Tutorial on Epistasis using MB-MDR software

Fentaw Abegaz and Kristel Van Steen

University of Liege
GIGA - BIO3

November 9, 2016

Outline

1. HapMap Data
2. Selection of SNPs: LD pruning
3. Epistasis using MB-MDR
4. Mapping SNPs to Genes: UCSC Browser
5. Dynamic gene network

Data

A subset of data from the HapMap project

- Population: CEU, HCB, JPT and YRI
- Number of SNPs: 9088
- Number of samples: 279
- Covariate: sex

```
> source("http://bioconductor.org/biocLite.R")
  > biocLite("gdsfmt")
  > biocLite("SNPRelate")
> library(SNPRelate)
```

HapMap data

```
 > library(SNPRelate)
 > genofile <- snpgdsOpen(snpgdsExampleFileName())
 > genofile
 > pop.group <- read.gdsn(index.gdsn(genofile,
     path="sample.annot/pop.group"))
> table(pop.group)
> #phenotype <- c(rep(0,93), rep(1,92)) #population substru
> phenotype <- c(rep(0,47), rep(1,47))
CEU HCB JPT YRI
92  47  47  93
 head(pop.group)
[1] "YRI" "YRI" "YRI" "YRI" "CEU" "CEU"
```

HapMap data

Extract genotype data:

```
> sex <- read.gdsn(index.gdsn(genofile,
    path="sample.annot/sex"))
> table(sex)
```

HapMap data

Covariate:

```
> genotype <- read.gdsn(index.gdsn(genofile, "genotype"))
> colnames(genotype) <-c(read.gdsn(index.gdsn(genofile,
                                "snp.rs.id")))
> head(genotype)
```

LD-based SNP pruning

To select SNPs which are not highly correlated to each other.

```
> set.seed(1000)
#LD
> LDcomp <- snpgdsLDMat(genofile, method="corr")
> LDmatrix <- LDcomp$LD
> LDmatrix[1:10, 1:10]
> image(LDmatrix)
 # LD pruning
> snpset <- snpgdsLDpruning(genofile, method="corr",
            ld.threshold=0.2)
> names(snpset)
> head(snpset$chr1)
> snpset.id <- unlist(snpset)
> snpset <- as.vector(snpset.id[1:length(snpset.id)])
```

Pruned SNP set

```
> genotype <- read.gdsn(index.gdsn(genofile, "genotype"))
> colnames(genotype) <-c(read.gdsn(index.gdsn(genofile,
                                "snp.rs.id")))
> pruned.genotype <- subset(genotype, select = snpset )
> dim(pruned.genotype)
> head(pruned.genotype)
> gen.data <- data.frame(pop.group, pruned.genotype)
#HCB and JPT populations
> popustrat <- rbind(gen.data[pop.group=="HCB",],
                     gen.data[pop.group=="JPT",])
> popustrat$pop.group <- phenotype
> dim(popustrat)
> popustrat[1:2,1:3]
```

Penalized regression and network

- glmnet
- permutation test

Large scale epistasis with MB-MDR

- MB-MDR is a method for identifying multi-locus genotypes and their interactions that are associated with a phenotype of interest, and allows to adjust for marginal and confounding effects.
- Software
  - MB-MDR C++ version developed by Kristel Lab (computationally efficient for the analysis of thousands of SNPs and their interactions)
  - mbmdr R package

- Install the **mbmdr** package from CRAN.
  - Menu based : click on

    packages → install packages) → select region → select package **mbmdr**
  - Command based:

```
install.packages("mbmdr",
                 repos= "http://cran.us.r-project.org",
                 dependencies = TRUE)
```

- Loading the package **mbmdr**

```
library(mbmdr)
```

- **MB-MDR C++ software**

```
  -bash-4.1$ ./mbmdr-4.4.1.out -- binary -d 2D
-a CODOMINANT -mt gammaMAXT
-o ./output_filename ./input_filename.txt
```

Description

- Phenotype: binary, continuous, survival
- d: order of interaction
- mt: multiple testing correction algorithm: NONE, MAXT, MINP, RAWP, STRAT1, STRAT2 or gammaMAXT (default)
- a: adjusting for main effects and covariates: CODOMINANT (default), ADDITIVE, ONESTEP or NONE
- Parrallel work flow for analysing big datasets using gammaMAXT.

