A Short Manual for Bayenv2.0

Torsten Günther & Graham Coop

August 2, 2018

Contents

1 Introduction

The program is an implementation of the MCMC algorithms described in:

Using Environmental Correlations to Identify Loci Underlying Local Adaptation. Coop G., Witonsky D., Di Rienzo A., Pritchard J.K. Genetics. 2010

and

Robust identification of local adaptation from allele frequencies. Günther T., Coop G. Genetics. 2013

The code is designed to be run on large SNP data sets. While the method attempts to control for the effects of population structure, it does not claim to fully control for population structure. Thus the statistics produced should NOT be treated at face value as measures of significance. Due to the fact that the residual effects of population structure will differ across environmental variables the statistics are not comparable across environmental variables. Rather the Bayes factors/Z/ρ provide a useful statistic for ranking SNPs by their allele frequency correlation with an environmental variable.

As with all MCMCs the estimates from the algorithm are subject to stochastic error that decreases with the number of iterations that the MCMC is run for. Therefore, the results will vary somewhat across independent runs of the program (differing in the random seed). You should check that the conclusions you reach are not sensitive to this variation by comparing results over multiple long runs of the program.

1.1 Disclaimer

The code is not written by professional programmers and therefore it neither looks nor behaves like professional code. However, it should provide the desired functionality.

We hope that the software is useful, but we provide it without any warranty for correctness, stability or anything else.

1.2 Contact

For questions or to report bugs, please contact `torsten.guenther@ebc.uu.se`

2 Input files

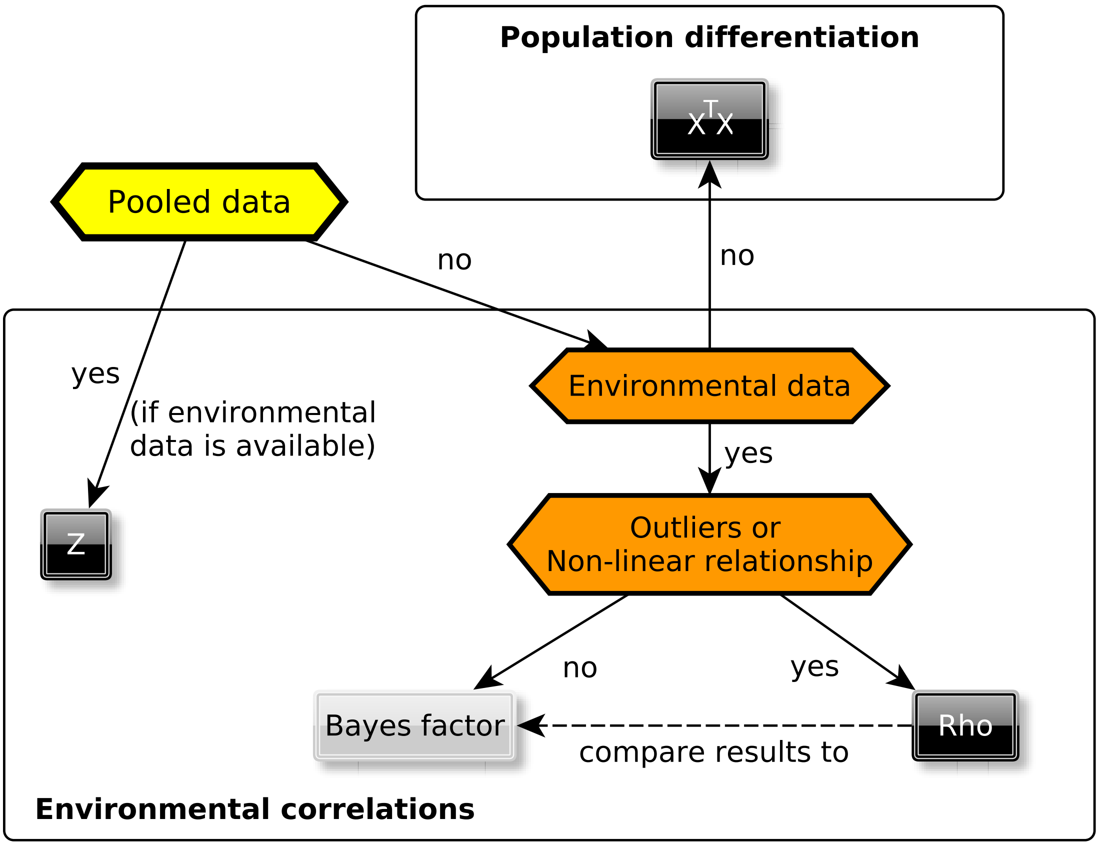
Population differentiation
X^TX
Pooled data
no
no
yes
(if environmental data is available)
Environmental data
yes
Z
Outliers or Non-linear relationship
no
yes
Bayes factor
compare results to
Rho
Environmental correlations

Figure 1: The options of Bayenv2.0 depending on the type of available data.

