Removing batch effects in gene expression array study

Chung-Hsing Chen

National Cancer Institute, National Health Research Institutes

Data background (1)

- Hedenfalk et al. measured **3226** genes in seven BRCA1 and eight BRCA2 mutation-positive tumor samples.
- The goal of the study was to identify genes that showed differential expression across breast cancer tumor subtypes defined by these germline mutations.
- Several genes with apparent outliers were removed. This left **3170** genes.

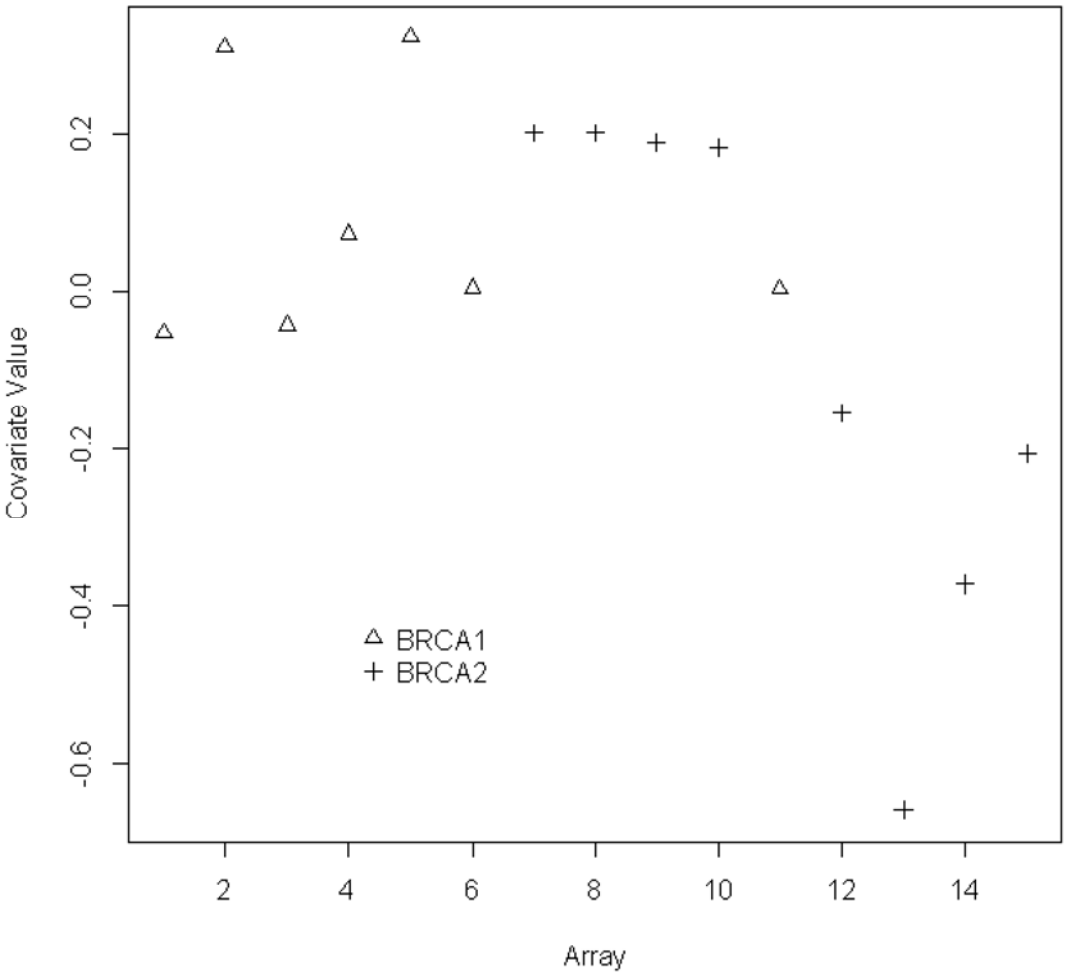
Data background (2)

```
> SVA.pheno
    NEJM-PatientID Mutation
V7               1    BRCA1
V8               5    BRCA1
V9               3    BRCA1
V10              7    BRCA1
V11              2    BRCA1
V12              4    BRCA1
V13             10    BRCA2
V14              9    BRCA2
V15              8    BRCA2
V16             10    BRCA2
V17              6    BRCA1
V18             13    BRCA2
V19             14    BRCA2
V20             11    BRCA2
V21             12    BRCA2
> head(SVA.exp.preprocessed)
      q-value    p-value fold-change (log base 2) s1996 s1822 s1714 s1224 s1252 s1510 s1900 s1787 s1721 s1486 s1905
[1,] 0.089529 0.01223344                   1.203  0.15  0.22  0.30  0.26  1.22  0.44  0.35  1.10  1.07  1.46  0.38
[2,] 0.213752 0.07611987                  -0.521  1.54  1.27  0.76  0.85  1.27  0.64  0.90  0.64  0.78  0.55  0.61
[3,] 0.672438 0.99530284                   0.002  1.72  1.57  2.13  1.09  1.98  0.74  1.71  1.16  1.33  1.46  2.43
[4,] 0.163987 0.04212934                   0.690  0.71  1.24  1.69  2.23  1.16  0.82  1.44  2.03  3.60  1.20  2.08
[5,] 0.637670 0.84745741                   0.053  0.94  1.53  1.87  1.19  1.16  1.54  1.05  0.91  0.85  1.22  1.01
[6,] 0.377451 0.25437224                  -0.227  0.80  0.95  1.53  1.37  1.02  1.22  0.78  0.96  0.65  1.02  1.09
     s1816 s1616 s1063 s1936
[1,]  0.73  0.63  0.77  0.66
[2,]  0.71  0.30  0.62  1.00
[3,]  1.71  1.26  1.41  3.00
[4,]  3.24  2.41  1.56  2.56
[5,]  3.25  2.20  1.09  1.29
[6,]  0.66  1.40  1.32  1.13
> head(exp.annotation)
  PlatePosition ImageCloneID                                                                                  Title
4         HK1A1        21652                                                catenin (cadherin-associated protein), alpha 1 (102kD)
5         HK1A2        22012                                                                 ADP-ribosylation factor 3
6         HK1A4        22293                                  uroporphyrinogen III synthase (congenital erythropoietic porphyria)
7         HK1A5        22493                                                                    ribosomal protein L26
8         HK1A6        23019 guanine nucleotide binding protein (G protein), alpha stimulating activity polypeptide 1
9         HK1A7        23132           pre-mRNA splicing factor SF3a (120 kDa subunit), similar to S. cerevisiae PRP21
```

Substructure detected

- Hierarchical clustering of the data reveals notable substructure within the BRCA2 samples.
- Applied **SVA** (Surrogate Variable Analysis) to identified a single surrogate variable that appears to capture this trend.