- **mbmdr** R package

```
mbmdr(y,
  data,
  order,
  covar = NULL,
  exclude = NA,
  risk.threshold = 0.1,
  output = NULL,
  adjust = c("none", "covariates", "main effects","both
  first.model = NULL,
  list.models = NULL,
  use.logistf = TRUE,
  printStep1 = FALSE, ...)
```

mbmdr input

- y : a vector of response values
- data: a data frame containing SNP values: 0, 1, or 2
- order: interaction order (eg. order=2)
- exclude: excluding missing Values (exclude=NA)
- risk.threshold: a threshold defining the risk category (risk.threshold=0.1)
- adjust: type of regression adjustment (eg. adjust="'covariates"')

mbmdr output

- SNP1...SNPx Names of snps in interaction.
- NH: Number of significant High risk genotypes in the interaction.
- betaH: Regresion coeficient for High risk exposition.
- WH: Wald statistic for High risk category.
- PH: P-value of the Wald test for the High risk category.
- NL: Number of significant Low risk genotypes in the interaction.
- betaL: Regresion coeficient in for Low risk exposition.
- WL: Wald statistic for Low risk category.
- PL: P-value of the Wald test for the Low risk category.
- MIN.P: Minimun p-value (min(PH,PL)) for the interaction model.

- Permutation test for correcting for multiple testing

```
mbmdr.PermTest(x, n, model = NULL, sig.level=1)
```

where

- x: an mbmdr object returned by mbmdr function.
- n: Number of permutations.

mbmdr model fiting

Objective: to investigate the effect of epistasis (interactions between genes or SNPs) on population substructure using mbmdr package.

```
> mbmdr.fit <- mbmdr(y=popustrat$pop.group,
      data=popustrat[,2:16], order=2,
            family=binomial(link=logit))
> mbmdr.fit$result
```

```
> print(mbmdr.fit)
       SNP1        SNP2 NH  betaH    WH       PH NL   betaL    WL       PL    MIN.P
 rs10864363   rs836755  1 0.9634 4.081 0.04338  1 -1.3125 7.729 0.005433 0.005433
   rs836755 rs12132314  0     NA    NA       NA  1 -2.0857 3.657 0.055842 0.055842
  rs7518506   rs583027  1 2.8675 3.310 0.06887  0      NA    NA       NA 0.068873
   rs836755  rs3766962  0     NA    NA       NA  1 -0.8314 3.184 0.074341 0.074341
   rs836755   rs583027  1 0.7703 2.983 0.08415  0      NA    NA       NA 0.084150
  rs7518506  rs1695824  1 2.7000 2.849 0.09141  0      NA    NA       NA 0.091413
   rs836755  rs1064721  0     NA    NA       NA  1 -0.6958 2.726 0.098729 0.098729
```

Permutation test

```
> models <- subset(mbmdr.fit$result, MIN.P <= 1.0, select = 1:order)
> mbmdr.PermTest(mbmdr.fit, 100, models)
```

SNP1	SNP2	NH	betaH	WH	NL	betaL	WL	Wmax	Perm.P
rs10864363	rs836755	1	0.9634	4.081	1	-1.3125	7.729	7.729	0.01
rs7518506	rs583027	1	2.8675	3.310	0	NA	NA	3.310	0.01
rs7518506	rs1695824	1	2.7000	2.849	0	NA	NA	2.849	0.02
rs836755	rs12132314	0	NA	NA	1	-2.0857	3.657	3.657	0.11
rs836755	rs583027	1	0.7703	2.983	0	NA	NA	2.983	0.2
rs836755	rs1064721	0	NA	NA	1	-0.6958	2.726	2.726	0.28
rs836755	rs3766962	0	NA	NA	1	-0.8314	3.184	3.184	0.33

Adjusting for Covariate(s)

```
      da <- popustrat[,2:16]
      rownames(da) <- NULL
      mbmdr.adj <- mbmdr(y=popustrat$pop.group,data=da,
             order=2, covar=popustrat$sex,
 adjust="covariates",
                    family=binomial(link=logit))
      mbmdr.adj$result

#Permutation test
     order <- 2
     models <- subset(mbmdr.adj$result, MIN.P <= 1.0,
                  select = 1:order)
     perm_adj <- mbmdr.PermTest(mbmdr.adj, 100, models)
                  #It takes sometime.
     perm_adj
```

Mapping SNPs to Genes

Mapping SNPs to the corresponding genes allows to have a better understanding and interpretation of the SNPs and their interactions.