SNPSFILE contains the allele counts across populations of each SNP are represented by two lines in the file, with the counts of allele one on the first line and the counts for second allele on the second. The counts of allele 1 and allele 2 are assumed to sum to the sample size typed at this SNP in this population (i.e. the total sample size excluding missing data).

The counts should be separated by tabs. An example file is hgdp_no_X_37_freqs, which contains the allele count data for the autosomal SNP data for the 2333 SNPs typed in the 52 human populations of the HGDP panel by Conrad et al. (2006 Nat. Gen).

For example the last 5 columns of the 1st 4 rows of hgdp_no_X_37_freqs are:

```
 0   0   0   0   0
42  24  16  22  32
 2   0   0   1   1
40  24  16  21  31
```

This represents the allele counts of 2 SNPS in 5 populations. At the second SNP in the first population allele 1 has been found 2 times, while allele 2 has been found 40 times. (Note that there should be just polymorphic sites in the data set.)

SNPFILE The input file for the data is identical in format, but now the file must contain the count data for a single SNP. The allele counts must be in the same population order that they appeared the in the covariance matrix, i.e. the order that they appeared in SNPSFILE. An example input file is given in the the file rs316.

ENVIRONFILE is the file of environmental variables that you wish to estimate correlation statistics for. Each environmental variable should be standardized, i.e. subtract the mean and then divided through by the standard deviation of the variable across populations. Each line of the file is an environmental variable across populations, values should be tab separated. The variables should be listed in the same population order as they appear in the allele count files.

MATRIXFILE a single covariance matrix, the entries in the matrix are separated by a tabs. This can be obtained by copying and pasting the last printed matrix from the output of the program. An example covariance matrix is given by: hgdp_matrix_1. If you are concerned about variation across the draws of the covariance matrix then you may consider using the mean covariance matrix over iterations here as input.

SAMPLEFILE is the file containing the sample sizes per population when the populations have been sequenced as pools (i.e. the number of chromosomes per pool). The file contains a single line of tab separated values. The sample sizes must be in the same population order that they appeared the in the covariance matrix, i.e. the order that they appeared in SNPSFILE. An example input file is given in the the file samplesizes.txt.

2.1 Converting files into Bayenv format

Since version 2.1.0.0, the popular data conversion tool PGDSpider (`http://www.cmpg.unibe.ch/software/PGDSpider/`; Lischer & Excoffier, *Bioinformatics* 2012) provides the option to convert from several widely used file formats (e.g. VCF, PED) into the file format used by Bayenv2.0. Please contact the authors of PGDSpider for support.

3 Matrix estimation

The first step for **all** Bayenv2.0 analyses is to estimate the covariance matrix. This should be done using a large number of markers with no or loose linkage disequilibrium between them. These SNPs should be well matched in ascertainment scheme to those that will be tested (see the papers for more discussion). Note that estimating the covariance matrix takes a time approximately linear in the number of SNPs used, thus if you are analyzing a very large data set you will probably want to estimate the matrix for a randomly chosen subset of the data. In general we have found that given a large number of SNPs (e.g. thousands) the program converges relatively quickly to a small set of very similar covariance matrices.

The command line for estimating the matrix is:

```
./bayenv2 -i SNPSFILE -p NUMPOPS -k 100000 -r 63479 > matrix.out
```

This outputs the current draw of the covariance matrix into matrix.out every 500 iterations.

The rows and columns in the covariance matrix appear in the same population order as they appeared in the allele count file. The MATRIXFILE used in subsequent steps of bayenv2 can be obtained by copying and pasting the last printed matrix from the output of the program. The covariance matrix can be visualized by the image command in R, the cov2cor function can be used to convert a covariance matrix into a correlation matrix. The correlation matrix computed from the estimated covariance matrix should be very correlated with the matrix of pairwise F_{ST} (see the paper for discussion). The matrix should be inspected for unexpected low or high correlations, as judged from pairwise F_{ST}, as these may indicate problems in the labeling of populations. Note that the entries of the covariance matrix are not forced to be positive, thus small negative entries in the covariance matrix are possible. Matrices should be compared within an across independent runs of the program to ensure that the matrix is well estimated.

