Expression Workshop

Humberto Ortiz-Zuazaga

November 14, 2014

Abstract

The bioconductor project publishes software and data for the analysis of functional genomics experiments, using a variety of techniques such as microarrays or second-generation sequencing. We will explore the bioconductor tools for expression analysis and pathway analysis and apply them to a demo dataset.

Installing bioconductor

Bioconductor comes with it's own installation procedures for software and data packages, the `biocLite()` function. You can use it to install any bioconductor package, along with all prerequisites. For example, to install the `limma` package, we can use code like this:

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

Once you have sourced the biocLite file in your R session, you can continue to use it to install additional packages, such as the pathview package:

```
biocLite("pathview")
```

Installing Data packages

Bioconductor includes many experimental data packages as well, that can be installed with `biocLite()` function as well:

```
biocLite("breastCancerMAINZ")
biocLite("hgu133a.db")
```

Loading packages

Once packages have been installed, you need to load the package into each session that will use them. R-Studio will keep the packages loaded into a workspace, even if you exit and restart the program, as long as you save your workspace.

```
library(Biobase)
```

```
## Loading required package: BiocGenerics
## Loading required package: parallel
##
## Attaching package: 'BiocGenerics'
##
## The following objects are masked from 'package:parallel':
##
```

```
##     clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##     clusterExport, clusterMap, parApply, parCapply, parLapply,
##     parLapplyLB, parRapply, parSapply, parSapplyLB
##
## The following object is masked from 'package:stats':
##
##     xtabs
##
## The following objects are masked from 'package:base':
##
##     anyDuplicated, append, as.data.frame, as.vector, cbind,
##     colnames, do.call, duplicated, eval, evalq, Filter, Find, get,
##     intersect, is.unsorted, lapply, Map, mapply, match, mget,
##     order, paste, pmax, pmax.int, pmin, pmin.int, Position, rank,
##     rbind, Reduce, rep.int, rownames, sapply, setdiff, sort,
##     table, tapply, union, unique, unlist, unsplit
##
## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.
```

```
library(limma)
```

```
##
## Attaching package: 'limma'
##
## The following object is masked from 'package:BiocGenerics':
##
##     plotMA
```

```
library(breastCancerMAINZ)
data(mainz)
```

Getting help.

bioconductor has extensive help available for almost all aspects.

```
?mainz
?ExpressionSet
?limma
```

Of particular note, are the package vignettes, pdf files with tutorial examples of almost all packages.

Examining the mainz breast cancer experiment

http://cancerres.aacrjournals.org/content/68/13/5405.abstract

```
dim(mainz)
```

```
## Features  Samples
##    22283      200
```

```
colnames(pData(mainz))
```

```
##  [1] "samplename"    "dataset"       "series"        "id"
##  [5] "filename"      "size"          "age"           "er"
##  [9] "grade"         "pgr"           "her2"          "brca.mutation"
## [13] "e.dmfs"        "t.dmfs"        "node"          "t.rfs"
## [17] "e.rfs"         "treatment"     "tissue"        "t.os"
## [21] "e.os"
```

