Two-stage genome-wide association study based on Adaptive Lasso

Version 1.0

Updated at 2022.10.19

Wenyu Yang ([yangwenyurain@126.com](mailto:yangwenyurain@126.com))

Jianbing Yan ([yjianbing@mail.hzau.edu.cn](mailto:yjianbing@mail.hzau.edu.cn))

Yingjie Xiao ([yxiao25@mail.hzau.edu.cn](mailto:yxiao25@mail.hzau.edu.cn))

**1 Introduction**

In this paper, a two-stage genome-wide association analysis (ALGWAS) method based on Adaptive Lasso is proposed. In the first stage, single nucleotide polymorphism (SNP) associated with target traits are screened by Adaptive Lasso, a variable selection method. In the second stage, SNPs selected from the first stage were put into the linear model as covariates for genome-wide scanning.

**2 GETTING STARTED**

ALGWAS is compatible with R, which is user-friendly and easily distributed with free licenses. The analysis of ALGWAS requires an independent ‘msgps’ package for variable selection.

**3 INPUT**

**3.1 Genotypic file**

The genotypic file requires a header row, which shows the names of individuals. The first row shows the names of molecular markers or SNPs. For each SNP, the genotypes, e.g., AA, AT, TT, are required to transformed to integer format as 0, 1 or 2 based on allelic frequency. For example, A is the major allele that AA and TT will be transformed to 0 and 2, the AT is transformed to 1 (Table 1). It means the copy number of minor alleles for a line at a SNP locus, relative to major allele.

**Table 1 Genotypic format of 5 SNPs in 4 individuals**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| line | chr1.s\_131317 | chr1.s\_296236 | chr1.s\_544904 | chr1.s\_605115 |
| Line\_1 | 0 | 0 | 0 | 0 |
| Line\_2 | 0 | 0 | 0 | 2 |
| Line\_3 | 0 | 0 | 0 | 0 |
| Line\_4 | 0 | 0 | 0 | 0 |
| Line\_5 | 1 | 0 | 0 | 2 |

**3.2 Agronomic trait file**

The agronomic trait file must contain a header row, which represents the names of traits. The first column shows the names of the individuals (Table 2). The trait values are numeric and missing values are indicated as NA, following the R coding requirement.

**Table 2 Agronomic format of 2 traits in 5 individuals**

|  |  |  |
| --- | --- | --- |
| line | Trait1 | Trait2 |
| Line\_1 | 71.30 | 177.23 |
| Line\_2 | 69.57 | 211.57 |
| Line\_3 | 71.81 | 224.40 |
| Line\_4 | 71.64 | 178.96 |
| Line\_5 | 71.30 | 213.96 |

**4 USAGE**

**4.1 Before analysis**

To run ALGWAS, the user requires to prepare three data files: genotypic file and agronomic data file. The user requires to keep all files in the same order of individuals before analysis.

**4.2 Read files**

The user could use following codes and demo data we provided to test the function of reading data from disk to R working space. It’s noted that, for novice users, the full path of data files that located in the computer is recommended to avoid possible errors.

***# read genotype***

Geno <- read.csv("./data/Geno\_Demo.txt",

header=TRUE,

sep="",

stringsAsFactors=0,

check.names=FALSE

)

***# read agronomic trait***

phe <- read.table("./data/Phe\_Demo.txt",

header=TRUE,

sep="",

stringsAsFactors=0,

check.names=FALSE

)

**4.3 Obtain the information about the genome**

We develop a user-friendly function ‘*Inf\_geno*’ for the analysis of this step. The tutorial code can be tested as followed:

infall <- Inf\_geno(geno,

win

)

*Notes:*

1. *geno: the genotype.*
2. *win: the scan window of ALGWAS.*

**4.4 Obtain the SNPs selected by Adaptive Lasso**

Based on the Adaptive Lasso, implemented in the ‘*msgps*’ package, we complete the variable selection task. We develop a user-friendly function ‘*ALGWAS\_S1*’ for the analysis of this step. The tutorial code can be tested as followed:

infq <- ALGWAS\_S1 (geno,

phe,

infall,

trid

)

*Notes:*

1. *geno: genotype.*
2. *phe: phenotype.*
3. *infall: the information of the genome.*
4. *trid: the serial number of the trait to be studied*

**4.5 Obtain the result of ALGWAS**

The user can use the function ‘*ALGWAS\_S2*’ to obtain the result of ALGWAS.

res <- ALGWAS\_S2 (geno,

phe,

infall,

infq

trid

)

*Notes:*

1. *geno: genotype.*
2. *phe: phenotype.*
3. *infall: the information of the genome.*
4. *infq: the information of the SNPs seleted by Adaptive Lasso*
5. *trid: the serial number of the trait to be studied*

**5 OUTPUT**

**5.1 Output of genomic information**

The function ‘Inf\_geno’ can give the information of genome. The user can use *infall$snp*, *infall$chr*, *infall$pos*, *infall$posx* to get the name of SNP, the chromosome where the SNP is located, and the location of the SNP, which will be used in the function ‘*ALGWAS\_S1*’.

**5.2 Output of the SNPs selected by Adaptive Lasso**

The function ‘ALGWAS\_S1’ gives the SNPs selected by Adaptive Lasso. The user can use *infq$snp*, *infq$chr*, *infq$pos*, *infq$posx* to get the name of the selected SNP, the chromosome where the selected SNP is located, and the location of the selected SNP, which will be used in the function ‘*ALGWAS\_S2*’.

**5.3 Output of the result of ALGWAS**

The function ‘*ALGWAS\_S2*’ can give the result of ALGWAS. The user can use *res$pval* to get the corresponding p-value corresponding of all SNPs.