Substructure detected



Significance analysis

- Included this surrogate variable in a significance analysis comparing BRCA1 and BRCA2 tumors.
- The number of genes differentially expressed between BRCA1 and BRCA2 before and after adjusting for surrogate variables.

	q-Value Threshold			
Analysis Type	0.01	0.025	0.05	0.10
Unadjusted	1	19	96	275
SVA adjusted	0	10	48	190

R package: sva (1)

- Create the full model matrix - including both the adjustment variables and the variable of interest (BRCA1/BRCA2).
- The null model contains only the adjustment variables.

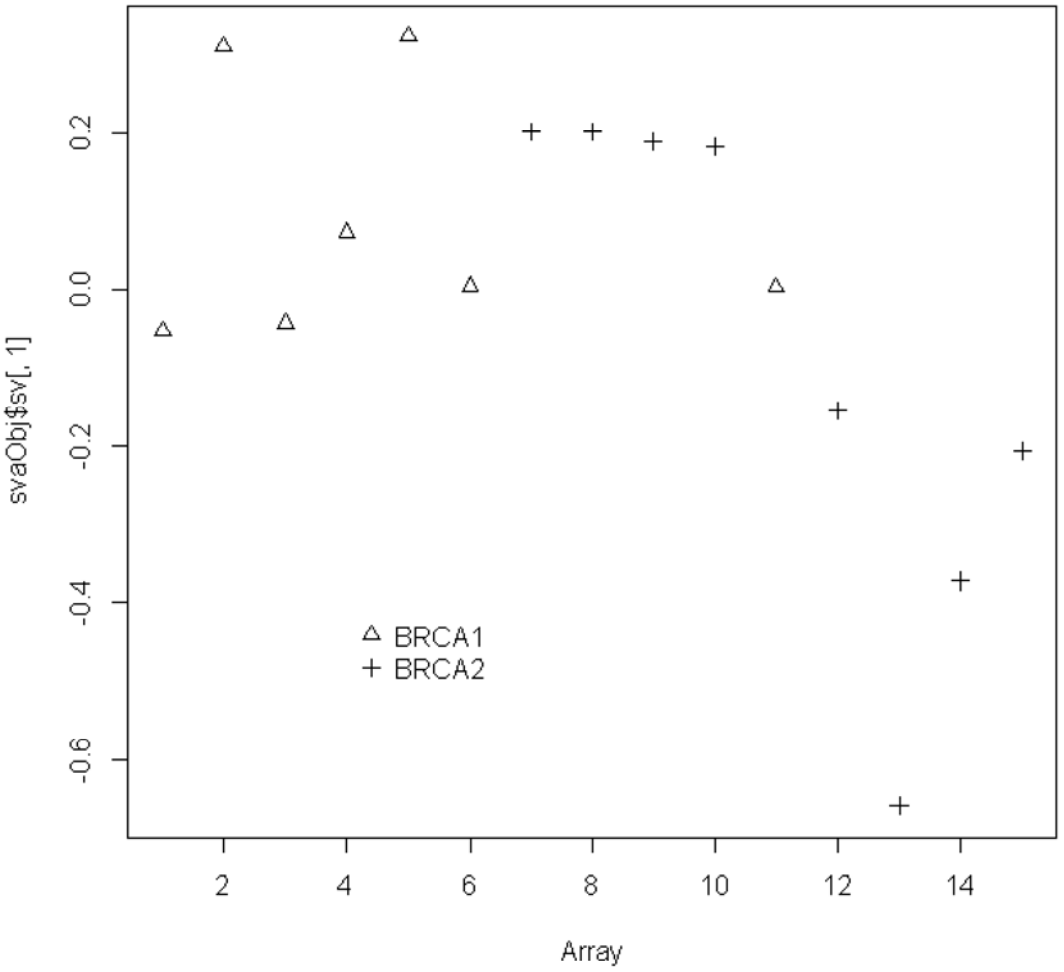
```
> mod=model.matrix(~as.factor(Mutation),data=SVA.pheno)
> mod0=model.matrix(~1,data=SVA.pheno)
> svaObj=sva(log2(SVA.exp.preprocessed[,-1:-3]),mod,mod0)
Number of significant surrogate variables is:  4
Iteration (out of 5 ):1  2  3  4  5  >
```

R package: sva (2)

- The f.pvalue function can be used to calculate parametric F-test p-values for each gene. The F-test compares the models mod and mod0.

```
> pValues.before=f.pvalue(log2(SVA.exp.preprocessed[,-1:-3]),mod,mod0)
> qValuesObj.before<-qvalue(pValues.before)
> qsummary(qValuesObj.before)

Call:
qvalue(p = pValues.before)

pi0:    0.6786508

Cumulative number of significant calls:

        <1e-04 <0.001 <0.01 <0.025 <0.05 <0.1   <1
p-value      9     62   228    392   565  832 3170
q-value      0      0     1     19    96  275 3170
```

R package: sva (3)



R package: sva (4)

