Summer Institute in Statistical Genetics 2016

Integrative Genomics

3. Gene Expression Exercises with R

ggibson.gt@gmail.com
http://www.gibsongroup.biology.gatech.edu

WIKIPEDIA

List of RNA-Seq bioinformatics tools

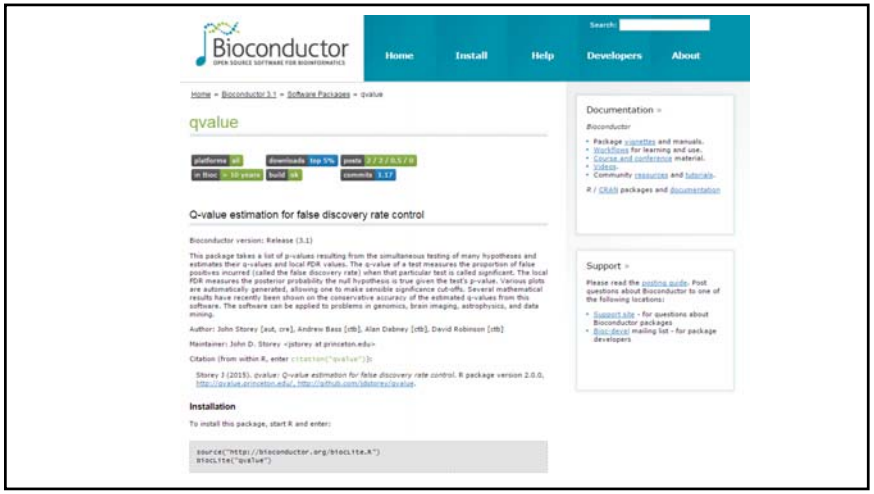
From Wikipedia, the free encyclopedia

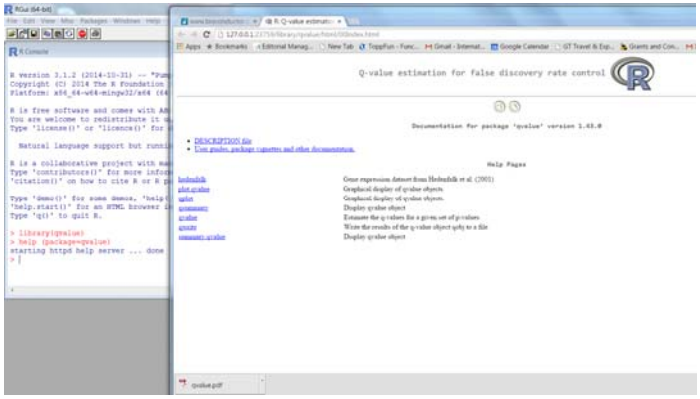
Contents [hide]

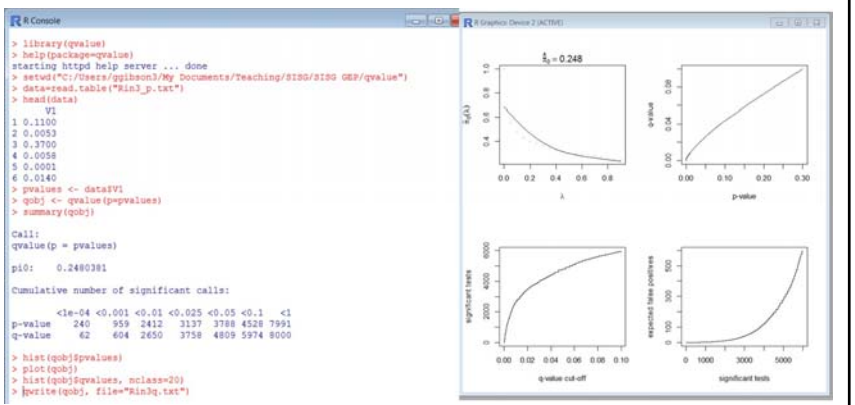
qvalue

qvalue is an R package for determining the false discovery rate from a list of p-values, adjusted for an estimate of the number of true nulls

Input: Rin3_p.txt




Q-value estimation for false discovery rate control
Documentation for package 'qvalue' version 1.40.0
DESCRIPTION file
User guides, package vignettes and other documentation.
Help Pages