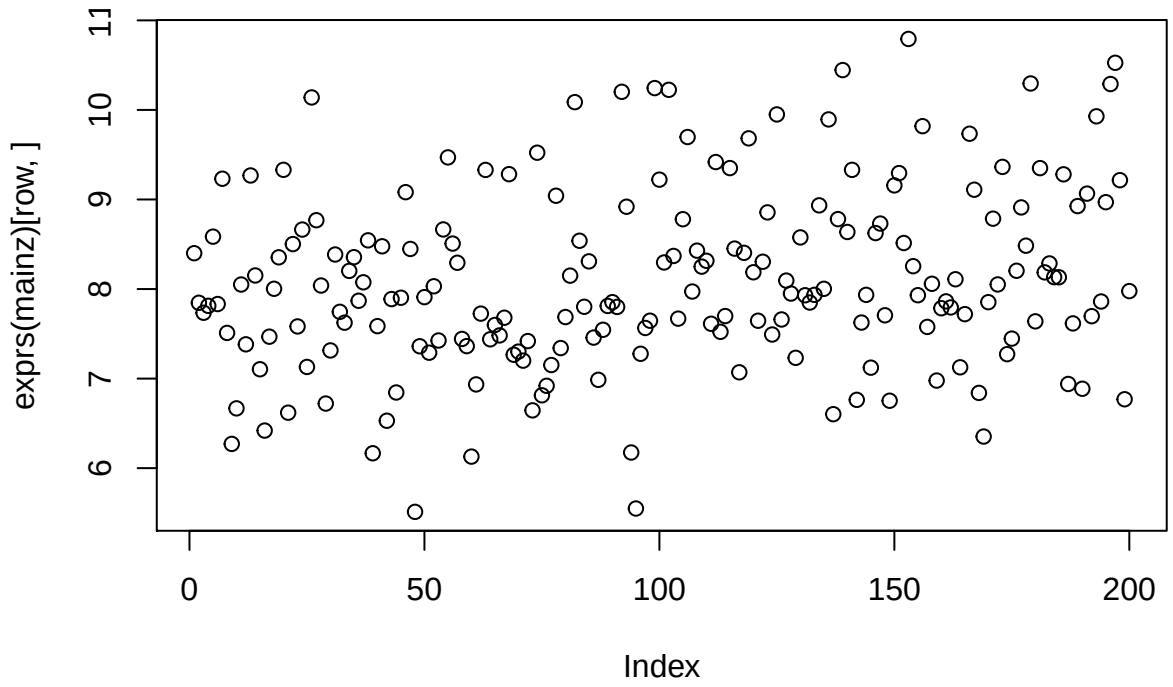
```
annotation(mainz)
```

```
## [1] "hgu133a"
```

Plot some data from mainz

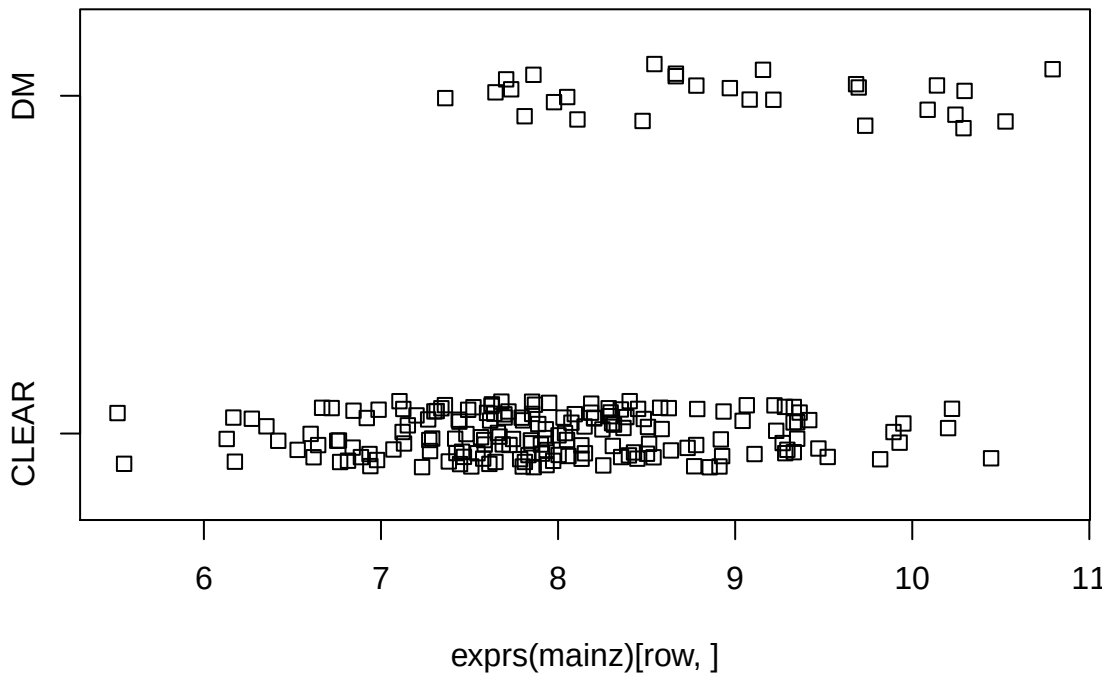
I'll pick a row from the 22,000 probes in the arrays and plot the expression in each array for that gene.

```
row = 21399
plot(exprs(mainz)[row,])
```