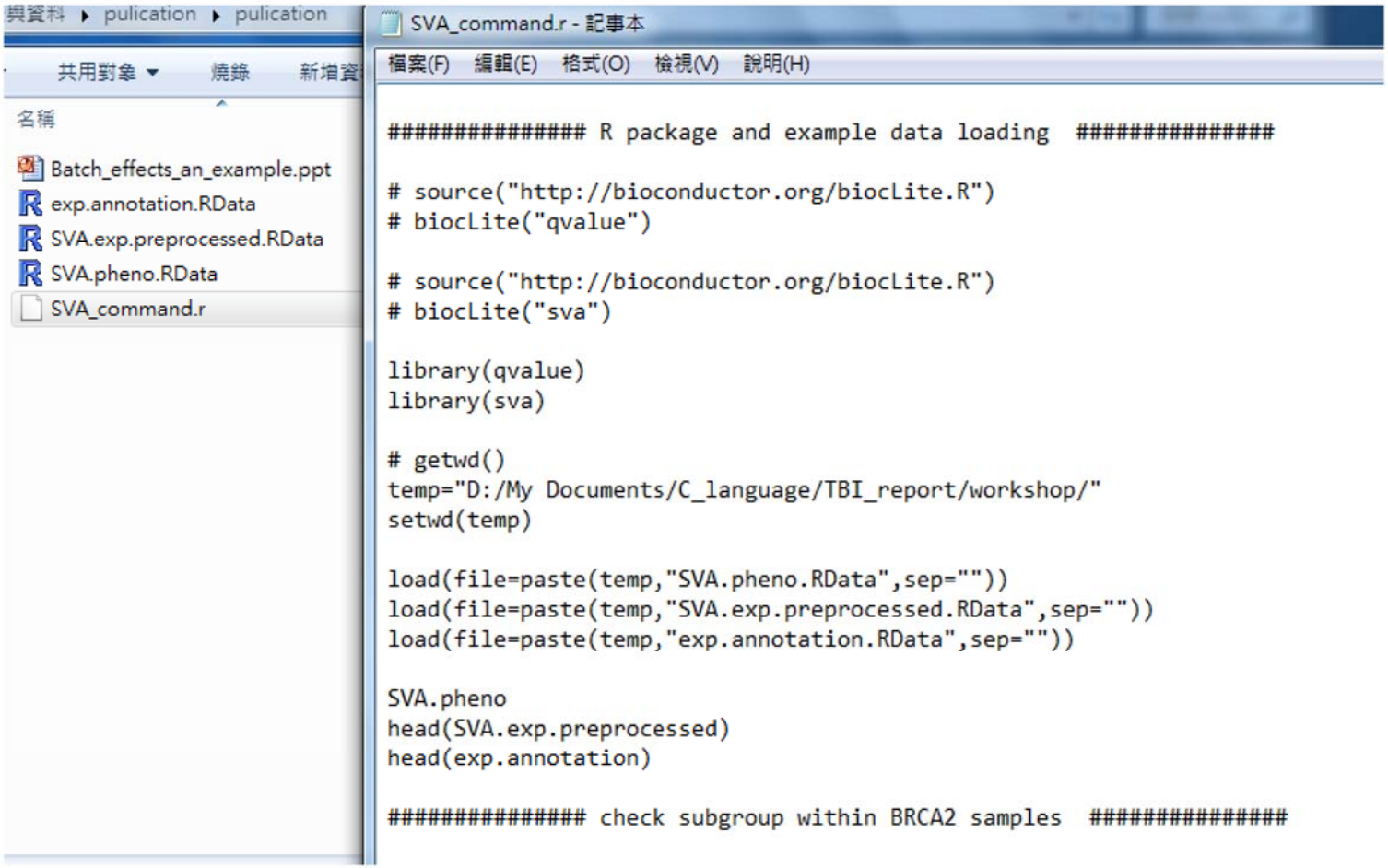
- Now we can perform the same analysis, but adjusting for surrogate variables. The first step is to include the surrogate variables in both the null and full models.

```
> svaObj=sva(log2(SVA.exp.preprocessed[,-1:-3]),mod,mod0)
Number of significant surrogate variables is:  4
Iteration (out of 5 ):1  2  3  4  5  >
> modSv=cbind(mod,svaObj$sv[,1])
> mod0Sv=cbind(mod0,svaObj$sv[,1])
```

- Then P-values and Q-values can be computed as before.

Step by Step: Open the text file

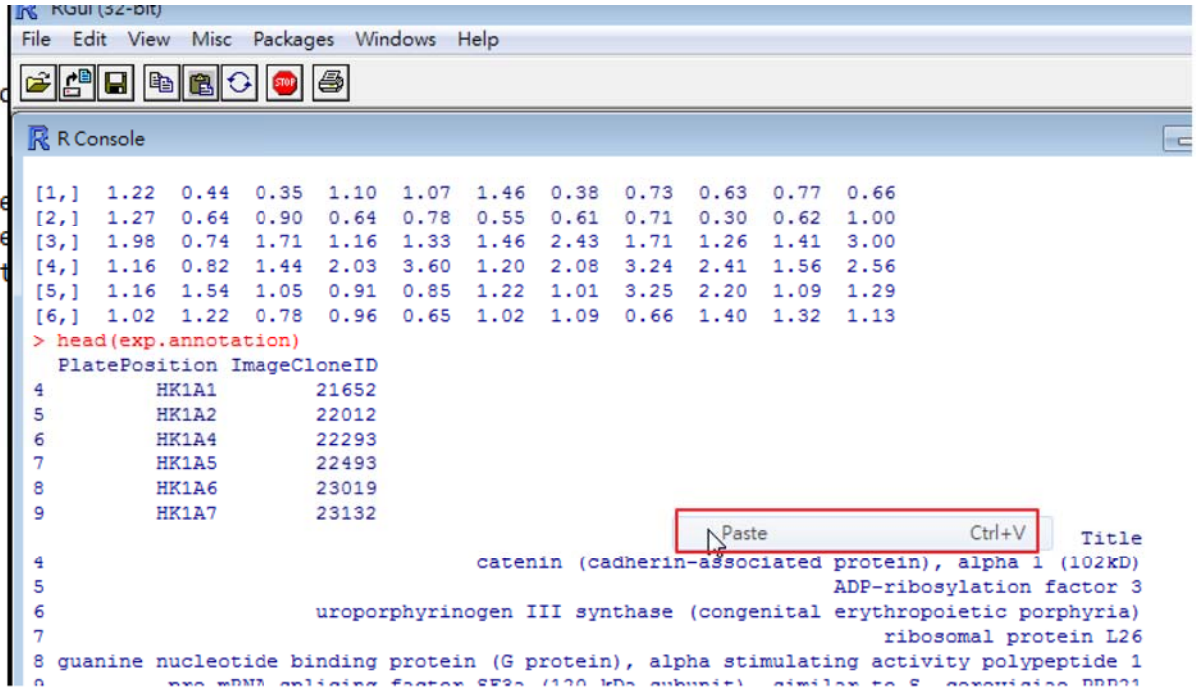
Use the Notepad to open the file named “SVA_command.r” (text file).



Step by Step: Install R packages (1)

