contents

1 Introduction

BalLeRMix is a Python script to perform scans with the mixture model-based likelihood ratio B statistics (Cheng and DeGiorgio, 2019). Operation of this package requires a UNIX environment with Python 2.7 and above.

Please cite it as:

X. Cheng and M. DeGiorgio. Robust and window-insensitive mixture model approaches for localizing balancing selection. *Submitted.*

If you experience any issues, please contact Xiaoheng Cheng at xh.cheng@psu.edu for further help.

2 Operation

2.1 Installation

We distribute *BalLeRMix* in compressed (tar.gz) format. The script included is designed to perform on a UNIX system. To unpack the package from the command line, go to the directory where it is stored, and enter

```
tar -xzvf BalLeRMix_v1.tar.gz
cd BalLeRMix_v1/
```

The first command will decompress the file and release the contents into folder `BalLeRMix_v1/` in the current directory. The second command will lead the user to the `BalLeRMix_v1/` directory, which contains the manual, `test/` directory, and the Python script `BalLeRMix.py`.

2.2 Quick Guide

To run *BalLeRMix*, the basic command (in one line) is

```
python BalLeRMix.py -i <input file> -o <output file> --spect <spect/config file>
   [--help] [Options to choose variant] [Options to generate helper files]
   [Options to customize scanning window] [Options to define parameter space]
```

In this command line, the paths and names of the input and output files follow the dash commands `-i` or `--input`, and `-o` or `--output`, respectively. The path and name of allele frequency spectrum file or polymorphism-substitution configuration file follow the command `--spect`. Note that these three files are the minimum requirements for the scan to run. Other commands can be used to specify the B variant of choice, generate helper files, customize the scanning window, or customize the parameter space for optimization.

The command `-h` or `--help` can be used to display details of each command. This help page, as shown below, would also appear if no arguments are provided.

```
Usage:  python BalLeRMix.py -i <input file> -o <output file> --spect <spect/config
file> [--help] [--nofreq] [--nosub] [--MAF] [--getSpect] [--getConfig] [--fixSize]
[--physPos] [--rec <recomb rate>] [-w <window size>] [--noCenter] [-s <step size>]
[--rangeA <min,max,step>] [--listA <A1,A2,..,Ak>] [--fixX <x>]
```

```
Options:
-h, --help            show this help message and exit
-i INFILE, --input=INFILE
                      Path and name of your input file.
-o OUTFILE, --output=OUTFILE
                      Path and name of your output file.

--spect=SPECTFILE     Path and name of the allele frequency spectrum file
                      or configuration file.

--getSpect            Option to generate frequency spectrum file from the
                      concatenated input file.  Use "-i" and "--spect"
                      commands to provide names and paths to input and
                      output files, respectively.  Indicate the input type
                      with "--MAF" and/or "--nosub".

--getConfig           Option to generate configuration file from the
                      concatenated input file.  Use "-i" and "--spect"
                      commands to provide names and paths to input and
                      output files, respectively.

--nofreq              Option to ignore allele frequency information.  All
                      polymorphic sites will be considered as equivalent.

--nosub               Option to not include substitution in input data.
--MAF                 Option to use minor allele frequency, instead of
                      polarized allele frequency.  The latter is default.

--physPos             Option to use physical positions instead of genetic
                      positions (in cM). Default is using genetic positions.

--rec=RRATE           The uniform recombination rate in cM/site.  Default
                      value is 1e-6 cM/site.  Only useful when choose to use
                      physical positions as coordinates.

--fixSize             Option to fix the size of scanning windows.  When true,
                      provide the length of window in nt with "-w" command.

-w R, --window=R      Number of sites flanking the test locus on either
                      side.  When choose to fix window size (--fixSize),
                      input the length of window in nt.

--noCenter            Option to have the scanning windows not centered on
                      informative sites.  Require that the window size ("-w")
                      in physical positions ("--physPos") is provided.
                      Default is True.

-s STEP, --step=STEP  Step size in nt (when using "--noCenter") or the
                      number of informative sites.  Default value is one
                      site or one nucleotide.

--fixX=X              Option to fix the presumed equilibrium frequency.
--rangeA=SEQA         Range of the values of the parameter A to optimize
                      over.  Format should follow <Amin>,<Amax>,<Astep> with
                      no space around commas.

--listA=LISTA         Manually provide a list of A values to optimize over.
                      Please separate the values with comma, no space.
```

2.3 Specify the B variant

With the necessary files provided, to run the scan, the user should specify the particular B statistic variant to use with commands `--nofreq`, `--nosub`, and `--MAF`. Specifically, when none of these commands are used, the script will perform the scan with the B_2 statistic, which requires both derived allele frequencies and substitutions in the input. Based on the type of input data the other variants consider, the user can use `--MAF` to indicate that the data contains the minor allele frequency, use `--nosub` to indicate the input does not contain substitutions, or use `--nofreq` to indicate that the data only include the polymorphic states for each site (1 for polymorphism, 0 for substitution), and does not present frequency information. Note that `--nofreq` and `--nosub` cannot be used together. The table in section 2.4 summarizes the commands to use for each variant.