R Console
> library(qvalue)
> help(package=qvalue)
starting httpd help server ... done
> setwd("C:/Users/ggibson3/My Documents/Teaching/SISG/SISG GEP/qvalue")
> data=read.table("Rin3_p.txt")
> head(data)
V1
1 0.1100
2 0.0053
3 0.3700
4 0.0058
5 0.0001
6 0.0140
> pvalues <- data$V1
> qobj <- qvalue(p=pvalues)
> summary(qobj)
Call:
qvalue(p = pvalues)
pi0: 0.2480381
Cumulative number of significant calls:
<1e-04 <0.001 <0.01 <0.025 <0.05 <0.1 <1
p-value 240 959 2412 3137 3788 4528 7991
q-value 62 604 2650 3758 4809 5974 8000
> hist(qobj$pvalues)
> plot(qobj)
> hist(qobj$qvalues, nclass=20)
> qwrite(qobj, file="Rin3q.txt")
R Graphics: Device 2 (ACTIVE)
q-value cut-off
significant tests

snm

snm is an R package for supervised normalization of microarray (or RNASeq) data, that simultaneously adjusts biological, technical and array effects

Input:	CAD8000.csv	data file
	CAD_bio.csv	biological variable
	CAD_adj.csv	adjustment variable (to fit)
	CAD_adjrm	adjustment variable (to remove)
	CAD_int	intensity-dependent (array) variable

DATA

		EUH02661	EUH01927	EUH02357	CLH00229	EUH02482	CLH00189	EUH02317	EUH02121	EUH02210	CLH00238	EUH02394
2	ILMN_238	14.68244	13.98202	14.46552	14.44341	13.87446	14.39205	13.8535	14.29878	14.03563	13.41536	13.69829
3	ILMN_166	14.51808	13.98157	15.04292	14.69002	13.74107	14.64559	13.61555	14.21511	13.5964	13.50156	13.47871
4	ILMN_168	14.09896	13.73664	13.98173	13.91768	13.85829	13.88826	13.88211	14.13509	13.9924	13.65327	13.81957
5	ILMN_210	14.88958	13.71523	14.93602	14.42121	13.55921	14.44894	13.45433	14.2198	13.53969	13.4686	13.42105
6	ILMN_224	14.35569	13.90495	14.45346	14.17866	13.62601	14.23437	13.51523	14.29185	13.69174	13.42759	13.58778
7	ILMN_212	14.67233	13.86557	15.08908	14.35686	13.54718	14.54751	13.61004	14.27196	13.5793	13.34899	13.41936
8	ILMN_223	14.15323	13.81096	14.06618	14.00923	13.7164	13.89014	13.54152	14.29942	13.57882	13.42607	13.56903
9	ILMN_209	13.92744	13.84332	13.89218	13.83482	13.81746	13.56429	13.85589	14.17429	13.79296	13.60181	13.64221
10	ILMN_222	13.81189	13.59864	13.79147	13.66698	13.80799	13.60064	13.73122	14.02785	13.92121	13.61169	13.75067

BIOL

		CVD_TYPE
2	EUH02661	ACUTE
3	EUH01927	ACUTE
4	EUH02357	ACUTE
5	CLH00229	ACUTE
6	EUH02482	ACUTE
7	CLH00189	ACUTE
8	EUH02317	ACUTE
9	EUH02121	ACUTE
10	EUH02210	FINE

ADJ_RM

		Study	Rin3
50	EUH02253	A	MOD
51	EUH02381	A	HIGH
52	EUH01996	A	MOD
53	CLH00187	A	MOD
54	GG_0372	B	MOD
55	GG_0380	B	HIGH
56	GG_0396	B	MOD
57	GG_0398	B	LOW
58	GG_0399	B	LOW

ADJ

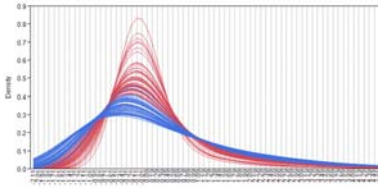
		Gender	Age	BMI
2	EUH02661	FEM	56	31
3	EUH01927	FEM	54	26.9
4	EUH02357	MAL	51	20.7
5	CLH00229	MAL	52	24.4
6	EUH02482	MAL	62	26.1
7	CLH00189	MAL	65	29.9
8	EUH02317	MAL	65	28.4
9	EUH02121	MAL	64	33.1
10	EUH02210	FEM	69	29.8

INT

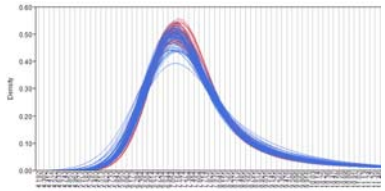
	Array
2	1
3	2
4	3
5	4
6	5
7	6
8	7
9	8
10	9

```
> summary(snmR.cad)

SNM Data and Model Summary

Total number of arrays: 100
Total number of probes: 8000

Final estimated proportion of null probes: 0.525

Cumulative number of significant calls:
          <1e-04 <0.001 <0.01 <0.025 <0.05 <0.1   <1
p-value       49    239   984   1505  1990 2777 8000
q-value        0      0    65    504  1143 1914 8000

Full model degrees of freedom: 5
Null model degrees of freedom: 4

Biological variables:
     EGN02641 EGN01927 EGN02357 CLN00229 EGN02482 CLN00189 EGN02317 EGN02121 EGN02210 CLN00238
[1,]        0        0        0        0        0        0        0        0        1        0
     EGN02394 EGN02085 EGN02438 CLN00143 EGN02342 EGN02300 EGN02512 EGN02392 EGN02433 EGN01977
[1,]        0        1        0        0        1        1        1        0        0        1
```