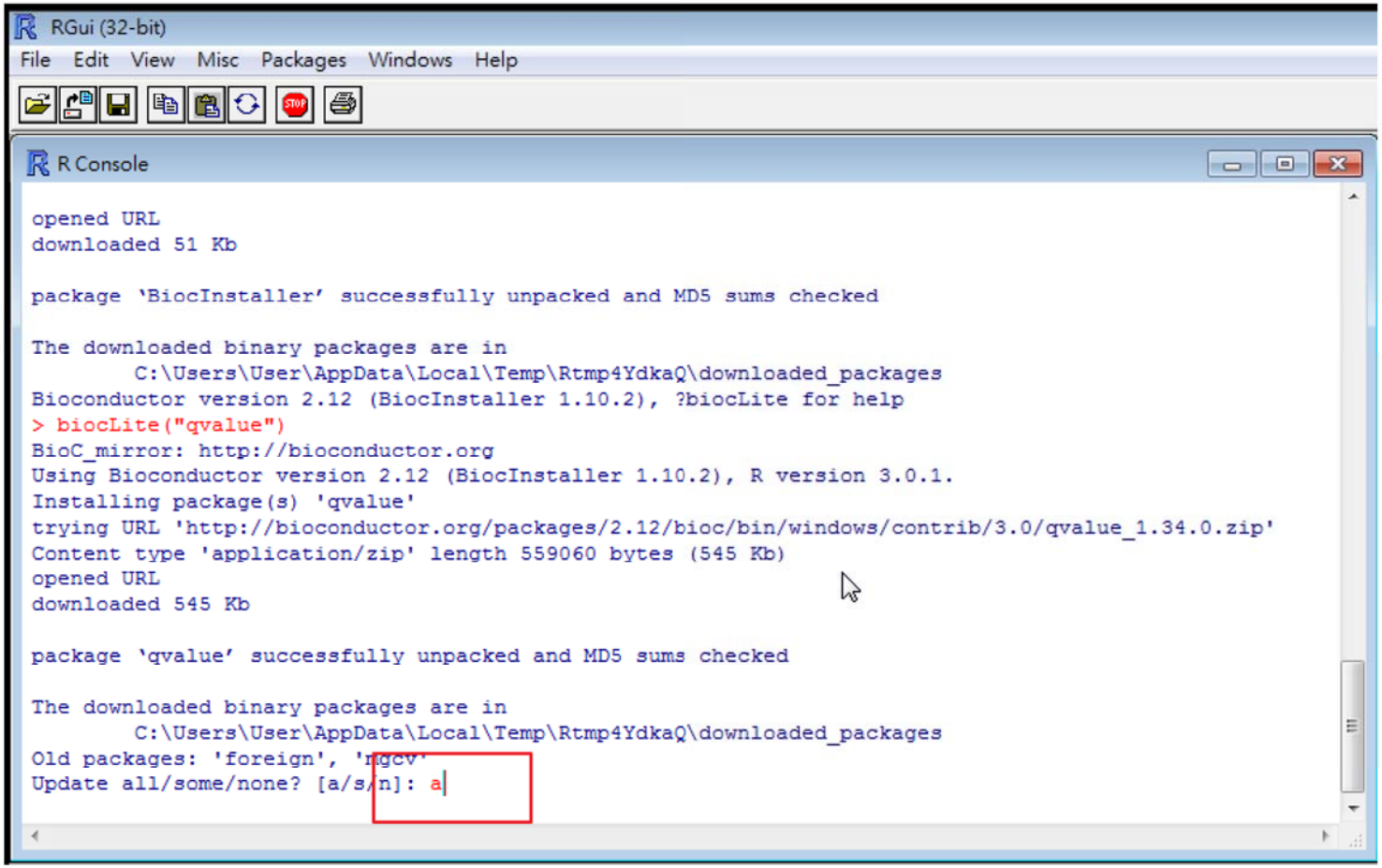
```
############### R package and example data loading  ###############

source("http://bioconductor.org/biocLite.R")
biocLite("qvalue")

source                                    ocLite.R")
biocLi

librar
librar

# getw
temp="                                    3I_repo
setwd(

load(t                                    ata",se
load(t                                    rocesse
load(t                                    on.RDat

SVA.pheno
head(SVA.exp.preprocessed)
head(exp.annotation)
```


1. Copy the command lines for installing the R package “qvalue”.



2. Paste to the R window.

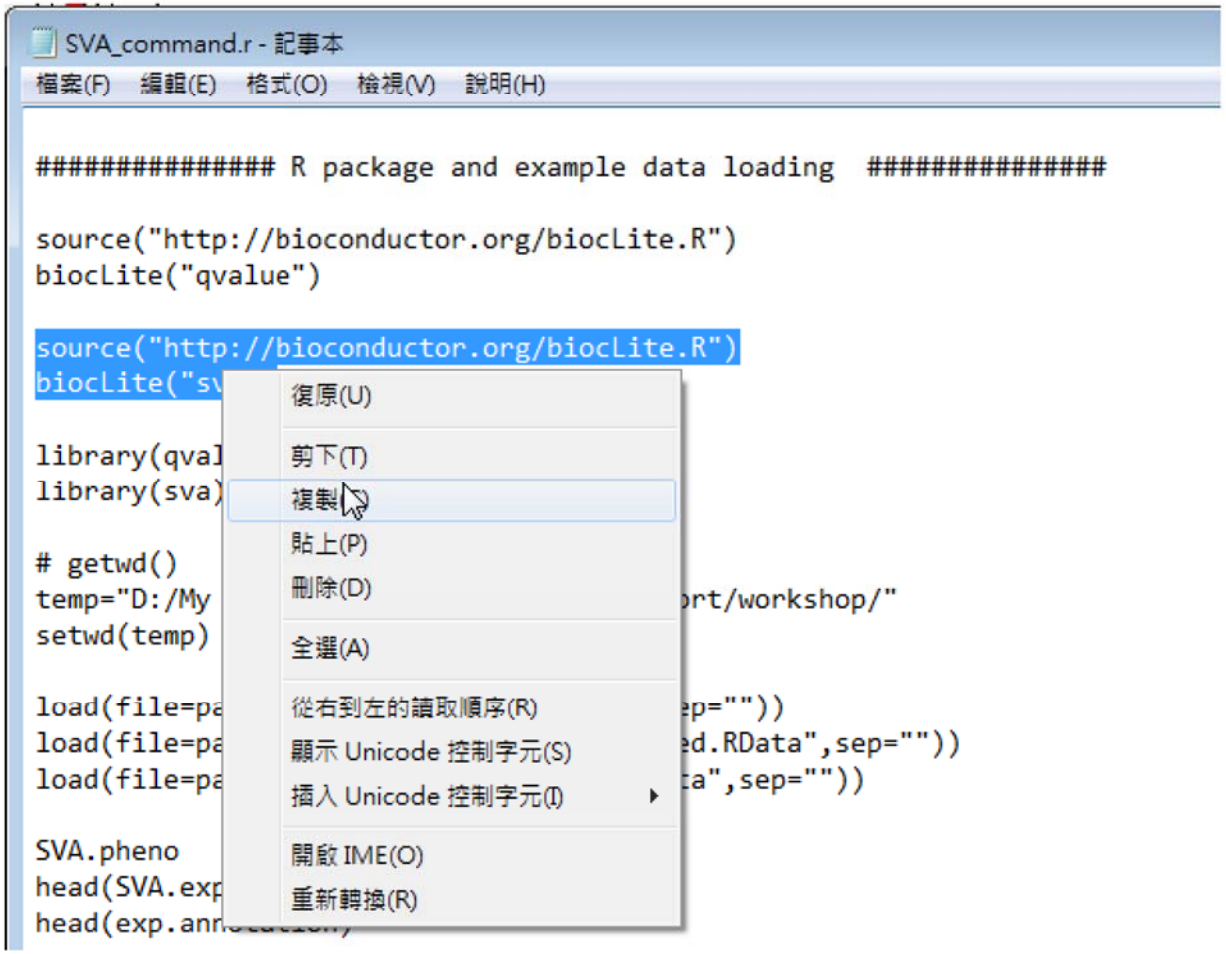
Step by Step: Install R packages (2)



