# MACAU User Manual

Xiang Zhou

March 15, 2017

## Contents

# 1 Introduction

## 1.1 What is MACAU

MACAU is the software implementing the Mixed model association for Count data via data AUgmentation algorithm. MACAU can be used to perform differential methylation analysis in bisulfite sequencing studies and differential expression anlaysis in RNA sequencing studies. It fits either a binomial mixed model (for bisulfite sequencing data) or a Poisson mixed model (for RNA sequencing data) to account for population stratification and structure and directly works with the raw read counts. It is computationally efficient for large scale studies and uses freely available open-source numerical libraries.

## 1.2 How to Cite MACAU

- Differential Methylation Analysis for Bisulfite Sequencing Studies

  Amanda J Lea, Jenny Tung and Xiang Zhou (2015). A flexible, efficient binomial mixed model for identifying differential DNA methylation in bisulfite sequencing data. PLoS Genetics 11: e1005650.

- Differential Expression Analysis for RNA Sequencing Studies

  Shiquan Sun, Michelle Hood, Laura Scott, Qinke Peng, Sayan Mukherjee, Jenny Tung and Xiang Zhou (2017). Differential Expression Analysis for RNAseq using Poisson Mixed Models. Nucleic Acids Research. in press.

## 1.3 The Model

### 1.3.1 Binomial Mixed Model

To detect differentially methylated sites in bisulfite sequencing studies, MACAU models each potential target of DNA methylation one site at a time. For each site, MACAU considers the following binomial mixed model (BMM):

$$y_i \sim \text{Bin}(r_i, \pi_i),$$

where $r_i$ is the total read count for $i$th individual; $y_i$ is the methylated read count for that individual, constrained to be an integer value less than or equal to $r_i$; and $\pi_i$ is an unknown parameter that represents the true proportion of methylated reads for the individual at the site. We use a logit

link to model $\pi_i$ as a linear function of parameters:

$$\begin{aligned}
\text{logit}(\pi_i) = \log(\lambda_i) &= \mathbf{w}_i^T \boldsymbol{\alpha} + x_i\beta + g_i + e_i, \\
\mathbf{g} = c(g_1, \cdots, g_n)^T &\sim \text{MVN}(0, \sigma^2 h^2 \mathbf{K}), \\
\mathbf{e} = c(e_1, \cdots, e_n)^T &\sim \text{MVN}(0, \sigma^2(1-h^2)\mathbf{I}_{n\times n}),
\end{aligned}$$

where logit denotes a logistic transformation $\text{logit}(\pi_i) = \log(\frac{\pi_i}{1-\pi_i})$; $\lambda_i = \frac{\pi_i}{1-\pi_i}$ is the odds; $\mathbf{w}_i$ is a $c$-vector of covariates including an intercept and $\boldsymbol{\alpha}$ is a $c$-vector of corresponding coefficients; $x_i$ is the predictor of interest and $\beta$ is its coefficient; $\mathbf{g}$ is an $n$-vector of genetic random effects that model correlation due to population structure or individual relatedness; $\mathbf{e}$ is an n-vector of environmental residual errors that model independent variation; $\mathbf{K}$ is a known $n$ by $n$ relatedness matrix that can be calculated based on a pedigree or genotype data and that has been standardized to ensure $tr(\mathbf{K})/n = 1$ (this ensures that $h^2$ lies between 0 and 1, and can be interpreted as heritability, see [4]); $\mathbf{I}$ is an $n$ by $n$ identity matrix; $\sigma^2 h^2$ is the genetic variance component; $\sigma^2(1-h^2)$ is the environmental variance component; $h^2$ is the heritability of the logit transformed methylation proportion (i.e. $\text{logit}(\pi)$); and MVN denotes the multivariate normal distribution.

### 1.3.2 Poisson Mixed Model

To detect differentially expressed genes in RNA sequencing studies, MACAU models one gene at a time. For each gene, MACAU considers the following Poisson mixed model (PMM):

$$y_i \sim \text{Poi}(N_i \lambda_i),$$

where for the $i$th individual, $y_i$ is the number of reads mapped to the gene (or isoform); $N_i$ is the total read counts for that individual summing read counts across all genes; and $\lambda_i$ is an unknown Poisson rate parameter. We model the log-transformed rate $\lambda_i$ as a linear combination of several parameters:

$$\begin{aligned}
\log(\lambda_i) &= \mathbf{w}_i^T \boldsymbol{\alpha} + x_i\beta + g_i + e_i, \\
\mathbf{g} = c(g_1, \cdots, g_n)^T &\sim \text{MVN}(0, \sigma^2 h^2 \mathbf{K}), \\
\mathbf{e} = c(e_1, \cdots, e_n)^T &\sim \text{MVN}(0, \sigma^2(1-h^2)\mathbf{I}_{n\times n}),
\end{aligned}$$