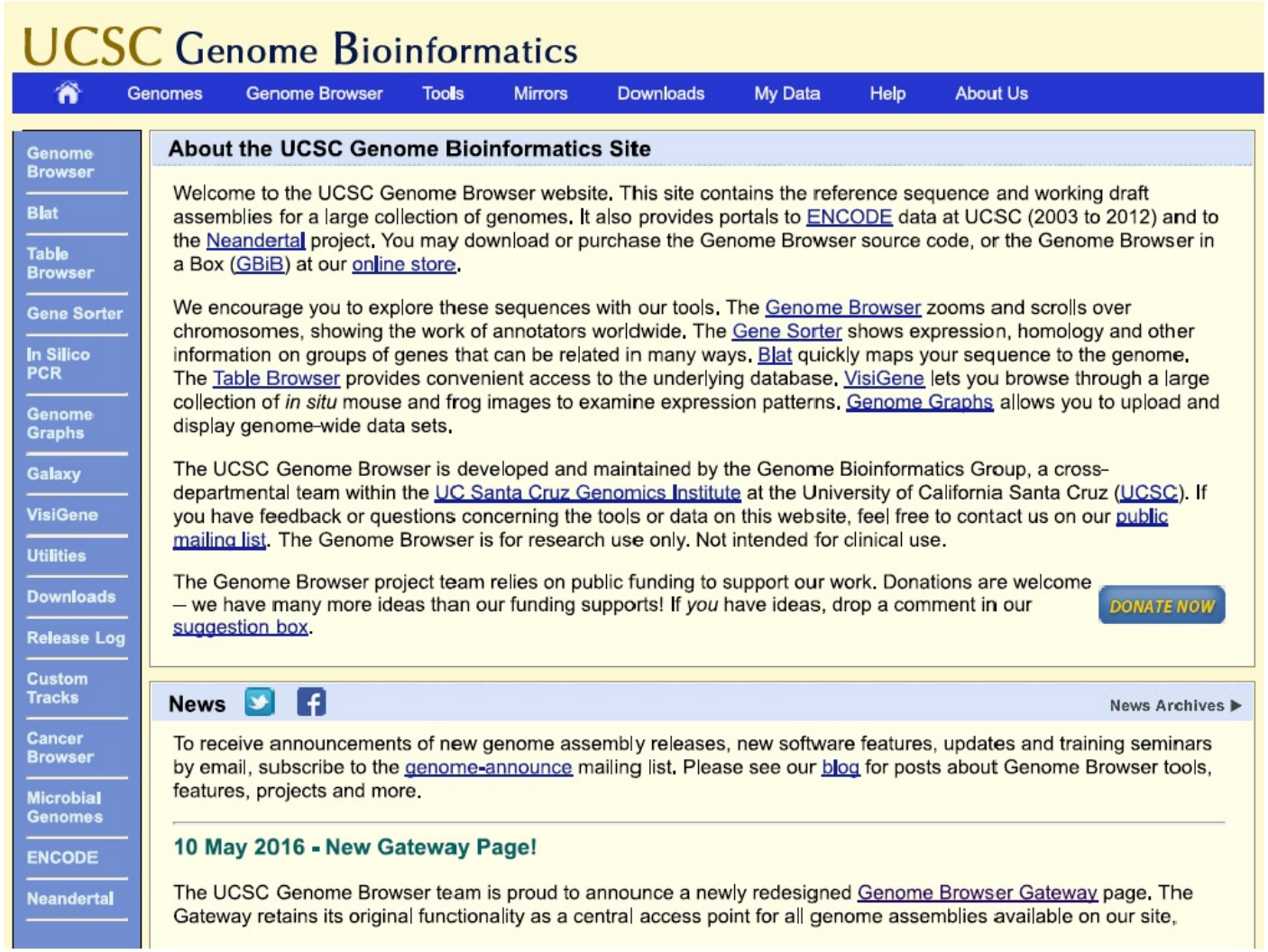
Use UCSC Genome browser: Open link
https://genome.ucsc.edu