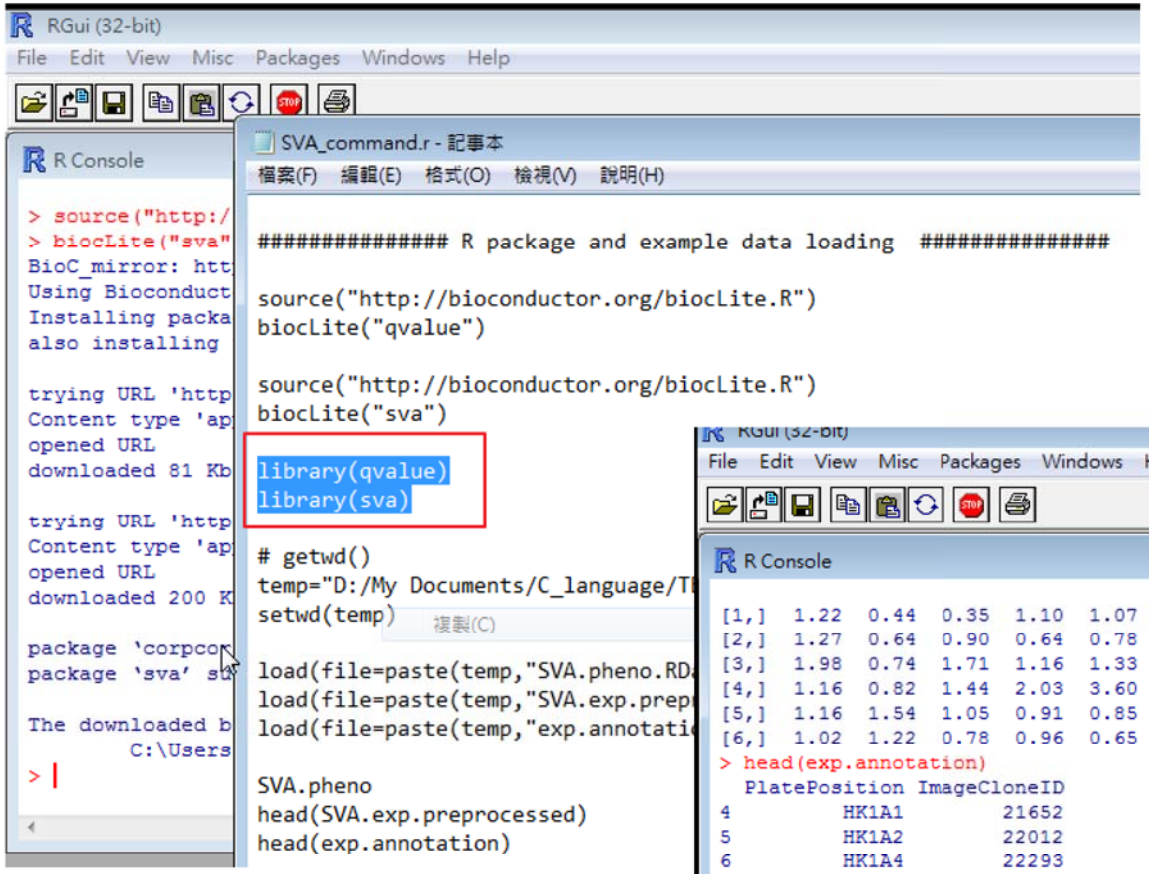
3. Type “a” for updating all old packages.

Step by Step: Install R packages (3)

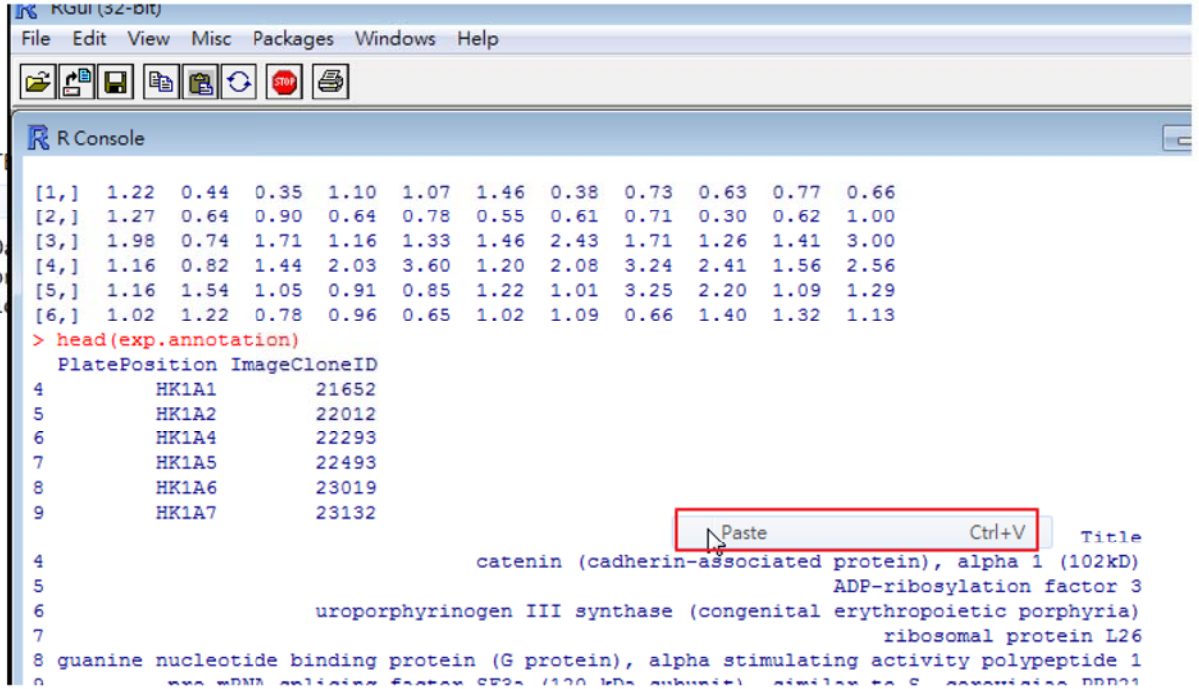
4. Repeat step 1-3. The same way to install the package “sva”.



Step by Step : Load R packages



1. Copy the command lines for loading two R packages, “qvalue” and “sva”.



2. Paste to the R window.

Step by Step : R environment

1. Change the current working directory for loading the example data files.
2. Replace the symbol “\” to “/”.

SVA_command.r - 記事本

檔案(F) 編輯(E) 格式(O) 檢視(V) 說明(H)

```
source("http://bioconductor.org/biocLite.R")
biocLite("qvalue")

source("http://bioconductor.org/biocLite.R")
biocLite("sva")

library(qvalue)
library(sva)

# getwd()
temp="C:\Users\User\Desktop\請安裝軟體與資料\pulication/pulication/"
setwd(temp)

load(file=paste(temp,"SVA.pheno.RData",sep=""))
load(file=paste(temp,"SVA.exp.preprocessed.RData",sep=""))
load(file=paste(temp,"exp.annotation.RData",sep=""))

SVA.pheno
head(SVA.exp.preprocessed)
head(exp.annotation)
```

