**1 Getting started**

* 1. **Downloading OptimGRAMMAR**

GRL can be downloaded https://github.com/YuxinSong-prog/OptimGRAMMAR. It can be installed as a regular R package.

* 1. **Installing OptimGRAMMAR**

GRL links to R packages Rcpp, RcppEigen and RcppArmadillo, and also imports R packages BEDMatrix and data.table. These dependencies should be installed before installing **OptimGRAMMAR**. In addition, OptimGRAMMAR requires PLINK2.0 Software (<http://www.cog-genomics.org/plink/2.0/>) with name “plink2.0” under your run directory. Here is an example for installing **OptimGRAMMAR** and all its dependencies in an R session(assuming none of the R packages other than the default has been installed):

install.packages(c("BEDMatrix ", " data.table ", "Rcpp", " RcppEigen ", " RcppArmadillo "), repos = "https://cran.r-project.org/")

system(“R CMD install OptimGRAMMAR \_1.0.tgz”)

**2 Input**

OptimGRAMMAR requires the phenotype and genotype files in an PLINK BED data frame which also called PLINK 1 binary file and the structure of these files is described in http://www.cog-genomics.org/plink/1.9/formats#bed. How to prepare these data are describe below.

**2.1 Phenotype**

Phenotype should place in the sixth column of “.fam” file. The missing phenotype value for quantitative traits is -9.

**2.2 Genotypes**

OptimGRAMMAR can only take genotype files in “.bed” and “.bim” format.

**3 Running OptimGRAMMAR**

If OptimGRAMMAR has been successfully installed, you can load it in an R session using:

library(OptimGRAMMAR)

and then loading "BEDMatrix "and " data.table "

library(BEDMatrix)

library(data.table)

We provide two functions in OptimGRAMMAR: **Data** function for basic data management including extract phenotypic values, sampling markers, calculate allele frequencies and calculating GRM, saved in an external file, used for GBLUP method or the transformed GRM for large scale dataset and **optimGRAMMAR** function for association tests using optimGRAMMARmethod, joint analysis based on the results of optimGRAMMAR, and outputting association result file then drawing Q-Q and Manhattan plot .

**3.1 Data function**

**Usage**

Data (filename, nsmar, msampling)

Arguments

filename An object class of character: the filename before the suffix of plink files. The three plink file must have a same filename, for example, “filename.bed”, “filename.bim” and “filename.bam”.

nsmar An optional numeric for the number of sampling markers. If this is unset, the “nsmar” default is 5,000.

msampling logicals. If FLASE, entire markers will be used to calculate GRM. If this is unset or for large scale dataset, the “msampling” default is TURE.

Here we provide two simple examples of data management for GRL using Data.

**3.2 optimGRAMMAR function**

**Usage**

optimGRAMMAR (Data, maxh2 = 0.5, opsm = 50000, ,Test = c("Separate","Joint") , Scan = c("Plink2","gcta"),QQ = T, Manh = T)

Arguments

Data An object class of list from the 1st step.

maxh2 Upper limit of polygenic heritability, which is set by default at 0.5.

opsm The total number of SNPs used for optimization, which is set by default at 50000.

AssoTst You can choose correlation analysis software “GCTA” or “PLINK2.0” as prefer, “PLINK2.0” by default.

Test An optional for association test, a test at once or joint analysis. You can choose “Separate” for a test at once and “Joint” for further joint analysis based on the result of “Separate”.

QQ logicals. If TURE, Q-Q plot would be drawn.

Manh logicals. If TURE, Q-Q plot would be drawn.

**4 Output**

**4.1 optimGRAMMAR result**

The Grammar function generates a plink association text output file called “Grammar.PHENO1.glm.linear”. Here we show the header and the first five rows of the example output:

#CHROM POS ID REF ALT A1 TEST OBS\_CT BETA SE T\_STAT P ERRCODE

1 8500 S1\_8500 G A A ADD 2648 0.0755563 0.188783 0.400227 0.689021 .

1 10390 S1\_10390 G A A ADD 2648 -0.00784112 0.238845 -0.0328293 0.973813 .

1 10590 S1\_10590 T A A ADD 2648 -0.202364 0.249746 -0.810281 0.417852 .