2.4 Generate helper files for each variant

Prior to performing the scan with a particular variant, the user can use `--getSpect` or `--getConfig` command to generate the corresponding frequency spectrum or configuration file. To this end, the user should provide the concatenated input for the whole-genome. The input and output helper files should be provided following `-i` and `--spect` commands, respectively. The type of B variant can be specified by combinations of `--nofreq`, `--nosub`, and `--MAF` commands, in the same way as specifying the variant for scanning. The table below summarizes the command combinations to use for each variant. Section 5 also provides example command lines for generating these files.

Variant	To run scan	To generate spect/config file
B_2:	-	`--getSpect`
$B_{2,\mathrm{MAF}}$:	`--MAF`	`--getSpect --MAF`
B_1:	`--nofreq`	`--getConfig`
B_0:	`--nosub`	`--getSpect --nosub`
$B_{0,\mathrm{MAF}}$:	`--nosub --MAF`	`--getSpect --nosub --MAF`

2.5 Specify the coordinates and scanning window

The arguments `--fixSize`, `--physPos`, `-w <window size>`, `--noCenter`, and `-s <step size>` can be used to customize position coordinates and the windows on which each test is computed. When none of these commands are used, *BalLeRMix* by default uses all the data based on their genetic positions in centi-Morgans (cM) for each test site.

The B statistics model the probability distributions under the null hypotheses as functions of the observed allele count of each informative site, as well as its distance from the test position. When implementing B statistics, *BalLeRMix* considers the distance between the two sites as their recombination distance in cM. By default, *BalLeRMix* takes in the genetic positions (in cM) of each informative site (see section 3 for input format) for computation. When the recombination map for the organism is unavailable, the user can use `--physPos` to instruct *BalLeRMix* to instead consider the physical positions (in the number of nucleotides [nt]). The user is then advised to provide a uniform recombination rate, in cM/nucleotide, following the command `--rec` to overwrite the default rate of 10^{-6} cM/nucleotide.

To fix the length of scanning windows (in nt), use `--fixSize`. When this command is present, *BalLeRMix* will automatically switch the coordinates to physical positions. Meanwhile, length of the window size (in nt), should be provided via `-w` or `--window`. By default, each test is computed at each informative site. To allow the fixed-length window to slide through the sequence with fixed-length steps, use `--noCenter` to overwrite the default, and provide the step size (in nt) following `-s` or `--step`. Without the `--noCenter` command, the number following `-s` would be the number of informative sites slid through at each step. The default value for step size is one, representing a round of computation every one nt or every one informative site, respectively, with or without `--noCenter`.

Alternatively, to fix the number of sites considered at either side of the test site, provide this number via `-w` only, without using `--fixSize`. By default, *BalLeRMix* will compute tests at each informative site. To skip sites, use `-s`, with the value being the number of sites skipped, including the previous test site. For example, `-s 2` indicates to compute every other informative site, and `-s 5` means to compute every five informative site. The following table summarizes the arguments for each type of sliding window.

Window span	Test position	Required commands	Optional commands
Fixed length in nt	Informative sites	`--fixSize -w <nt> -s <sites> [--physPos]`	`[--rec <cM/nucleotide>]`
	evenly-spaced locations	`--fixSize -w <nt> --noCenter -s <nt> [--physPos]`	`[--rec <cM/nucleotide>]`
Fixed number of informative sites	Informative sites	`-w <sites in radius> -s <sites per steps>`	`[--physPos] [--rec <cM/nucleotide>]`
All input data	Informative sites	`-`	`[-s <sites/step>] [--physPos] [--rec <cM/nucleotide>]`

2.6 Customize the parameter space

The likelihood for the alternative hypothesis in B statistics is maximized over the parameter A (which characterizes the width of the genomic footprint left by balancing selection) and equilibrium frequency x, before being considered for the likelihood ratio. *BalLeRMix* implements the optimization process across a grid of discrete x and A values, with $x \in \{0.05, 0.10, \ldots, 0.50\}$ and the values considered for A being 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1100, 1200, 1400, 1600, 1800, 2000, 2200, 2400, 2500, 3000, 3500, 4000, 5000, 6000, 7000, 8000, 9000, 10^4, 10^6, and 10^8.