BEFORE SNM



AFTER SNM1



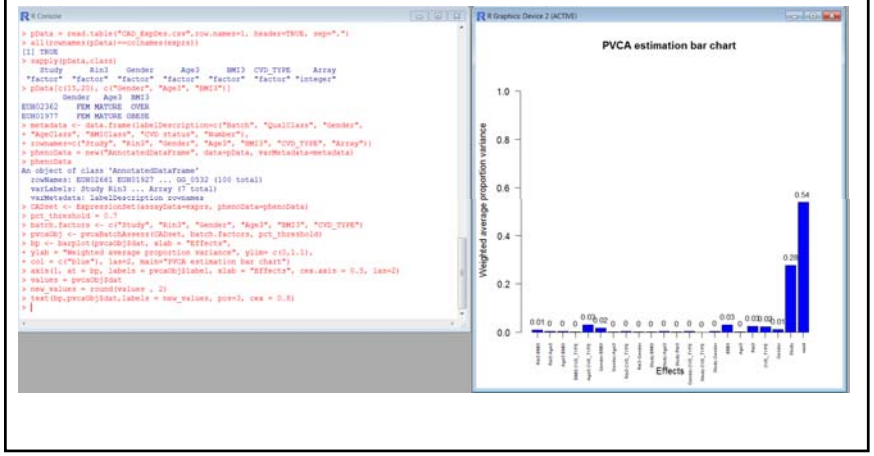
pvca

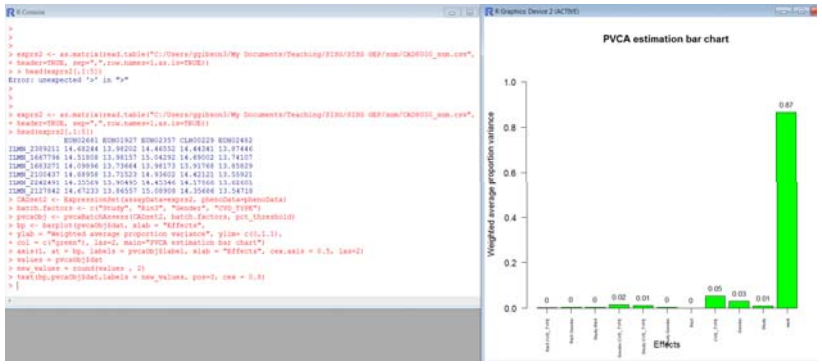
pvca is an R package for assessing the principal variance components in a gene expression dataset, namely the proportion of the variance in the first n PC that is due to covariates of interest

Input: CAD8000.csv
CAD_ExpDes

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

```
> setwd("C:/Users/ggibson3/My Documents/Teaching/SISG/SISG GEP/pvca")
> library("Biobase")
Loading required package: BiocGenerics
Loading required package: parallel

> library(pvca)
> exprs <- as.matrix(read.table("CAD8000.csv", header=TRUE, sep=",",row.names=1,as.is=TRUE))
> class(exprs)
[1] "matrix"
> dim(exprs)
[1] 8000  100
> colnames(exprs)
  [1] "EUH02661" "EUH01927" "EUH02357" "CLH00229" "EUH02482" "CLH00189" "EUH02317" "EUH02121"
  [9] "EUH02210" "CLH00238" "EUH02394" "EUH02085" "EUH02638" "CLH00143" "EUH02362" "EUH02300"
 [17] "EUH02512" "EUH02392" "EUH02433" "EUH01977" "EUH02063" "EUH02614" "EUH02054" "EUH02430"
 [25] "EUH02465" "EUH02404" "EUH02686" "EUH02468" "EUH02391" "EUH02158" "EUH02271" "EUH02296"
 [33] "EUH02287" "EUH02659" "EUH02475" "EUH02520" "CLH00094" "EUH02070" "EUH02396" "EUH02297"
 [41] "EUH02276" "EUH02113" "EUH02062" "EUH02366" "CLH00190" "EUH02265" "EUH02102" "EUH02253"
 [49] "EUH02381" "EUH01996" "CLH00187" "GG_0372"  "GG_0380"  "GG_0396"  "GG_0398"  "GG_0399"
 [57] "GG_0400"  "GG_0401"  "GG_0413"  "GG_0419"  "GG_0420"  "GG_0423"  "GG_0425"  "GG_0428"
 [65] "GG_0431"  "GG_0433"  "GG_0434"  "GG_0435"  "GG_0436"  "GG_0437"  "GG_0442"  "GG_0443"
 [73] "GG_0445"  "GG_0452"  "GG_0454"  "GG_0464"  "GG_0469"  "GG_0473"  "GG_0475"  "GG_0478"
 [81] "GG_0484"  "GG_0489"  "GG_0491"  "GG_0503"  "GG_0504"  "GG_0506"  "GG_0507"  "GG_0508"
 [89] "GG_0509"  "GG_0510"  "GG_0511"  "GG_0514"  "GG_0516"  "GG_0517"  "GG_0518"  "GG_0519"
 [97] "GG_0520"  "GG_0525"  "GG_0531"  "GG_0532"
> head(exprs[,1:5])
           EUH02661 EUH01927 EUH02357 CLH00229 EUH02482
ILMN_2389211   14.091   13.892   13.747   14.056   14.243
ILMN_1667796   13.872   13.903   14.273   14.276   14.125
ILMN_1683271   13.419   13.665   13.180   13.487   14.252
ILMN_2100437   14.237   13.638   14.160   14.004   13.945
ILMN_2242491   13.692   13.830   13.667   13.756   14.015
ILMN_2127842   13.992   13.794   14.287   13.926   13.941
>
```