1 128159 S1\_128159 C T T ADD 2648 0.15431 0.215808 0.715033 0.474652 .

1 128373 S1\_128373 C T T ADD 2648 0.033819 0.124565 0.271498 0.78603 .

…

In addition to the above file, a QTN candidate file, named “QTNs”, with Bonferroni as the threshold is also output. The header and the first five rows are:

#CHROM POS ID REF ALT A1 TEST OBS\_CT BETA V2

1 222963574 S1\_222963574 G A A ADD 2648 -0.462216 7.65251920613318

2 6597140 S2\_6597140 T C C ADD 2648 1.16828 17.6731608197186

2 6597170 S2\_6597170 G T T ADD 2648 1.91808 43.6572192493634

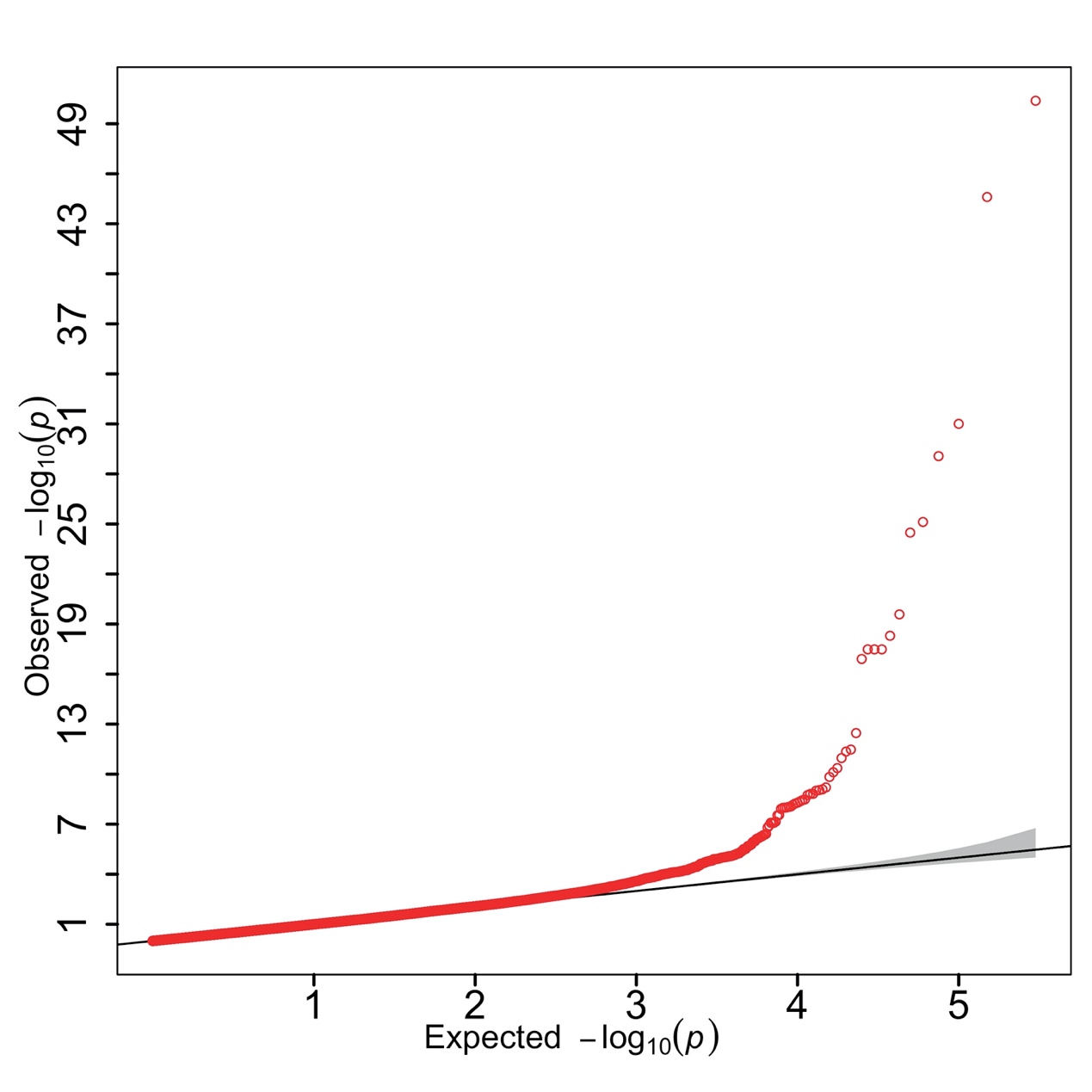
3 165856383 S3\_165856383 T C C ADD 2648 0.489627 7.28229257925678