Step by Step : Load example files

SVA_command.r - 記事本

檔案(F) 編輯(E) 格式(O) 檢視(V) 說明(H)

```
source("http://bioconductor.org/biocLite.R")
biocLite("qvalue")

source("http://bioconductor.org/biocLite.R")
biocLite("sva")

library(qvalue)
library(sva)

# getwd()
temp="C:/Users/User/Desktop/請安裝軟體
setwd(temp)

load(file=paste(temp,"SVA.pheno.RData"
load(file
load(file

SVA.phenc
head(SVA.
head(exp.
```

復原(U)
剪下(T)
複製(C)
貼上(P)
刪除(D)

1. Copy the command lines for loading example files.

RGui (32-bit)

File Edit View Misc Packages Windows Help

R Console

```
[1,]  1.22  0.44  0.35  1.10  1.07  1.46  0.38  0.73  0.63  0.77  0.66
[2,]  1.27  0.64  0.90  0.64  0.78  0.55  0.61  0.71  0.30  0.62  1.00
[3,]  1.98  0.74  1.71  1.16  1.33  1.46  2.43  1.71  1.26  1.41  3.00
[4,]  1.16  0.82  1.44  2.03  3.60  1.20  2.08  3.24  2.41  1.56  2.56
[5,]  1.16  1.54  1.05  0.91  0.85  1.22  1.01  3.25  2.20  1.09  1.29
[6,]  1.02  1.22  0.78  0.96  0.65  1.02  1.09  0.66  1.40  1.32  1.13
> head(exp.annotation)
  PlatePosition ImageCloneID
4         HK1A1        21652
5         HK1A2        22012
6         HK1A4        22293
7         HK1A5        22493
8         HK1A6        23019
9         HK1A7        23132
```

Paste Ctrl+V

```
                                                                  Title
4                                  catenin (cadherin-associated protein), alpha 1 (102kD)
5                                                        ADP-ribosylation factor 3
6                          uroporphyrinogen III synthase (congenital erythropoietic porphyria)
7                                                        ribosomal protein L26
8 guanine nucleotide binding protein (G protein), alpha stimulating activity polypeptide 1
```

2. Paste to the R window.

Step by Step : Command lines

The package “sva” command lines.

```
mod=model.matrix(~as.factor(Mutation),data=SVA.pheno)

mod0=model.matrix(~1,data=SVA.pheno)

svaObj=sva(log2(SVA.exp.preprocessed[,-1:-3]),mod,mod0)
```

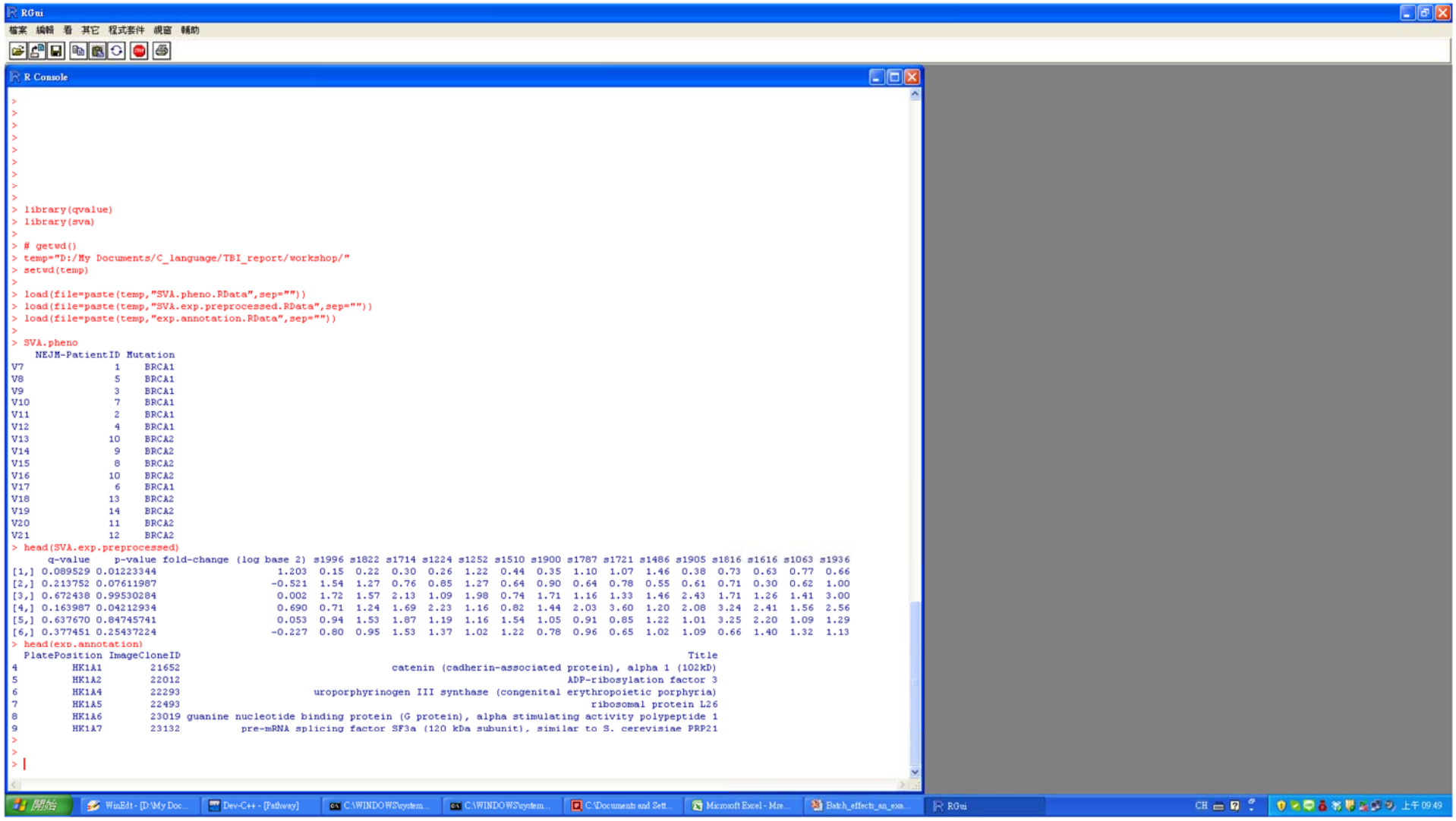
The command for plotting the graph.

```
symbol<-rep(2,15)
symbol[BRCA2]<-3
svaObj$sv

plot(1:15,svaObj$sv[,1],xlab="Array",ylab="Covariate Value",pch=symbol)
legend(4,-0.4,c("BRCA1","BRCA2"),pch=c(2,3),bty = "n")
```

The command for computing F-test and q-value.

```
################ compare q-value before/after batch effect  ################


pValues.before=f.pvalue(log2(SVA.exp.preprocessed[,-1:-3]),mod,mod0)
qValuesObj.before<-qvalue(pValues.before)

qsummary(qValuesObj.before)

modSv=cbind(mod,svaObj$sv[,1])
mod0Sv=cbind(mod0,svaObj$sv[,1])

pValues.after=f.pvalue(log2(SVA.exp.preprocessed[,-1:-3]),modSv,mod0Sv)
qValuesObj.after<-qvalue(pValues.after)

qsummary(qValuesObj.after)
```