Go to **Tools** $\rightarrow$ **Variant Annotation Integrator**

Mapping the selected SNPs using UCSC genome browser:

rs10864363 $\longrightarrow$ PER3 gene

rs836755 $\longrightarrow$ RERE gene


UCSC Genome Bioinformatics
Genomes
Genome Browser
Tools
Mirrors
Downloads
My Data
Help
About Us
Genome Browser
Blat
Table Browser
Gene Sorter
In Silico PCR
Genome Graphs
Galaxy
VisiGene
Utilities
Downloads
Release Log
Custom Tracks
Cancer Browser
Microbial Genomes
ENCODE
Neandertal
About the UCSC Genome Bioinformatics Site
Welcome to the UCSC Genome Browser website. This site contains the reference sequence and working draft assemblies for a large collection of genomes. It also provides portals to ENCODE data at UCSC (2003 to 2012) and to the Neandertal project. You may download or purchase the Genome Browser source code, or the Genome Browser in a Box (GBiB) at our online store.
We encourage you to explore these sequences with our tools. The Genome Browser zooms and scrolls over chromosomes, showing the work of annotators worldwide. The Gene Sorter shows expression, homology and other information on groups of genes that can be related in many ways. Blat quickly maps your sequence to the genome. The Table Browser provides convenient access to the underlying database. VisiGene lets you browse through a large collection of in situ mouse and frog images to examine expression patterns. Genome Graphs allows you to upload and display genome-wide data sets.
The UCSC Genome Browser is developed and maintained by the Genome Bioinformatics Group, a cross-departmental team within the UC Santa Cruz Genomics Institute at the University of California Santa Cruz (UCSC). If you have feedback or questions concerning the tools or data on this website, feel free to contact us on our public mailing list. The Genome Browser is for research use only. Not intended for clinical use.
The Genome Browser project team relies on public funding to support our work. Donations are welcome – we have many more ideas than our funding supports! If you have ideas, drop a comment in our suggestion box.
DONATE NOW
News
News Archives
To receive announcements of new genome assembly releases, new software features, updates and training seminars by email, subscribe to the genome-announce mailing list. Please see our blog for posts about Genome Browser tools, features, projects and more.
10 May 2016 - New Gateway Page!
The UCSC Genome Browser team is proud to announce a newly redesigned Genome Browser Gateway page. The Gateway retains its original functionality as a central access point for all genome assemblies available on our site,

Genomes Genome Browser Tools Mirrors Downloads My Data Help About Us

UCSC Genome Bioinformatics

Annotated Variants in VEP/HTML format

Variants: Variant Identifiers
Transcripts: GENCODE v22 Comprehensive Transcript Set (only Basic displayed by default) (hg38.knownGene)
dbSNP: Simple Nucleotide Polymorphisms (dbSNP 146) (hg38.snp146)
Keys for Extra column items:
VEST: Variant Effect Scoring Tool (VEST) (scores [0–1] predict confidence that a change is deleterious
SIFT: SIFT (D = damaging, T = tolerated)
PP2HVAR: PolyPhen-2 with HumVar training set (D = probably damaging, P = possibly damaging, B = benign)
PP2HDIV: PolyPhen-2 with HumDiv training set (D = probably damaging, P = possibly damaging, B = benign)
MUTTASTER: MutationTaster (A = disease causing automatic, D = disease causing, N = polymorphism, P = polymorphism automatic)
MUTASSESSOR: MutationAssessor (high or medium: predicted functional; low or neutral: predicted non-functional)
LRT: Likelihood ratio test (LRT) (D = deleterious, N = Neutral, U = unknown)
INTERPRO: InterPro protein domains
GERPRS: GERP++ Rejected Substitutions (RS)
GERPNR: GERP++ Neutral Rate (NR)