where $\mathbf{w}_i$ is a $c$-vector of covariates including an intercept and $\boldsymbol{\alpha}$ is a $c$-vector of corresponding coefficients; $x_i$ is the predictor of interest and $\beta$ is its coefficient; $\mathbf{g}$ is an $n$-vector of genetic random effects that model correlation due to population structure or individual relatedness; $\mathbf{e}$ is an n-vector of environmental residual errors that model independent variation; $\mathbf{K}$ is a known $n$ by $n$ relatedness matrix that can be calculated based on a pedigree or genotype data and that has been standardized to ensure $tr(\mathbf{K})/n = 1$ (this ensures that $h^2$ lies between 0 and 1, and can be

interpreted as heritability, see [4]); $\mathbf{I}$ is an $n$ by $n$ identity matrix; $\sigma^2 h^2$ is the genetic variance component; $\sigma^2(1-h^2)$ is the environmental variance component; $h^2$ is the heritability of the logit transformed methylation proportion (i.e. $\text{logit}(\pi)$); and MVN denotes the multivariate normal distribution.

### 1.4 Hypothesis Test

MACAU tests the null hypothesis $H_0 : \beta = 0$ for each unit (site or gene) in turn. It uses a sampling based approach to compute an approximate maximum likelihood estimate $\hat{\beta}$, its standard error $se(\hat{\beta})$ and the corresponding $p$ value.

### 1.5 Missing Data

No missing data is allowed in the count table. Missing data is allowed in the predictor variable file and covariate file, but individuals with missing data will not be included in the analysis.

## 2 Installing and Compiling MACAU

If you have downloaded a binary executable, no installation is necessary. In some cases, you may need to use "chmod a+x macau" before using the binary executable. In addition, notice that the end-of-line coding in Windows (DOS) is different from that in Linux, and so you may have to convert input files using the utility *dos2unix* or *unix2dos* in order to use them in a different platform. Sometimes you can use the following command in Linux to convert files

```
tr '\r' '\n' < input.txt > output.txt
```

If you want to compile MACAU by yourself, you will need to download the source code, and you will need a standard C/C++ compiler such as GNU gcc, as well as the GSL and LAPACK libraries. You will need to change the library paths in the Makefile accordingly. A sample Makefile is provided along with the source code. For details on installing GSL library, please refer to `http://www.gnu.org/s/gsl/`. For details on installing LAPACK library, please refer to `http://www.netlib.org/lapack/`.

# 3 Input File Format

MACAU requires four input files containing methylated read counts, total read counts, relatedness matrix, predictor variable of interest. One can also provide an optional covariate file.

## 3.1 Count File

This file contains a table of read counts with a header. The first column lists site IDs while the first row lists individual IDs. An example file with two sites and four individuals is as follows:

```
site idv1 idv2 idv3 idv4
site1 2 4 3 8
site2 3 0 15 9
```

Both methylated read count file and total read count file are in the above format. The gene expression read count file is also in the above format.

## 3.2 Predictor File

This file contains the predictor variable of interest. Each line is a number indicating the phenotype value for each individual in turn, in the same order as in the count files. Notice that only numeric values are allowed and characters will not be recognized by the software. Missing phenotype information is denoted as NA. The number of rows should be equal to the number of individuals in the mean genotype file. An example predictor file with four individuals as follows:

```
1.2
NA
2.7
-0.2
```

One can include multiple predictors as multiple columns in the phenotype file, and specify a different column for the association tests by using "-n [num]", where "-n 1" uses the original first column as phenotypes, and "-n 2" uses the second column, and so on and so forth. An example predictor file with four individuals and three predictor variables is as follows:

```
1.2  -0.3  -1.5
NA    1.5   0.3
2.7   1.1   NA
-0.2 -0.7   0.8
```

### 3.3 Relatedness Matrix File

MACAU, as a mixed model software, requires a relatedness matrix file. It contains a $n \times n$ matrix, where each row and each column corresponds to individuals in the same order as in the count file or in the predictor file, and $i$th row and $j$th column is a number indicating the relatedness value between $i$th and $j$th individuals. An example relatedness matrix file with three individuals is as follows:

```
0.3345  -0.0227   0.0103
-0.0227  0.3032  -0.0253
0.0103  -0.0253   0.3531
```

### 3.4 Covariates File (optional)

One can provide an optional covariates file for fitting BMM. The covariates file is similar to the above BIMBAM multiple predictor file. An example covariates file with four individuals and three covariates (the first column is the intercept) is as follows:

```
1  1  -1.5
1  2   0.3
1  2   0.6
1  1  -0.8
```

If a column of 1s is not provided, then the software will automatically add one at the end of the covariate matrix.

