# Efficient and Effective Control of Confounding in eQTL Mapping Studies through Joint Differential Expression and Mendelian Randomization Analyses

**Yue Fan and Xiang Zhou**

**2020-06-02**

## Introduction

This vignette provides an introduction to the ECCO package. R package ECCO implements ECCO, an alternative way to determine the optimal number of PEER factors used for eQTL mapping studies. The package can be installed with the following commands:

```
library(devtools)
install_github("fanyue322/ECCO")
```

Load the package using the following command:

```
library(ECCO)
```

## Calculate the gene expression residuals with PEER pacakge

Input: the gene expression data dt, an N*P matrix; peer, the number of PEER factors to be remove

| Geneid | Gene1 | Gene2 | .. |
|---|---|---|---|
| indiv1 | 4.91 | 4.63 | .. |
| indiv2 | 13.78 | 13.14 | .. |
| .. | .. | .. | .. |

```
library(peer)  ## library the PEER package
pc=10          ## Set the number of PEER factors
model = PEER()
  PEER_setPhenoMean(model,as.matrix(dt))
  dim(PEER_getPhenoMean(model))
  PEER_setAdd_mean(model, TRUE)
  PEER_setNk(model,peer)
  PEER_getNk(model)
  PEER_update(model)
  factors = PEER_getX(model)
  factors=factors[,-1]
  residuals = PEER_getResiduals(model)
```

```
  write.table(residuals, paste(tissue,'_peer', pc, ".txt", sep=""), quote=F, col.names=F,
row.names=F)
```

Output: the gene expression residuals, an N*P matrix

# Fit ECCO using simulated data

```
data(exampledata)
attach(exampledata)
ind=1
genename=gene_name[ind]
gene=M_matrix[,ind]
geno=snp_raw[[ind]]
```

### Select the instrumental variable for each gene with ecco0

gene expression data format:

| Geneid | Gene1 | Gene2 | .. |
|---|---|---|---|
| indiv1 | 4.91 | 4.63 | .. |
| indiv2 | 13.78 | 13.14 | .. |
| .. | .. | .. | .. |

genotype data format:

| snp id | indiv1 | indiv2 | .. |
|---|---|---|---|
| snp1 | 1 | 0 | .. |
| snp2 | 0 | 2 | .. |
| .. | .. | .. | .. |

```
iv_snp=c()
  for (ind in 1:P) {
    tryCatch({
      gene <- M_matrix[ind,]
      #geno is a P*N matrix containing all the cis-SNPs for the ind th gene
      ivsnp=ecco0(gene,genename,gene_name,geno,ind)
      iv_snp=rbind(iv_snp,ivsnp)
          },
    error=function(e){})
    print (ind)
  }
  save(iv_snp,file=paste0("./cissnp/",tissue,"/",chr,".RData"))
#Output: a matrix containing the cis-SNPs for all P genes
```

### Estimate $\overline{\beta}$, $\tilde{\beta}$ and p-values for $\overline{\beta}$.

```
peerlist=c(0,1,2,5,10,15,20,30,40,50,60,70,80,90,100)
for(num_peer in 1:length(peerlist))
```

```
{
tryCatch({
summary<-ecco(pheno,peer[[num_peer]],gene_name,iv_snp,peerlist[num_peer])
},
error=function(e){})
summary_total=rbind(summary_total,summary)
res=rbind(res,c(cor(as.numeric(summary[,4]),as.numeric(summary[,5])),peerlist[num_peer]))
}
res=data.frame(res)
```

output format for ecco:

| Gene | PEER | p-value | beta_hat | beta_tilde |
|---|---|---|---|---|
| Gene1 | 1 | .. | .. | .. |
| Gene2 | 1 | .. | .. | .. |
| Gene3 | 1 | .. | .. | .. |
| .. | .. | .. | .. | .. |

Until now, we obtain the effect sizes:$\overline{\beta}$ and $\tilde{\beta}$,

### Determine the optimal number of PEER factors

```
optimal_num_peer=res[which(res[,1]==max(res[,1])),2]
```

…

# Example

### A toy example for testing purposes only:

```
data(exampledata)
attach(exampledata)
N=length(gene_name)
iv_snp=c()
for(ind in 1:N)
{
tryCatch({
gene=M_matrix[,ind]
geno=snp_raw[[ind]]
genename=gene_name[ind]
ivsnp=ecco0(gene,genename,gene_name,geno,ind)
iv_snp=rbind(iv_snp,ivsnp)
},
error=function(e){})
}
res=c()
for(num_peer in 1:length(peerlist))
{
tryCatch({
pheno=Y
gene=M_matrix
```

```
geno=snp_raw
gene_name=gene_name
peerlist=c(1,2,5)
summary<-ecco(pheno,peer[[num_peer]],gene_name,iv_snp,peerlist[num_peer])
},
error=function(e){})
summary_total=rbind(summary_total,summary)
res=rbind(res,c(cor(as.numeric(summary[,4]),as.numeric(summary[,5])),peerlist[num_peer]))
}
res=data.frame(res)
optimal_num_peer=res[which(res[,1]==max(res[,1])),2]
```