MACAU 2.0 User Manual

Shiquan Sun, Jiaqiang Zhu, and Xiang Zhou
Department of Biostatistics, University of Michigan
shiquans@umich.edu and xzhousph@umich.edu

April 9, 2017

Contents

1 Introduction

1.1 What is MACAU 2.0

MACAU 2.0 is an R package implementing the Mixed model association for Count data via penalized quasi-likelihood [2]. MACAU 2.0 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 2.0

- Sun, S., et al. MACAU 2.0: Efficient Mixed Model Analysis of Count Data in Large-Scale Genomic Sequencing Studies, bioR$\mathcal{X}$iv.
- Sun, S., et al. (2017) Differential expression analysis for RNAseq using Poisson mixed models, Nucleic Acids Res, In press.
- Lea, A.J., et al. (2015) A flexible, efficient binomial mixed model for identifying differential DNA methylation in bisulfite sequencing data, Plos Genet, 11, e1005650.

1.3 The Model

1.3.1 Binomial Mixed Model

To detect differentially methylated sites in bisulfite sequencing studies, MACAU 2.0 models each potential target of DNA methylation one site at a time. For each site, MACAU 2.0 considers the following binomial mixed model (BMM):

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

where r_i is the total read count for ith 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 π_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 π_i as a linear function of parameters:

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

where logit denotes a logistic transformation $\mathrm{logit}(\pi_i) = \log(\frac{\pi_i}{1-\pi_i})$; $\lambda_i = \frac{\pi_i}{1-\pi_i}$ 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 β 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 [1]); $\mathbf{I}$ 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. $\mathrm{logit}(\pi)$); and MVN denotes the multivariate normal distribution.

1.3.2 Poisson Mixed Model

To detect differentially expressed genes in RNA sequencing studies, MACAU 2.0 models one gene at a time. For each gene, MACAU 2.0 considers the following Poisson mixed model (PMM):

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

where for the ith 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 λ_i is an unknown Poisson rate parameter. We model the log-transformed rate λ_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 β 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 [1]); $\mathbf{I}$ 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 2.0 tests the null hypothesis $H_0 : \beta = 0$ for each unit (site or gene) in turn. It uses a quasi-likelihood based approach to compute an approximate maximum likelihood estimate $\hat{\beta}$, its standard error $se(\hat{\beta})$ and the corresponding p value.

2 Installation

The R package of MACAU2 is freely available at http://www.xzlab.org/software.html. It can be installed as a regular R package on a UNIX/Linux operating system. Currently, MACAU2 do not support Mac or Windows operating system.

2.1 Packages Dependence

MACAU2 depends on several existing packages which we list as follows:

- C++ library Armadillo: http://arma.sourceforge.net/
- R package Rcpp: https://cran.r-project.org/web/packages/Rcpp/index.html
- R package RcppArmadillo: https://cran.r-project.org/web/packages/RcppArmadillo/index.html
- R package foreach: https://cran.r-project.org/web/packages/foreach/index.html
- R package doParallel: https://cran.r-project.org/web/packages/doParallel/index.html
- R package INLA: http://www.r-inla.org/

Note: If you install RcppArmadillo package, the Rcpp package will be installed automatically. Whatever you use parallel setting, the package foreach must be installed before you runing MACAU 2.0.

2.2 Installing MACAU 2.0

Before installing MACAU2 package, please make sure your system has installed the packages mentioned above Section (Section 2.1).

To install MACAU 2.0 into ”your directory”, in an R session you can use

```
> install.packages("macau-2.00.tar.gz", lib = "your_directory")
```

Alternatively, the development version can be installed from GitHub:

```
Note: INLA package is required for the MACAU2 installation
#install.packages("INLA", repos="https://www.math.ntnu.no/inla/R/stable")
#install.packages("devtools")
devtools::install_github("jakyzhu/MACAU2")
```

This installation might fail if some of the dependency libraries are not yet installed. If so, please run the following lines and repeat the installation.

```
#install.packages("INLA", repos="https://www.math.ntnu.no/inla/R/stable")
#install.packages("Rcpp")
#install.packages("RcppArmadillo")
#install.packages("foreach")
#install.packages("doParallel")
```

3 Running MACAU 2.0

If MACAU 2.0 has been successfully installed into ”your directory”, you can load it in an R session using

```
> library(MACAU2, lib.loc = "your directory")
```

We provide one function (macau2) and two example datasets (ExampleRNAseq and ExampleBSseq) in MACAU2. Details about how to use this function, its argument and returned values can be found in the R help document of MACAU2. For example, to learn more about macau2 function, in a R session you can type

```
?macau2
```