edgeR

edgeR is an R package for normalization of RNASeq counts using the negative binomial distribution to adjust for high variance at low expression

Input:	EM1_week2.txt	data file
	EM1_week2.txt	data file
	EM1_week2.txt	data file
	targets_EM1.txt	design file

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

```
> library(edgeR)
Loading required package: limma
Warning messages:
1: package 'edgeR' was built under R version 3.1.3
2: package 'limma' was built under R version 3.1.3
> setwd("C:/Users/ggibson3/My Documents/Teaching/SISG/SISG GEP/edgeR")
> targets <- readTargets("targets_EM1.txt")
> targets
          files  group           description
1 EM1_Week1.txt   Sick Respiratory infection
2 EM1_Week2.txt   Sick Respiratory infection
3 EM1_Week3.txt Better    Health improvement
> ed <- readDGE(targets, skip=2, comment.char = "!")
> ed$samples
          files  group           description lib.size norm.factors
1 EM1_Week1.txt   Sick Respiratory infection 40396653            1
2 EM1_Week2.txt   Sick Respiratory infection 47774582            1
3 EM1_Week3.txt Better    Health improvement 38294470            1
> head(ed$counts)
      1   2   3
42064 1081  903 1623
42065   24  25  21
42066  170  173  129
42067    1   2   1
42068 1270 1276 1283
42069 7655 8545 7388
> summary(ed$counts)
       1                2                3
 Min.   :     0   Min.   :     0   Min.   :     0
 1st Qu.:     0   1st Qu.:     0   1st Qu.:     0
 Median :    40   Median :    49   Median :    37
 Mean   :  1556   Mean   :  1840   Mean   :  1475
 3rd Qu.:  1041   3rd Qu.:  1280   3rd Qu.:   975
 Max.   :1093284   Max.   :1295951   Max.   :315122
> dim(ed)
[1] 25963     3
```

```
> keep <- rowSums(cpm(ed)>50) >= 2
> ed <- ed[keep,,keep.lib.sizes=FALSE]
> head(ed$counts)
      1    2    3
42069  7655  8545  7388
42070  6431  7028  6784
42071  5553  5945  4441
42248  3845  5314  3461
42249 13088 15971 12261
42253 11005 14132  8969
> dim(ed)
[1] 4356    3
```

```
>
>
>
>
> plotBCV(ed)
> et <- exactTest(ed)
> topTags(et, n=5)
Comparison of groups:  Sick-Better
          logFC    logCPM       PValue          FDR
ALAS2  2.371452  6.594378 5.787006e-40 2.520820e-36
HBB    2.036585 14.530549 7.837729e-34 1.707037e-30
SLC4A1 2.140032  5.407104 4.738589e-25 6.880432e-22
HBA1   1.634547 10.727079 1.623237e-22 1.767705e-19
HBA2   1.569068 12.949427 2.692825e-21 2.345989e-18
> detags <- rownames(topTags(et, n=5))
> cpm(ed)[detags,]
             1         2         3
ALAS2    126.86217  137.2773   25.52189
HBB    31500.42618 31785.7160 7712.67945
SLC4A1    72.43504  50.7175   14.87007
HBA1    2114.74321  2264.7934  705.26609
HBA2    9869.95144 10437.2443 3422.02752
> summary(de <- decideTestsDGE(et, p=0.05, adjust="BH"))
   [,1]
-1  233
0  4014
1  109
> detags <- rownames(ed)[as.logical(de)]
> plotSmear(et, de.tags=detags)
> abline(h = c(-0.5, 0.5), col = "blue")
>
> write.table(detags, file="detags.csv", sep=",", col.names=NA)
>
```