4 Environmental correlations

Detecting environmental correlations is the default mode, which has been implemented for Bayenv1.0. To estimate the Bayes factor for a SNP (input: SNPFILE) for an environmental variable (input: ENVIRONFILE) the each the program must be run separately on each SNP.

Example: `./bayenv2 -i SNPFILE -m MATRIXFILE -e ENVIRONFILE -p NUMPOPS -k 100000 -n 5 -t -r 429`

This example estimates Bayes factors for the allele frequencies of SNPFILE and five environmental variables in ENVIRONFILE. If no specific OUTFILE is requested (`-o`), the resulting Bayes factors for the SNP are appended to a file called bf_environ.ENVIRONFILE. The first column gives the SNPFILE subsequent columns give the estimated Bayes factor for each environmental variable, in the order that they appear in the file. The output is appended to the file, and so runs for different SNPs will appear sequentially in the file. When selecting candidate loci, please also consider the face values of BFs, see for example Kass and Raftery (Journal of the American Statistical Association, 1995) on the interpretation of Bayes factors.

4.0.1 A small bash script: calculate BFs for all SNPs of SNPSFILE

The archive of Bayenv2 contains a small bash script (calc_bfs.sh) to calculate Bayes factors for all SNPs in SNPSFILE. It splits up all SNPs of SNPSFILE into separate files (numbered in the same order as in SNPSFILE) and then runs Bayenv2 (if bayenv2 is located in the same directory). The separate files are automatically deleted directly after running of the script.

Example: `./calc_bfs.sh SNPSFILE ENVIRONFILE MATRIXFILE NUMPOPS NUMITER NUMENVIRON`

Please note that the order of input parameters is important for this script.

4.1 Non-parametric tests

Sometimes the linear model underlying the Bayes factors might not be correct or outliers might misguide the model. For such cases, Bayenv2.0 calculates standardized allele frequencies (X), from which the covariance structure among populations was removed. Bayenv2.0 calculates the non-parametric Spearman's rank correlation coefficient ρ in addition to Bayes factors, when the `-c` flag is set.

Example: `./bayenv2 -i SNPFILE -m MATRIXFILE -e ENVIRONFILE -p NUMPOPS -k 100000 -n 1 -t -c -r 24542`

This example calculates Bayes factors, Spearman's ρ and Pearson's correlation coefficient (for comparison) for a single environmental variable. If no specific OUTFILE is requested (`-o`), the resulting Bayes factors and correlation coefficients for the SNP are appended to a file called bf_environ.ENVIRONFILE. The first column gives the name of SNPFILE followed by a column showing the Bayes factor (see above), a column showing Spearman's ρ and a column showing Pearson's correlation coefficient r_S. If more than one environmental variables are used, additional column trios (BF, ρ and r_S) are appended to the output file. For users worried about outliers, we suggest to compare the results of rho and BFs. SNPs highly ranked in BF lists might be affected by outliers, but if they are also supported by a high ρ, the signal can be considered to be robust. As the linear method underlying BF has slightly higher power, we suggest to consider all SNPs as robust candidates that are among the top x % of BF and among the top y % of the absolute values of ρ ($x < y$, please note that ρ ranges between -1 and 1, and high absolute values suggest non-zero correlation).

The standardized allele frequencies could be used in arbitrary tests for correlation. For instance, some researchers might be interested in testing multiple environmental variables in the same linear model. Bayenv2.0 offers the `-f` flag which writes the standardized allele frequencies X into a file called SNPFILE.freqs (which will be placed in the same directory as SNPFILE). Each line represents one sample of X from the MCMC with the columns in the same order as populations in SNPFILE. We suggest to calculate test statistics for each line of X separately and then average across all values. This utilizes the exploration of the parameter space of the MCMC and should integrate the uncertainty of allele frequency estimates due to finite sample sizes. Additionally, we write the transformed environmental variable Y' (please see our paper for more details) into a file standardized.env. It should be straightforward to read this file into other analysis softwares (e.g. R) for further analyses.