Uploaded Variation	Location	Allele	Gene	Feature	Feature type	Consequence	Position in cDNA	Position in CDS	Position in protein	Amino acid change	Codon change	Co-located Variation	Extra
rs836755	chr1:7786467	C	PER3	uc001aon.3	Transcript	intron_variant	-	-	-	-	-	rs836755	INTRON=3/9
rs836755	chr1:7786467	C	PER3	uc001aop.5	Transcript	intron_variant	-	-	-	-	-	rs836755	INTRON=3/21
rs836755	chr1:7786467	C	PER3	uc057bxq.1	Transcript	intron_variant	-	-	-	-	-	rs836755	INTRON=3/22
rs836755	chr1:7786467	C	PER3	uc057bxr.1	Transcript	intron_variant	-	-	-	-	-	rs836755	INTRON=2/20
rs836755	chr1:7786467	C	PER3	uc057bxs.1	Transcript	intron_variant	-	-	-	-	-	rs836755	INTRON=2/20
rs836755	chr1:7786467	C	PER3	uc057bxt.1	Transcript	intron_variant	-	-	-	-	-	rs836755	INTRON=1/3
rs836755	chr1:7786467	C	PER3	uc057bxu.1	Transcript	upstream_gene_variant	-	-	-	-	-	rs836755	DISTANCE=855
rs10864363	chr1:8707252	G	RERE	uc001ape.4	Transcript	intron_variant	-	-	-	-	-	rs10864363	INTRON=2/23
rs10864363	chr1:8707252	G	RERE	uc001apf.4	Transcript	intron_variant	-	-	-	-	-	rs10864363	INTRON=1/22
rs10864363	chr1:8707252	G	RERE	uc001aph.2	Transcript	intron_variant	-	-	-	-	-	rs10864363	INTRON=1/10
rs10864363	chr1:8707252	G	RERE	uc057bzn.1	Transcript	upstream_gene_variant	-	-	-	-	-	rs10864363	DISTANCE=4033

Dynamic gene network

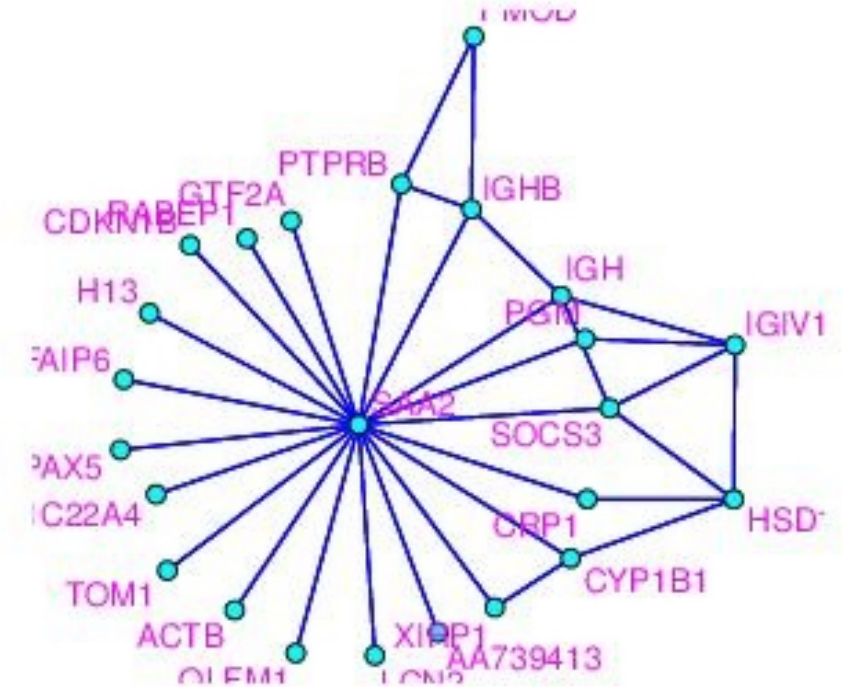
Gene network reconstruction from time course data