To customize the set of parameter values for *BalLeRMix* to optimize over, use `--fixX <x>` for equilibrium minor allele frequency x, and either `--rangeA <min,max,step>` or `--listA <A1,A2,..,Ak>` parameter A. The `--fixX` command allows the user to assign a certain equilibrium minor allele frequency-of-interest for the scan. For example, `--fixX 0.35` would instruct *BalLeRMix* to only consider the likelihood for balancing selection with the equilibrium frequency of $x = 0.35$. Note that the assigned x value should not exceed 0.5, as it represents the equilibrium minor allele frequency. With `--rangeA`, the user can specify an evenly-spaced range of values as the possible values of A for *BalLeRMix* to consider, with the first two numbers setting the minimum and maximum values, and the third specifying the space between the neighboring values. For example, for A to be considered among 1500, 2000, 2500, or 3000, the command would be `--rangeA 1500,3000,500`. Alternatively, the user can also list out all the values, delimited by comma, following the `--listA` command: *e.g.*, `--listA 1500,2000,2500,3000`.

3 Input format

BalLeRMix requires allele count data with both physical and genetic positions. To present such data, input files must be tab-delimited plain text files, with four columns: physical position, genetic position, allele count, and sample size. The physical positions should be integers in bases, and the genetic positions should be in centi-Morgans (cM). When the user does not have recombination maps for reference, the seceond column can be `NA`s as long as the user makes sure to use `--physPos`. When using physical positions as coordinates,

BalLeRMix assumes a uniform recombination rate of 10^{-6} cM per nucleotide, close to the estimated value in humans (Payseur and Nachman, 2000). If the species of interest does not have this recombination rate, then the user must specify this rate with `--rec`, and ensure its unit is cM per nucleotide.

All file names are recommended to be presented in absolute paths. If their absolute paths are unavailable, then users can move `BalLeRMix.py` to the same working directory, and follow the instructions in sections 2 or 5.

If the input file has been previously edited in an operating system environment other than UNIX, then we advise users to use the following command to ensure this file is readable in a UNIX environment:

```
dos2unix <file>
```

3.1 Input file format

In the `test/` folder, the user should find example input files for B_2, $B_{2,\mathrm{MAF}}$, B_0, and $B_{0,\mathrm{MAF}}$. They follow the same layout, with the observed allele count `x` being different. For B_2 and B_0, `x` should be the derived allele frequency; for $B_{2,\mathrm{MAF}}$ and $B_{0,\mathrm{MAF}}$, it should be the minor allele frequency. The two example files below are examples for derived allele count (left) and minor allele count (right), respectively. Note that although the sites of substitution are represented by `x=0` when using minor allele frequencies, the user should not include monomorphic sites that are the same allele as the outgroup sequence.

physPos	genPos	x	n
79	0.000079	50	50
178	0.000178	50	50
256	0.000256	50	50
267	0.000267	28	50
361	0.000361	19	50
432	0.000432	2	50
...			

physPos	genPos	x	n
79	0.000079	0	50
178	0.000178	0	50
256	0.000256	0	50
267	0.000267	22	50
361	0.000361	19	50
432	0.000432	2	50
...			

Nonetheless, because *BalLeRMix* automatically folds the allele count when opted to use minor allele frequencies, the left file can also be the input for $B_{2,\mathrm{MAF}}$ and $B_{0,\mathrm{MAF}}$. Similarly, inputs for B_2 variants can be used by B_0 too, as *BalLeRMix* will not consider any substitutions when reading the data. The user should, however, provide different frequency spectrum files to B_2 and B_0 variants.

3.2 Frequency spectrum format

For B_2 and B_0 variants, the user should provide the allele frequency spectrum file. This file should be tab-delimited, have no header, and consist of three columns: allele count, sample size, and proportion in the genome (normalized such that the sum of values in the third column sum to one across rows with the same sample size). Following section 5, the user should be able to generate a full frequency spectrum file, whose beginning is as below. The user should generate the frequency spectrum for each variant of the B statistics, separately.

```
1   50   0.0514846766351
2   50   0.0291495363996
3   50   0.0194707349246
4   50   0.0150561808571
    ...
```

3.3 Configuration format

The configuration file records the proportions of substitutions and polymorphisms, and is required for computing the B_1 statistic. Similar to the frequency spectrum file, the configuration file is tab-delimited and has no headers. The three columns correspond to the sample size, the proportion of substitution, and the proportion of polymorphism, with the proportions of substitutions and polymorphisms summing to one. An example file is shown below.

```
50   0.73883844754   0.26116155246
```

