iGREX: Quantifying the impact of genetically regulated expression on complex traits and diseases

CAI Mingxuan

2018/10/12

Introduction

This vignette provides an introduction to the `iGREX` package. R package `iGREX` implements iGREX, a statistical modelfor quantifying the impact of genetically regulated expression on complex traits and diseases. The package can be installed with the command:

```
library(devtools)
```

```
install_github("mxcai/iGREX")
```

The package can be loaded with the command:

```
library("iGREX")
```

Fit CoMM using simulated data

We first demonstrate the workflow of iGREX using simulated data. The genotype matrix can be simulated using *genRawGeno* function:

```
library(mvtnorm)
```

```
## Warning: package 'mvtnorm' was built under R version 3.4.4
```

```
set.seed(10)
n1 <- 1000
n2 <- 4000
p <- 100   #number of SNPs in each gene
G <- 200   #number of genes
maf <- runif(p*G, 0.05, 0.5)
rho <- 0.8
X <- genRawGeno(maf,p,G,rho,n1+n2)
X1 <- X[1:n1,]
X2 <- X[(n1+1):(n1+n2),]
```

Then, cis effect sizes are generated from standard Gaussian distribution with sparse structure. We set only 20% of cis effects as non-zero, and control the cellular heritability level at $h_y^2 = 0.3$.

```
sb2_true <- 0.3
sy2_true <- 0.7
Y0 <- matrix(0,n1+n2,G)
for(g in 1:G){
  nonzero <- rbinom(p,1,0.2)
  if(sum(nonzero)==0) nonzero[1] <- 1
  beta <- rnorm(p,0,sqrt(sb2_true/sum(nonzero)))
  beta[nonzero==0] <- 0
  Y0[,g] <- X[,((g-1)*p+1):(g*p)] %*% beta
}
Y <- Y0[1:n1,] + matrix(rnorm(n1*G,0,sqrt(sy2_true)),n1,G)
```

Finally, the phenotype data is generated as the generative model of iGREX with variance components specified as $h^2_{GREX} = 0.2$, $h^2_{Alternative} = 0.3$. GREX effects are also sparse with only 20% non-zero:

```
sg2_true <- 0.2
sa2_true <- 0.3
t <- mean(diag(Y0[(n1+1):(n1+n2),]%*%t(Y0[(n1+1):(n1+n2),])))
nonzero <- rbinom(G,1,0.2)
if(sum(nonzero)==0) nonzero[1] <- 1
alpha <- as.matrix(rnorm(G,0,sqrt(sg2_true/t*G/sum(nonzero))))
alpha[nonzero==0] <- 0
z0 <- Y0[(n1+1):(n1+n2),] %*% alpha
gamma <- as.matrix(rnorm(p*G,0,sqrt(sa2_true/(p*G))))
z1 <- X2%*%gamma
z <- z0 + z1 + rnorm(n2,0,sqrt(var(z0+z1)))
```

Stage one iGREX solves G linear mixed models using PX-EM algorithm and calculate K_{GREX} and $K_{Alternative}$.

```
X <- scale(X)
X1 <- X[1:n1,]
X2 <- X[(n1+1):(n1+n2),]
K <- K0 <- 0
for(g in 1: G){
  y_g <- Y[,g]
  X1tmp <- X1[,((g-1)*p+1):(g*p)]
  X2tmp <- X2[,((g-1)*p+1):(g*p)]
  W1 <- matrix(1,n1,1)
  fit_g <- iGREX_Kg(y_g,X1tmp,X2tmp,W1,1e-5,500)
  K <- K + fit_g$K_g
  K0 <- K0 + fit_g$K_g0
}
Ka <- X2 %*% t(X2) / ncol(X2)
K <- K/mean(diag(K))
K0 <- K0/mean(diag(K0))
```

Stage two iGREX fit variance components model using either REML or MoM:

```
  # REML
  REML <- REML_3var(K,Ka,z,verbose=F)

  # REML without acocunting for uncertainty
  REML0 <- REML_3var(K0,Ka,z,verbose=F)

  # exact estimate by MoM
  MoM <- MoM_3var(K,Ka,z)

  # exact estimate by MoM without accounting for uncertainty
  MoM0 <- MoM_3var(K0,Ka,z)
```