# 4 Running MACAU

## 4.1 Example Datasets

On the software website, you will find two example datasets: a bisulfite sequencing data and a RNA sequencing data.

The bisulfite sequencing data contains 438,865 methylation sites on 20 chromosomes for 50 baboons. The detailed processing steps are described in [1]. The" mcounts_chr[num]_n50.txt" file contains methylated read counts for each site; the "counts_chr[num]_n50.txt" file contains total read counts, the "predictor_n50.txt" file contains the predictor variable of interest (age), and the "relatedness_n50.txt" file contains the kinship matrix estimated from genotypes using GEMMA [5, 4, 6]. To run MACAU on this data set, simply type

```
./bin/macau -g example/mcounts_chr20_n50.txt -t example/counts_chr20_n50.txt
-p example/predictor_n50.txt -k example/relatedness_n50.txt -bmm -o chr20
```

For convenience, the analyses results for chr20 are also included inside the output folder in the same directory.

The RNA sequencing data contains 12,018 genes for 63 baboons. The detailed processing steps are described in [3, 2]. The"expression.txt" file contains read counts for each gene; the "predictor.txt" file contains the predictor variable of interest (sex); the "covariates.txt" file contains the covariates (intercept plus gene expression PCs); and the "kinship.sXX.txt" file contains the kinship matrix estimated from genotypes using GEMMA [5, 4, 6]. To run MACAU on this data set, simply type

```
./bin/macau -g example/RNAseq/expression.txt -p example/RNAseq/predictor.txt
-c example/RNAseq/covariates.txt -k example/RNAseq/kinship.sXX.txt
-pmm -o rna_wcvt
```

For convenience, the analyses results are included inside the output folder in the same directory.

## 4.2 Estimate Relatedness Matrix from Genotypes

MACAU requires a pre-computed relatedness matrix. If you only have SNP genotypes, you can use GEMMA [5, 4, 6] to compute the relatedness matrix. For details, please refer to the GEMMA manual available at `www.xzlab.org/software.html`.

## 4.3 Association Tests with Binomial Mixed Models for Bisulfite Sequencing Studies

### 4.3.1 Basic Usage

The basic usages for association analysis are:

```
./macau -g [filename] -t [filename] -p [filename] -k [filename] -bmm -o [prefix]
```

where the "-g [prefix]" specifies the methylated reads count file name; "-t [filename]" specifies the total reads count file name; "-p [filename]" specifies the predictor variable file name; "-k [filename]" specifies relatedness matrix file name; "-o [prefix]" specifies output file prefix.

The software computes approximate maximum likelihood estimate, its standard error and the corresponding $p$-value for each site via a sampling-based algorithm. One can increase the number of sampling iterations by changing "-s [num]" to improve accuracy. One can also change the proposal distributions by assigning "-hscale [num]" and "-sscale [num]" to improve mixing.

The software only analyzes sites where the total read counts are non-zero for at least two individuals. In addition, the software also filters out sites where the likelihood is not informative. This is done by filtering out sites where the ratio between the posterior variance and prior variance of $\beta$ is above a certain threshold. Use "-ratio [num]" to change this threshold.

#### 4.3.2 Output Files

There will be two output files, both inside an output folder in the current directory. The "prefix.log.txt" file contains some detailed information about the running parameters and computation time. The "prefix.assoc.txt" contains the results. An example assoc file with a few sites is shown below:

```
id      n acpt_rate    beta    se_beta    pvalue       h       se_h    sigma2   se_sigma2   alpha0   se_alpha0
4_1035 22 4.469e-01 1.559e+00 1.050e+00 1.377e-01 4.898e-01 2.808e-01 1.046e+01 8.264e+00 -5.882e-01 9.475e-01
4_1038 22 4.509e-01 1.234e+00 1.506e+00 4.126e-01 4.335e-01 2.787e-01 2.713e+01 1.612e+01 -8.427e-01 1.187e+00
4_1049 22 4.533e-01 9.742e-01 7.089e-01 1.694e-01 5.092e-01 2.771e-01 3.673e+00 2.394e+00 1.032e+00 6.738e-01
```

The columns are: site id, number of individuals analyzed at the given site, acceptance rate, $\hat{\beta}$, $se(\hat{\beta})$, $p$-value, $\hat{h}$, $se(\hat{h})$, $\hat{\sigma}^2$, $se(\hat{\sigma}^2)$, estimates and the corresponding standard errors for other coefficients. The estimates for other coefficients are in the same order as provided in the covariate file; in the case when a column of 1s is not provided in the covariate file, then the intercept is provided at the end.