3.1 Fitting PMM

Here we provide a simple example of fitting PMM using MACAU2:

```
# Attach the RNAseq example data
> data(ExampleRNAseq)
> attach(ExampleRNAseq)
> count[1:3,1:5]
# An example read count file with three genes and five individuals is as follows:
       idv1 idv2 idv3 idv4 idv5
 gene1  301 1194 1343  782  385
 gene2  195 3352  678  579  667
 gene3  260 2030  264  807  624

> predictor[1:5]
# An example predictor file with five individuals is as follows:
 [1] -1.13618508  0.01460988  0.50888758  0.52874195  0.48353657

> relatednessmatrix[1:5,1:5]
# An example relatedness matrix file with five individuals is as follows:
             V1           V2          V3          V4          V5
 1  1.980908800 -0.008376381  0.12271015 -0.01753495 -0.51374676
 2 -0.008376381  2.512448718 -0.02862503  0.06850543 -0.07018207
 3  0.122710154 -0.028625033  1.80948916  0.01192036 -0.03083115
 4 -0.017534952  0.068505433  0.01192036  1.28732173 -0.04061807
 5 -0.513746759 -0.070182069 -0.03083115 -0.04061807  2.34092438

> totalcount[1:5]
# An example total count file with five individuals is as follows:
 [1] 25456 80821 97346 62081 45666

# Fit PMM
> fit = macau2(RawCountDataSet=count,Phenotypes=predictor,RelatednessMatrix=relatednessmatrix,
                 LibSize=totalcount,fit.model="PMM",numCore=1)
## number of total individuals:  200
## number of total genes/sites:  100
## number of adjusted covariates:  0
# fitting Poisson mixed model ...
```

```
> fit[1:3,]
# An example of the output with three genes is as follows:
      numIDV        beta    se_beta    pvalue        h2    sigma2 converged
gene1    200  0.02718814 0.04374406 0.5342525 0.2052934 0.3340768      TRUE
gene2    200 -0.01014800 0.04287225 0.8128869 0.5175004 0.3189550      TRUE
gene3    200  0.02816845 0.04565769 0.5372686 0.3542878 0.3555320      TRUE
```

The columns in the output are: gene ID, number of individuals analyzed at the given gene, $\hat{\beta}$, $se(\hat{\beta})$, p-value, $\hat{h}^2$, $\hat{\sigma}^2$, whether the algorithm is converged or not.
The total count file is `optional` for fitting PMM. It will be calculated automatically if not provided.

3.2 Fitting BMM

Here we provide a simple example of fitting BMM using MACAU2. The methylated read count file and total read count file are in the same format.

```
# Attach the BSseq example data
> data(ExampleBSseq)
> attach(ExampleBSseq)
> mcount[1:3,1:5]
# An example read count file with three sites and five individuals is as follows:
      idv1 idv2 idv3 idv4 idv5
site1   23    2    7   10   18
site2    3   14    0    0    0
site3    5    3   17   17   15

> totalcount[1:3,1:5]
# An example total count file with three sites and five individuals is as follows:
      idv1 idv2 idv3 idv4 idv5
site1   31    9   11   15   24
site2    5   15    0    1    2
site3   15    6   28   20   21
> mfit = macau2(RawCountDataSet=mcount,Phenotypes=predictor,RelatednessMatrix=relatednessmatrix,
                  LibSize=totalcount,fit.model="BMM",numCore=1)
## number of total individuals:  200
## number of total genes/sites:  100
## number of adjusted covariates:  0
# fitting binomial mixed model ...
> mfit[1:3,]
# An example of the output with three genes is as follows:
      numIDV         beta    se_beta     pvalue        h2    sigma2 converged
site1    200 -0.016345736 0.03565928 0.64667468 0.2534432 0.9498826      TRUE
site2    196  0.009537827 0.03272432 0.77069959 0.1328485 0.6713365      TRUE
site3    199 -0.080598486 0.03686595 0.02879671 0.3597304 0.8431537      TRUE
```

The columns in the output are: site ID, number of individuals analyzed at the given site, $\hat{\beta}$, $se(\hat{\beta})$, p-value, $\hat{h}^2$, $\hat{\sigma}^2$, whether the algorithm is converged or not. For each site, only the individuals with non-zero total read count are analyzed.
Note: make sure the order of individuals in the relatedness matrix, the order of individuals in the read count file matches the order of individuals in the phenotype file.

3.3 Covariates File (optional)

One can provide an optional covariates file (not provided in the example dataset) for fitting the models using MACAU2.

```
cfit = macau2(RawCountDataSet=mcount,Phenotypes=predictor, Covariates=covariate,
              RelatednessMatrix=relatednessmatrix, LibSize=totalcount,
              fit.model="BMM",numCore=1)
```

An example covariates file with five 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
1  1   1.2
```

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

4 Advanced options

4.1 Parallel parameter

By default we set the number of cores to 1:

```
numCore = 1
```

4.2 Quality assessment parameter

By default at least two individuals that read count (RNAseq data) are the larger than 5, otherwise we filter them out.

```
filtering=TRUE
```

4.3 Model fitting parameters

By default we set the maximum number of iteration to 500 and tolerance to declare convergence to 1e-5:

```
fit.maxiter = 500, fit.tol = 1e-5
```

References

[1] Zhou, X., Carbonetto, P. and Stephens, M.(2013) Polygenic modeling with Bayesian sparse linear mixed models. PLoS Genetics. 9(2): e1003264.

[2] Breslow, N.E. and Clayton, D.G. (1993) Approximate inference in generalized linear mixed models, J Am Stat Assoc, 88, 9-25.

[3] Zhou, X. and Stephens, M. (2012) Genome-wide efficient mixed-model analysis for association studies, Nat Genet, 44, 821-824.

[4] Zhou, X. and Stephens, M. (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nature Methods.

[5] Lea, A.J., et al. (2015) A flexible, efficient binomial mixed model for identifying differential DNA methylation in bisulfite sequencing data, Plos Genet, 11, e1005650.