The $P\hat{V}E_{iGREX}$ and $P\hat{V}E_{Alternative}$ as well as their standard errors are returned:

```
REML$H[,1:2]
```

```
##                      h_R        h_D
## heritability 0.19839126 0.31953878
## se           0.02414782 0.03229657
```

```
REML0$H[,1:2]
```

```
##                      h_R        h_D
## heritability 0.13724202 0.35646760
## se           0.01684552 0.03103538
```

```
MoM$H[,1:2]
```

```
##                      h_R        h_D
## heritability 0.22309468 0.30862957
## se           0.03919848 0.05517908
```

```
MoM0$H[,1:2]
```

```
##                      h_R        h_D
## heritability 0.15451975 0.34415728
## se           0.02715762 0.05277561
```

Fit iGREX using GWAS and eQTL data

We provide an example of fitting iGREX using GWAS and eQTL data in plink binary format. We first specify the input file names:

```
file_qtlgeno <-  "GTEx_qc_hm3"
file_GWASgeno <- "NFBC_filter_mph10"
file_expression <- "Liver_gene_expression.txt"
file_qtlcov <- "Liver_cov.txt"
file_GWAScov <- "NFBC_hm3_filter.eigenvec"
load("K_NFBC_hm3.RData")
```

Here, file_qtlgeno is the prefix for eQTL genotype data in plink binary format, file_GWASgeno is the GWAS data in plink binary format, file_expression is the gene expression file with extended name, file_qtlcov and file_GWAScov are covariates file for eQTL and GWAS data, respectively. For gene expresion file, it must have the following format (rows for genes and columns for individuais and note that it must be tab delimited):

lower	up	genetype1	genetype2	TargetID	Chr	HG00105	HG00115
59783540	59843484	lincRNA	PART1	ENSG00000152931.6	5	0.5126086	0.7089508
48128225	48148330	protein_coding	UPP1	ENSG00000183696.9	7	1.4118007	-0.0135644
57846106	57853063	protein_coding	INHBE	ENSG00000139269.2	12	0.5755268	-1.0162217
116054583	116164515	protein_coding	AFAP1L2	ENSG00000169129.8	10	1.1117776	0.0407033
22157909	22396763	protein_coding	RAPGEF5	ENSG00000136237.12	7	0.2831573	-0.1772559
11700964	11743303	lincRNA	RP11-434C1.1	ENSG00000247157.2	12	0.2550282	-0.2831573

The covariate files must have the following format (rows for individuals and the first two columns are FID and IID):

V1	V2	V3	V4	V5	V6	V7	V8
GTEX-11DXY	GTEX-11DXY	0.0109	-0.0027	-0.0789	-0.0552724	0.0469389	0.0488902
GTEX-11DXZ	GTEX-11DXZ	-0.1053	-0.0126	0.0027	-0.0097383	0.0542790	-0.0918273
GTEX-11EQ9	GTEX-11EQ9	0.0148	-0.0078	-0.0057	0.0362291	0.0474735	-0.0502703
GTEX-11GSP	GTEX-11GSP	0.0168	-0.0052	-0.0019	-0.0505170	-0.0960899	-0.0554980
GTEX-11NUK	GTEX-11NUK	0.0154	-0.0025	0.0049	-0.0733902	0.0009193	0.0227932
GTEX-11NV4	GTEX-11NV4	-0.1152	-0.0170	-0.0223	-0.0619178	0.0035294	0.0361403

The alternative genetic correlation matrix (`Ka`) is just the kinship matrix and should be manually computed (which can be done by plink or GEMMA) and supplied to the function. Finally, the analysis can be done by (`iGREX`) function:

```
fit <- iGREX(file_qtlgeno,file_GWASgeno,file_expression,file_qtlcov,file_GWAScov,Ka=K,
             whCol=6,bw=500000,subsample = 0,method="MoM")
```

The estimated PVEs are stored in the outcome list and can be accessed by (`fit$fit_iGREX$H`)

	h_R	h_D	h_all	h_med
heritability	0.1917895	0.4069149	0.5987045	0.3203409
se	0.0408129	0.1001861	0.0937858	0.0823769