Appendix: print screen (1)



Appendix: print screen (2)

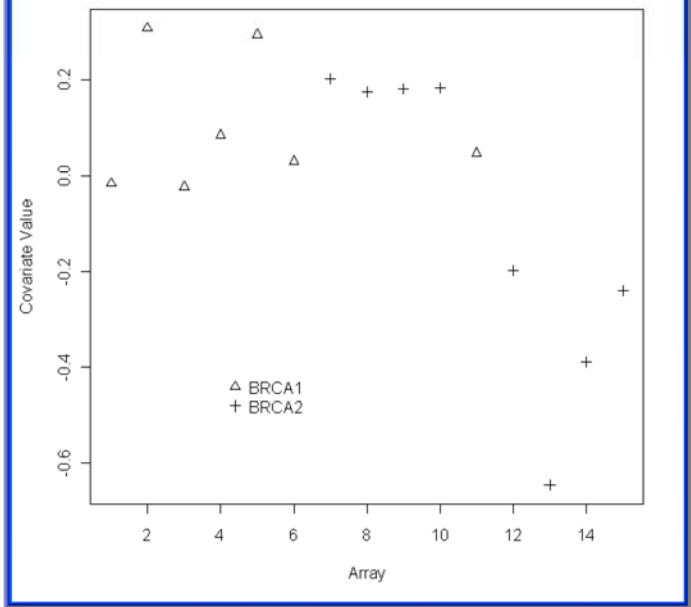
RGui

檔案 命令歷程 重設大小 視窗

R Console

```
V19             14    BRCA2
V20             11    BRCA2
V21             12    BRCA2
> head(SVA.exp.preprocessed)
      q-value    p-value fold-change (log base 2) s1996 s1822 s1714 s1224 s1252 s1510 s1900 s1787 s1721 s1486 s1905 s1816 s1616 s1063 s1936
[1,] 0.089529 0.01223344                   1.203  0.15  0.22  0.30  0.26  1.22  0.44  0.35  1.10  1.07  1.46  0.38  0.73  0.63  0.77  0.66
[2,] 0.213752 0.07611987                  -0.521  1.54  1.27  0.76  0.85  1.27  0.64  0.90  0.64  0.78  0.55  0.61  0.71  0.30  0.62  1.00
[3,] 0.672436 0.99530284                   0.002  1.72  1.57  2.13  1.09  1.98  0.74  1.71  1.16  1.33  1.46  2.43  1.71  1.26  1.41  3.00
[4,] 0.163987 0.04212934                   0.690  0.71  1.24  1.69  2.23  1.16  0.82  1.44  2.03  3.60  1.20  2.08  3.24  2.41  1.56  2.56
[5,] 0.637670 0.84745741                   0.053  0.94  1.53  1.87  1.19  1.16  1.54  1.05  0.91  0.85  1.22  1.01  3.25  2.20  1.09  1.29
[6,] 0.377451 0.25437224                  -0.227  0.80  0.95  1.53  1.37  1.02  1.22  0.78  0.96  0.65  1.02  1.09  0.66  1.40  1.32  1.13
> head(exp.annotation)
  PlatePosition ImageCloneID                                                                                  Title
4         HK1A1        21652                                        catenin (cadherin-associated protein), alpha 1 (102kD)
5         HK1A2        22012                                                        ADP-ribosylation factor 3
6         HK1A4        22293                              uroporphyrinogen III synthase (congenital erythropoietic porphyria)
7         HK1A5        22493                                                              ribosomal protein L26
8         HK1A6        23019 guanine nucleotide binding protein (G protein), alpha stimulating activity polypeptide 1
9         HK1A7        23132          pre-mRNA splicing factor SF3a (120 kDa subunit), similar to S. cerevisiae PRP21
>
> ################ check subgroup within BRCA2 samples  ################
>
>
> BRCA1=which(SVA.pheno[,"Mutation"]=="BRCA1")
> BRCA2=which(SVA.pheno[,"Mutation"]=="BRCA2")
>
> mod=model.matrix(~as.factor(Mutation),data=SVA.pheno)
>
> mod0=model.matrix(~1,data=SVA.pheno)
>
> svaObj=sva(log2(SVA.exp.preprocessed[,-1:-3]),mod,mod0)
>
> symbol<-rep(2,15)
> symbol[BRCA2]<-3
> svaObj$sv
             [,1]        [,2]         [,3]        [,4]
 [1,] -0.01550500 -0.06346957 -0.295757985  0.28139222
 [2,]  0.30975202  0.22338364  0.106212414  0.07891533
 [3,] -0.02298093 -0.56946215 -0.009682216 -0.51731850
 [4,]  0.08520060 -0.20558439  0.169768683 -0.06873269
 [5,]  0.29635047  0.32762443  0.045267743 -0.03384718
 [6,]  0.03086823 -0.08592774  0.302552340  0.35604415
 [7,]  0.20187282 -0.10329907  0.164672789  0.13410345
 [8,]  0.17522354 -0.05996738  0.051943643 -0.36904042
 [9,]  0.18119710 -0.05810417 -0.305412223 -0.18798714
[10,]  0.10405557 -0.05661062  0.362025595  0.27419233
[11,]  0.04739771 -0.11128608 -0.705597310  0.32237227
[12,] -0.19748974  0.54013239 -0.046892171 -0.30015913
[13,] -0.64699967 -0.15469287  0.139191736  0.19800041
[14,] -0.39011196  0.03909485  0.060417068 -0.04994333
[15,] -0.23883078  0.33817674 -0.038710105 -0.11799178
>
> plot(1:15,svaObj$sv[,1],xlab="Array",ylab="Covariate Value",pch=symbol)
> legend(4,-0.4,c("BRCA1","BRCA2"),pch=c(2,3),bty = "n")
>
>
```

R Graphics: Device 2 (ACTIVE)



開始 WinEdt - [D:\My Doc... Dev-C++ - [Pathway] C:\WINDOWS\system... C:\WINDOWS\system... C:\Documents and Sett... Microsoft Excel - Mic... Batch_effects_an_exa... RGui CH 上午 10:02

Appendix: print screen (3)