3 165858424 S3\_165858424 G A A ADD 2648 -0.642925 11.2704149095184

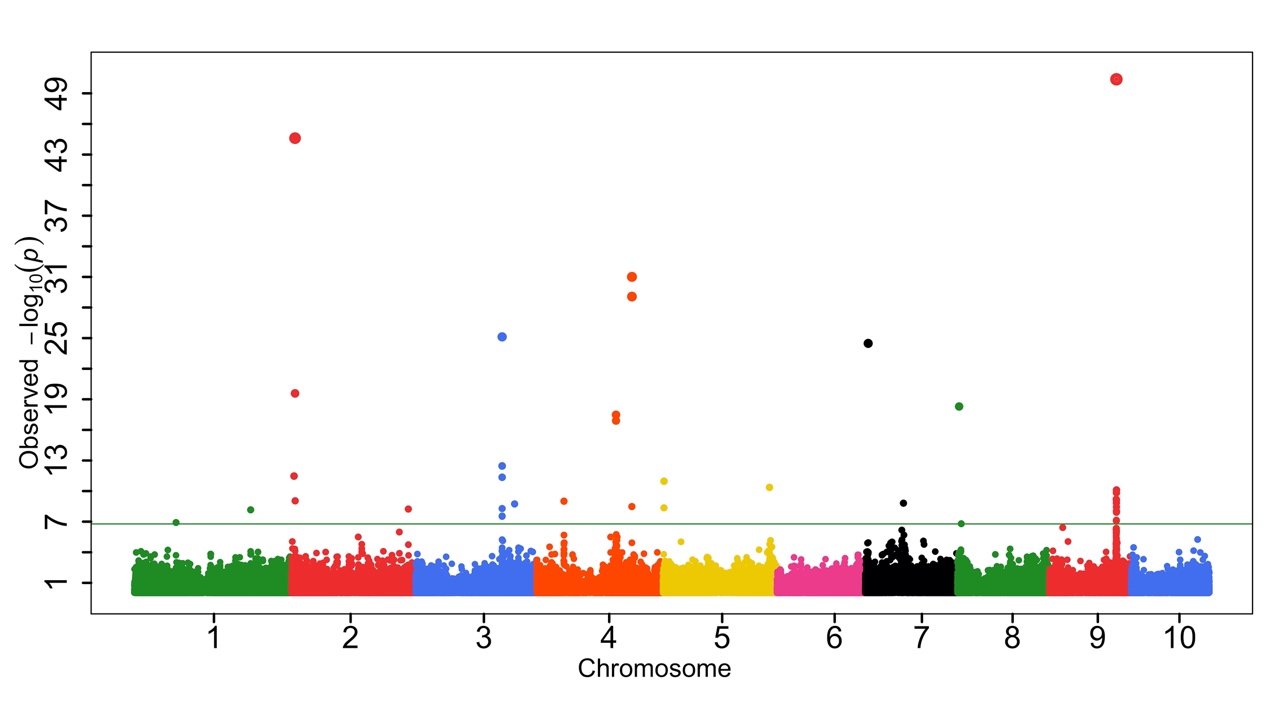
…

If joint analysis, a file, named “jointQTNs”, will display candidate QTNs resulting from joint analysis, which is the same format as “QTNs”.

**4.2 Q-Q plot**

****

**4.3 Manhattan plot**



**5 Example**

library(optimGRAMMAR)

library(BEDMatrix)

library(data.table)

pathdata<-"/Users/songyuxin/Desktop/CFWmice\_binary"

setwd(pathdata)

plinkfilename = "maize"

gdata <- Data(plinkfilename)

optimGRAMMAR (gdata)