4 Output format

BalLeRMix writes output as a tab-delimited table with five columns, respectively for physical positions, genetic positions, maximum log likelihood ratios, and the values of x and A that yielded this likelihood ratio. The only exception is when choosing fixed-size sliding windows that do not center on the informative sites, where the header would be

```
midPos   genPos   LR   xhat   Ahat
```

where the first column is instead the physical center position of the respective window. All other output files will have the below header.

```
physPos   genPos   LR   xhat   Ahat
```

5 Examples

To further illustrate the usage of *BalLeRMix*, we include example input files and the concatenated input file (for deriving helper files) in the subfolder `test/`. The input files for each variant of the B statistics begin with

```
==> test/B0_input.txt <==
physPos   genPos     x    n
267       0.000267   28   50
361       0.000361   19   50
432       0.000432   2    50
...
==> test/B0maf_input.txt <==
physPos   genPos     x    n
267       0.000267   22   50
361       0.000361   19   50
432       0.000432   2    50
...
==> test/B2_input.txt <==
physPos   genPos     x    n
8         0.000008   50   50
79        0.000079   50   50
178       0.000178   50   50
...
==> test/B2maf_input.txt <==
physPos   genPos     x   n
8         0.000008   0   50
79        0.000079   0   50
178       0.000178   0   50
...
```

Additionally, in the `test/` folder there is a file named `Concatenated_input.txt`, which is concatenated from the parsed inputs of 500 replicates of a neutrally-evolving 50 kilobase (kb) sequence. In practice, the user should concatenate inputs from all chromosomes or contigs, and use them to infer whole-genome level of variation, which we consider as neutral.

5.1 Generating helper files

BalLeRMix provides built-in functions to generate helper files (see section 2.4). Here, the concatenated input file is sufficient to generate helper files for all variants of B statistics. After navigating into the `BalLeRMix_v1/` folder, the following commands can generate the spectrum files for B_2, $B_{2,\mathrm{MAF}}$, B_0, and $B_{0,\mathrm{MAF}}$, respectively.

```
# for B2
python BalLeRMix.py -i test/Concatenated_input.txt --getSpect \
   --spect Spect_DAF.txt

# for B2,MAF
python BalLeRMix.py -i test/Concatenated_input.txt --getSpect \
   --spect Spect_MAF.txt --MAF

# for B0
python BalLeRMix.py -i test/Concatenated_input.txt --getSpect \
   --spect Spect_DAF-nosub.txt --nosub

# for B0,MAF
python BalLeRMix.py -i test/Concatenated_input.txt --getSpect \
   --spect Spect_MAF-nosub.txt --MAF --nosub

# for B1
python BalLeRMix.py -i test/Concatenated_input.txt --getConfig \
   --spect Config.txt
```

5.2 Performing scans

With the helper files ready, we can move on to perform scans with different B variants and different sliding windows.

To run $B_{0,\mathrm{MAF}}$ with genetic positions as coordinates, considering all the polymorphic sites in the input on every polymorphic site:

```
python BalLeRMix.py -i test/B0maf_input.txt \
   --nosub --MAF --spect Spect_MAF-nosub.txt
```

To run B_0 with physical positions as coordinates, using as much data as possible, with the recombination rate being 10^{-5} cM/nt:

```
python BalLeRMix.py -i test/B0_input.txt \
   --nosub --spect Spect_DAF-nosub.txt \
   --physPos --rec 1e-5
```

To run B_1 with physical positions as coordinates, employing 10 kb windows and a 500-base step size, with the default recombination rate of 10^{-6} cM/nt (note that either input for B_2 or B_0 will be compatible with B_1 computation):

```
python BalLeRMix.py -i test/B2_input.txt \
   --nofreq --spect Config.txt \
   --noCenter --fixSize -w 1e4 -s 500
```

To run B_2 with genetic positions, using 100 informative sites at either side of the test site (*i.e.*, 200 sites in total), and computing a test at each informative site:

```
python BalLeRMix.py -i test/Concatenated_input.txt \
   --spect Spect_MAF.txt -w 100
```

To run $B_{2,\mathrm{MAF}}$ with 100 kb windows centered on every three informative sites (*i.e.*, skip two sites at every step):

```
python BalLeRMix.py -i test/Concatenated_input.txt \
   --MAF --spect Spect_MAF.txt \
   --fixSize -w 1e5 -s 3
```

Note that in the case above, *BalLeRMix* will automatically consider physical positions as coordinates, and adopt the default recombination rate of 10^{-6} cM/nt. Use `--rec` to assign other values if needed.

References

X. Cheng and M. DeGiorgio. Robust and window-insensitive mixture model approaches for localizing balancing selection. *Submitted*, 2019.

B. A. Payseur and M. W. Nachman. Microsatellite variation and recombination rate in the human genome. *Genetics*, 156(3):1285–1298, 2000.