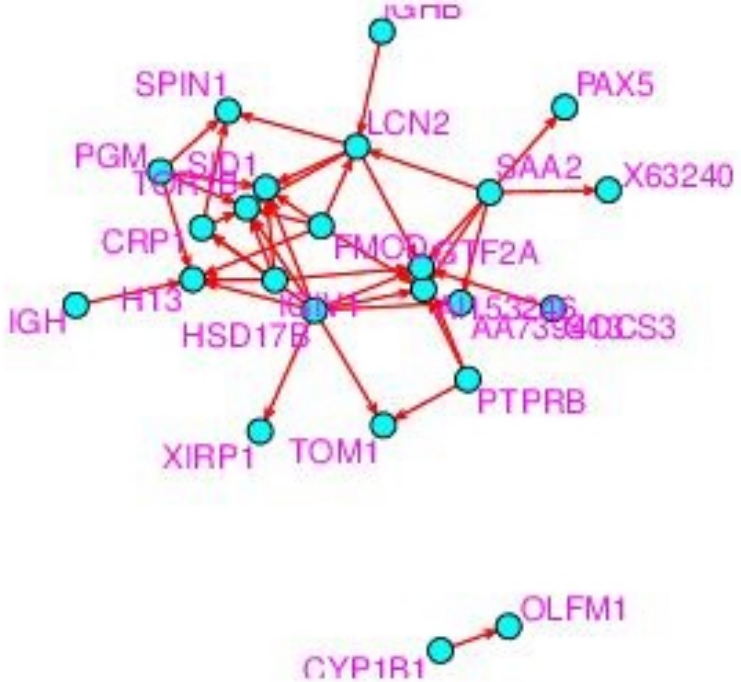
```
> library(SparseTSCGM)
> library(longitudinal)
> data(mammary)
> mammary
Longitudinal data:
 30 variables measured at 18 different time points
 Total number of measurements per variable: 54
 Repeated measurements: yes
 To obtain the measurement design call 'get.time.repeats()
```

Longitudinal data format

	SID1	S1C22A4	CDKN1B	RABEP1	TOM1
1-1	6.337710	3.830813	2.442347	3.655840	4.714025
1-2	6.143542	3.881564	2.351375	4.069027	4.112512
1-3	6.190725	3.935740	2.602690	3.594569	4.144721
2-1	6.154433	3.456317	2.261763	3.543854	4.578826
2-2	6.462249	3.113515	2.517696	3.706228	5.225747
2-3	6.097400	3.015535	2.219203	3.152736	4.631812
:					
.					
18-1	6.434707	3.261935	2.541602	3.901973	5.101694
18-2	6.407705	3.321432	2.468100	3.837299	4.955123
18-3	6.131009	3.795489	2.667228	3.600048	4.374498

```
 > mammary.sc <- dyn.scale(mammary, center=TRUE, scale=FALS
 > lam1 <- seq(0.9, 0.3, -0.10)
 > lam2 <- seq(0.9, 0.1, -0.15)
 > mammary.tscgm <- sparse.tscgm(data=mammary.sc,
         lam1=lam1, lam2=lam2, nlambda=NULL,
 model="ar1", optimality="bic",
         control=list(maxit.out = 10, maxit.in = 100))
#Graphical visualization
> plot.tscgm(mammary.tscgm, mat="precision",
        main="Undirected gene network",
        pad=0.01, label.pad=0.3, label.col=6,
vertex.col=5, vertex.cex=1.5, edge.col=4)
> plot.tscgm(mammary.tscgm, mat="autoregression",
        main="Directed gene network",
        pad=0.01, label.pad=0.01, label.col=6,
vertex.col=5, vertex.cex=2, edge.col=2)
```


Undirected gene network
PTPRB
IGHB
GTF2A
CDKN1B
H13
IGH
IGIV1
SOCS3
HSD
CYP1B1
TOM1
ACTB
AA739413


Directed gene network
SPIN1
PAX5
LCN2
PGM
SAA2
X63240
CRP
TF2A
IGH
H13
HSD17B
PTPRB
XIRP1
TOM1
OLFM1

Practical Exercise

1. Consider the data frame "'gen.data"' to investigate epistasis on population substructure (0=YRI, 1=CEU). Only using the first 10 SNPs:

   a. identify significant 2-way interactions.
   b. interpret your results at gene level.
   c. Correct for main (lower order) effects.
   d. Correct for the effect of sex.

2. Using the mammary data and time lag 2, construct time series chain networks.