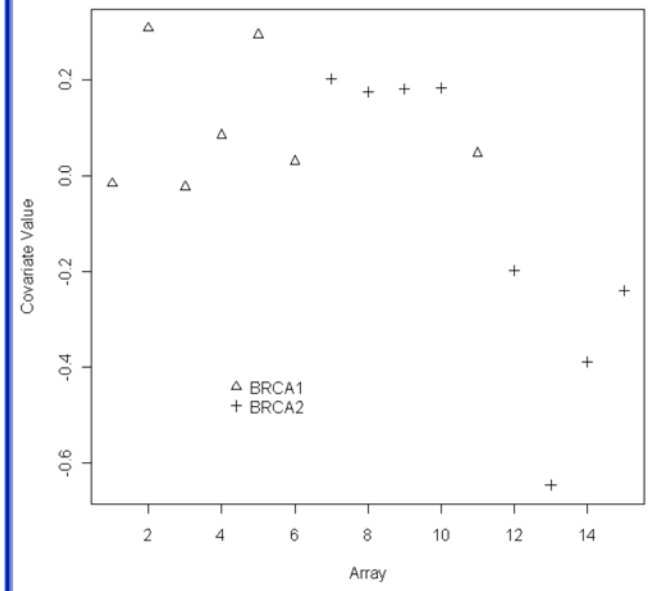
RGui

檔案 編輯 看 其它 程式套件 視窗 輔助

R Console

```
 [7,]  0.20107202 -0.10329907  0.164672709  0.13410345
 [8,]  0.17522354 -0.05996738  0.051943643 -0.36904042
 [9,]  0.18119710 -0.05810417 -0.305412223 -0.18798714
[10,]  0.18405557 -0.05661862  0.362025595  0.27419233
[11,]  0.04739771 -0.11128608 -0.705597310  0.32237227
[12,] -0.19748974  0.54013239 -0.046892171 -0.30015913
[13,] -0.64699967 -0.15469287  0.139191736  0.19800041
[14,] -0.39011196  0.03909485  0.060417068 -0.04994333
[15,] -0.23883078  0.33817674 -0.038710105 -0.11799178
>
> plot(1:15,svaObj$sv[,1],xlab="Array",ylab="Covariate Value",pch=symbol)
> legend(4,-0.4,c("BRCA1","BRCA2"),pch=c(2,3),bty = "n")
>
>
> ################ compare q-value before/after batch effect  ################
>
>
> pValues.before=f.pvalue(log2(SVA.exp.preprocessed[,-1:-3]),mod,mod0)
> qValuesObj.before<-qvalue(pValues.before)
>
> qsummary(qValuesObj.before)

Call:
qvalue(p = pValues.before)

pi0:    0.6786508

Cumulative number of significant calls:

        <1e-04 <0.001 <0.01 <0.025 <0.05 <0.1   <1
p-value      9     62   220    392   565  832 3170
q-value      0      0     1     19    96  275 3170

>
> modSv=cbind(mod,svaObj$sv[,1])
> mod0Sv=cbind(mod0,svaObj$sv[,1])
>
> pValues.after=f.pvalue(log2(SVA.exp.preprocessed[,-1:-3]),modSv,mod0Sv)
> qValuesObj.after<-qvalue(pValues.after)
>
> qsummary(qValuesObj.after)

Call:
qvalue(p = pValues.after)

pi0:    0.6804928

Cumulative number of significant calls:

        <1e-04 <0.001 <0.01 <0.025 <0.05 <0.1   <1
p-value     10     40   188    322   484  729 3170
q-value      0      0     0     10    35  149 3170

>
>
```

R Graphics: Device 2 (ACTIVE)



開始 WinEdt - [D:\My Doc... Dev-C++ - [Pathway] C:\WINDOWS\system... C:\WINDOWS\system... C:\Documents and Sett... Microsoft Excel - Mic... Batch_effects_an_exa... RGui CH 上午 10:02

Appendix: print screen (4)

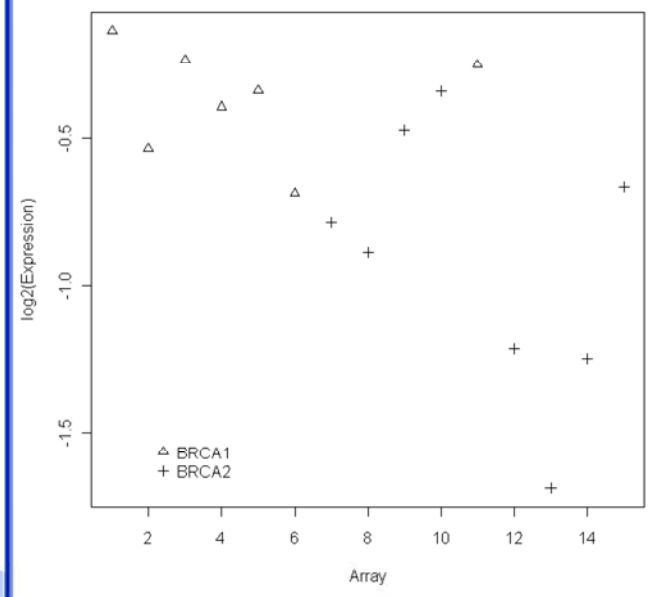
RGui

檔案 編輯 看 其它 程式套件 視窗 輔助

R Console

```
pi0:    0.6706500

Cumulative number of significant calls:

        <1e-04 <0.001 <0.01 <0.025 <0.05 <0.1   <1
p-value      9     62   228    392   565  832 3170
q-value      0      0     1     19    96  275 3170

>
> modSv=cbind(mod,svaObj$sv[,1])
> mod0Sv=cbind(mod0,svaObj$sv[,1])
>
> pValues.after=f.pvalue(log2(SVA.exp.preprocessed[,-1:-3]),modSv,mod0Sv)
> qValuesObj.after<-qvalue(pValues.after)
>
> qsummary(qValuesObj.after)

Call:
qvalue(p = pValues.after)

pi0:    0.6916102

Cumulative number of significant calls:

        <1e-04 <0.001 <0.01 <0.025 <0.05 <0.1   <1
p-value      9     37   186    314   476  720 3170
q-value      0      0     0      9    34  143 3170

>
>
>
> ################ a example (eukaryotic translation factor 2 (EIFS2))  ################
>
>
> gene.no<-43
>
> exp.annotation[gene.no,]
   PlatePosition ImageCloneID                                                Title
46         HK1E7        41511 eukaryotic translation initiation factor 2, subunit 2 (beta, 38kD )
>
> plot(1:15,log2(SVA.exp.preprocessed[gene.no,-1:-3]),xlab="Array",ylab="log2(Expression)",pch=symbol)
> legend(2,-1.5,c("BRCA1","BRCA2"),pch=c(2,3),bty = "n")
>
>
> pValues.before[gene.no]
[1] 0.01060463
> qValuesObj.before$qvalue[gene.no]
[1] 0.09440496
>
>
> pValues.after[gene.no]
[1] 0.03754076
> qValuesObj.after$qvalue[gene.no]
[1] 0.2022719
>
>
```

R Graphics: Device 2 (ACTIVE)



開始 WinEdt - [D:\My Doc... Dev-C++ - [Pathway] C:\WINDOWS\system... C:\WINDOWS\system... C:\Documents and Sett... Microsoft Excel - Mic... Batch_effects_an_exa... RGui CH 上午 09:58