# Simultaneous Detection of Signal Regions Using Quadratic Scan Statistics With Applications in Whole Genome Association Studies

**Zilin Li**

**2020-05-15**

The `QSCAN` package accompanies the paper "Simultaneous Detection of Signal Regions Using Quadratic Scan Statistics With Applications in Whole Genome Association Studies" and is designed for detecting rare-variant association regions in whole genome sequencing studies.

## Data Set-up: Generating simulated data

The genotype data was generated by simulating 10,000 samples for a 10 Mb region using COSI. We analyzed the 302,737 low frequency and rare variants (minor allele frequeny, MAF < 0.05) in the sequence.

```
library("Matrix")
library(QSCAN)
genotype <- example$genotype
samplesize <- dim(genotype)[1]
maf <- example$maf
snploc <- example$snploc
sum((maf>0)*(maf<0.05))
```

```
## [1] 302737
```

Next we generated phenotype data. We first generated two covariates, one is continous, the other one is binary.

```
set.seed(937)
X0 <- rep(1,samplesize)
X1 <- rnorm(samplesize,0,1)
X2 <- rbinom(samplesize,1,1/2)
epi <- rnorm(samplesize,0,1)
```

We randomly selected two signal regions (genotype-phenotype association regions) across the 10Mb region. The number of variants in the signal region were ramdonly selected between 50 and 80.

```
## number of signal regions
n0 <- 2

## generate signal region location
```

```
pp <- floor(dim(genotype)[2]/n0)

## location od signal region
# In snese of variants order
sigloc <- matrix(rep(0,3*n0),ncol=3)
for(i in 1:n0)
{
    begnum <- (i-1)*pp + 1
    endnum <- i*pp - 1000
    sigloc[i,1] <- sample(begnum:endnum,1)
    sigloc[i,3] <-  sample(50:80,1)
    sigloc[i,2] <- sigloc[i,1]  +  sigloc[i,3] - 1
}
```

Then the location of the two signal regions in sense of variants order were as follows (last column is the number of variants in the signal region):

```
sigloc
```

```
##        [,1]   [,2] [,3]
## [1,]  12279  12354   76
## [2,] 176079 176146   68
```

Let p0 be the number of variants in the signal regions. We set the sparsity index xi=2/3 and randomly select p0^2/3 variants in the signal region to be causal.

```
percen <-  2/3

sigploc <- matrix(rep(0,60*n0),ncol=60)
sigloctemp <- c()

for(ii in 1:n0)
{
    mafid <- (sigloc[ii,1]+1):(sigloc[ii,2]-1)
    p0 <- ceiling(sigloc[ii,3]^percen)

    sigploc[ii,1:p0] <- c(sigloc[ii,1],sort(sample(mafid,p0-2,replace=FALSE)),sigloc[ii,2])
}

sigloc <- cbind(sigloc,sigploc)
```

We set the effect sizes as a decreasing function of MAFs, beta=0.185*log10(MAF). The sign of coefficients were randomly and independently set as 50% positive and 50% negative. The phenotype was generated through linear model.

```
# signal strength
c0 <- 0.185
# direction of the signals
protect <- 0.5

phenotype <- 0.5*X1 + 0.5*X2
```

```
for(ii in 1:n0)
{
    sigloctemp <- sigloc[ii,4:(dim(sigloc)[2]-1)]
    sigloctemp <- sigloctemp[sigloctemp>0]
    # beta
    beta <- c0*log(maf[sigloctemp])/log(10)

    betadir <- rbinom(length(beta),1,protect)*2-1
    beta <- beta*betadir
    phenotype <- phenotype + genotype[,sigloctemp]%*%beta
    phenotype <- as.vector(phenotype)
}

phenotype <- phenotype + epi
X <- cbind(X1,X2)
```

## Analyzing the simulated data using Q-SCAN

We set the parameter of Q-SCAN procedure. We set the smallest number of variants of searching windows as 40 and the maximum number of searching windows as 200. In this example, we consider the continuous trait and set the regression type as “gaussian”.

```
Lmax <- 200
Lmin <- 40
family <- "gaussian"
```

We then applied Q-SCAN procedure to analyze the simulated data set. It takes around 20 mins to analyze 10Mb sequence on 10,000 individuals.

```
# Q_SCAN_40_200 <- Q_SCAN(genotype,phenotype,X,family,Lmax,Lmin)
```

Q-SCAN detected 2 significant signal regions, which were listed as follows:

```
Q_SCAN_40_200$SCAN_res
```

```
##          [,1]   [,2]   [,3]  [,4]
## [1,] 31.07485  12261  12300 0.000
## [2,] 13.75784 176079 176121 0.012
```

Recall the true signal regions:

```
sigloc[,1:3]
```

```
##        [,1]   [,2] [,3]
## [1,]  12279  12354   76
## [2,] 176079 176146   68
```

The two detected region of Q-SCAN was overlapped with the two true signal region, respectively.

We also applied M-SCAN to analyze the same data.

```
# M_SCAN_40_200 <- M_SCAN(genotype,phenotype,X,family,Lmax,Lmin)
```

M-SCAN only detected 1 significant signal regions, which was listed as follows:

```
M_SCAN_40_200$SCAN_res
```

```
##          [,1]  [,2]  [,3] [,4]
## [1,] 39.79448 12284 12323    0
```