### 4.4 Association Tests with Poisson Mixed Models for RNA Sequencing Studies

#### 4.4.1 Basic Usage

The basic usages for association analysis are:

```
./macau -g [filename] -p [filename] -k [filename] -bmm -o [prefix]
```

where the "-g [prefix]" specifies the gene expression reads count file name; "-p [filename]" specifies the predictor variable file name; "-k [filename]" specifies relatedness matrix file name; "-o [prefix]" specifies output file prefix.

The software computes approximate maximum likelihood estimate, its standard error and the corresponding $p$-value for each site via a sampling-based algorithm. One can increase the number of sampling iterations by changing "-s [num]" to improve accuracy. One can also change the proposal distributions by assigning "-hscale [num]" and "-sscale [num]" to improve mixing.

The software only analyzes sites where the read counts are non-zero for at least one individual. In addition, the software also filters out sites where the likelihood is not informative. This is done by filtering out sites where the ratio between the posterior variance and prior variance of $\beta$ is above a certain threshold. Use "-ratio [num]" to change this threshold.

### 4.4.2 Output Files

There will be two output files, both inside an output folder in the current directory. The "prefix.log.txt" file contains some detailed information about the running parameters and computation time. The "prefix.assoc.txt" contains the results. An example assoc file with a few sites is shown below:

```
id      n  acpt_rate   beta    se_beta   pvalue       h       se_h     sigma2    se_sigma2  alpha0   se_alpha0
MARCH1 63 4.694e-01 4.238e-02 6.919e-02 5.402e-01 4.242e-01 2.769e-01 5.396e-02  1.360e-02 -1.111e+01 1.106e-01
SEPT11 63 4.488e-01 1.835e-02 5.623e-02 7.441e-01 4.336e-01 2.605e-01 4.136e-02  9.108e-03 -1.003e+01 9.110e-02
```

The columns are: site id, number of individuals analyzed at the given site, acceptance rate, $\hat{\beta}$, $se(\hat{\beta})$, $p$-value, $\hat{h}$, $se(\hat{h})$, $\hat{\sigma}^2$, $se(\hat{\sigma}^2)$, estimates and the corresponding standard errors for other coefficients. The estimates for other coefficients are in the same order as provided in the covariate file; in the case when a column of 1s is not provided in the covariate file, then the intercept is provided at the end.

# 5 Options

File I/O Related Options

- -g [prefix] specify the methylated read counts file name
- -t [filename] specify the total read counts file name
- -p [filename] specify the predictor variable file name
- -n [num] specify which predictor variable column to use analysis (default 1)
- -k [filename] specify the kinship/relatedness matrix file name
- -v [filename] specify the residual error variance matrix file name (optional; default uses an identity matrix)
- -c [filename] specify the covariates file name (optional)
- -o [prefix] specify output file prefix (default "result")
- -ratio [num] specify the filtering ratio threshold (default 0.95)
- -outdir [pathname] specify output path (optional; default "./output")
- -pace [num] specify terminal display update pace (optional; default 1,000 sites)
- -silence silence terminal display (optional; default 1,000)

Algorithm Options

- -vb [num] specify prior variance for beta (default 1)
- -h [num] [num] [num] specify prior values for h: min, max, num (default 0.01 0.99 10)
- -hscale [num] specify proposal, scale for h (default min(2/sqrt(n),1) )
- -sscale [num] specify proposal, standard deviation for log(sigma2) (default min(4/sqrt(n),1) )
- -w [num] specify burn-in steps (default 100)
- -s [num] specify sampling steps (default 1,000)
- -mh [num] specify number of MH steps in each iteration (default 10)
- -seed [num] specify random seed (a random seed is generated by default)

## References


[1] Amanda J. Lea, Jenny Tung, and Xiang Zhou. A flexible, efficient binomial mixed model for identifying differential DNA methylation in bisulfite sequencing data. *PLoS Genetics*, 11:e1005650, 2015.

[2] Shiquan Sun, Michelle Hood, Laura Scott, Qinke Peng, Sayan Mukherjee, Jenny Tung, and Xiang Zhou. Differential expression analysis for RNAseq using Poisson mixed models. *Nucleic Acids Research*, 2017.

[3] Jenny Tung, Xiang Zhou, Susan C Alberts, Matthew Stephens, and Yoav Gilad. The genetic architecture of gene expression levels in wild baboons. *eLife*, 4:e04729, 2015.

[4] Xiang Zhou, Peter Carbonetto, and Matthew Stephens. Polygenic modelling with Bayesian sparse linear mixed models. *PLoS Genetics*, 9:e1003264, 2013.

[5] Xiang Zhou and Matthew Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, 44:821–824, 2012.

[6] Xiang Zhou and Matthew Stephens. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature Methods*, 11:407–409, 2014.