4.2 Pooled NGS data

One of the new features of Bayenv2.0 is the support of pooled NGS data. Pooled sequencing of multiple individuals from the same population is a cost efficient way of estimating allele frequencies with increasing popularity. To deal with the different levels of sampling variance (sampling of individuals and sequencing reads), we designed the new test statistic Z. Z ranges between 0 and 0.5, where $Z = 0.5$ corresponds to full support of a non-zero correlation between environmental variables and allele frequencies (across all MCMC iterations) whereas $Z = 0$ corresponds to full support for no correlation.

Example:

```
./bayenv2 -i SNPSFILE -s SAMPLEFILE -p NUMPOPS -k 200000 -x -r 83556
```
(Matrix Estimation)

```
./bayenv2 -i SNPFILE -s SAMPLEFILE -p NUMPOPS -n 1 -e ENVIRONFIL -t -k 200000 -x -r 436216
```
(Test mode)

The general usage of Bayenv2.0 with pooled data is very similar to the use with genotype data. To enter the pooled mode, the flag `-x` has to be set for matrix estimation and test mode. **SNPSFILE/SNPFILE** for the data is identical in format to the files described, but now the file must contain the read count data for the two alleles per SNP. The read counts must be in the same population order that they appeared the in the covariance matrix, i.e. the order that they appeared in SNPSFILE. Additionally, a file including the sample sizes (SAMPLEFILE) per pool (i.e.number of sampled chromosomes per population) is required. The output is written into a file named environ_corr.ENVIRONFILE.

Our experience is, that more MCMC iterations are needed for pooled data. We suggest to double the number of iterations of the non-pooled mode.

Please note: Z cannot be calculated for more than one environmental variable in a single run.

Please note: The current version of Bayenv2.0 does not support the usage of any statistics based on the standardized allele frequencies (X, ρ, X^TX) for pooled data. The output files written when the options `-f`, `-c` or `-X` are used with pooled data do not contain reliable results.

5 Population differentiation

In addition to environmental correlations, Bayenv2.0 calculates a new population differentiation statistic called X^TX. This test is similar to the classical F_{ST} as highly differentiated SNPs might be driven by local adaptation. Environmental correlations provide high power when the driving environmental variable is known, whereas differentiation statistics also detect responses to other (maybe unknown) environmental conditions. As this statistic is based on the standardized allele frequencies X, it is powerful to identify loci that are more differentiated than expected under pure drift among populations. The expected value for X^TX equals the number of populations. However, we observed strong deviations in our data analyses so we suggest to rank SNPs empirically. If a reliable demographic model is available, simulations could be used to obtain a null distribution of $X^T X$.

Example: `./bayenv2 -i SNPFILE -m MATRIXFILE -e ENVIRONFILE -n 1 -p NUMPOPS -k 100000 -t -X -r 13258`

$X^T X$ will be calculated when the `-X` flag is set in test mode. The current version of Bayenv2.0 writes the $X^T X$ statistics into a file XtX_out.ENVIRONFILE. The first column gives the SNPFILE and the second column gives the estimate for $X^T X$ averages across multiple samples from the MCMC. The output is appended to the file, and so runs for different SNPs will appear sequentially in the file. Currently, Bayenv2.0 requires to read an environmental file although the environmental variables are not used for the calculation. If no environmental data is available, ENVIRONFILE could simply include some dummy values.

6 Summary of options

Option	Description
`-i SNPFILE`	input file (always required)
`-e ENVIRONFILE`	environmental file (required for test mode)
`-m MATRIXFILE`	matrix file (output of Bayenv2.0, required for test mode)
`-s SAMPLEFILE`	file of sample sizes per population (only required for pool mode)
`-k NUMRUNS`	number of iterations (always required)
`-r SEED`	random seed (integer)
`-p NUMPOPS`	number of populations (always required)
`-n NUMENVIRON`	number of environmental variables (required for test mode)
`-t`	rest mode (not matrix estimation), calculates Z, BF or ρ for single SNPs
`-x`	pool mode (to be used when the input data originates from pooled sequencing of populations), calculate Z instead of BF (when in test mode, only one environmental variable)
`-o OUTFILE`	(optional) custom name of output file (only available in test mode)
`-f`	write standardized allele frequencies X into a file
`-c`	calculate ρ in addition to BF
`-X`	calculate $X^T X$
`-z`	calculate Z in test mode for unpooled data (only one environmental variable)