Let's divide the arrays into two groups, patients that developed distant metastases within 5 years of diagnosis and those that did not. We can then examine the expression of the selected gene in these groups.

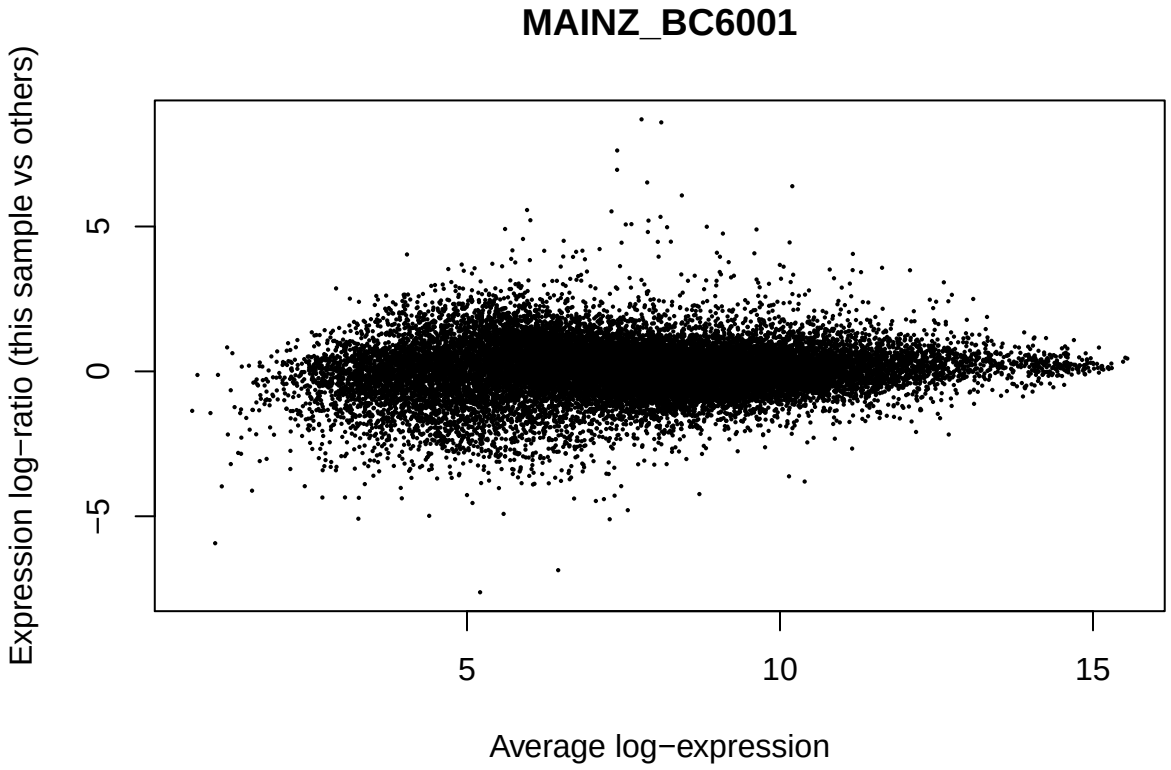
```
mainz.class <- (pData(mainz)$e.dmfs & pData(mainz)$t.dmfs < 5*365)
mainz.fac <- factor(mainz.class, labels= c("CLEAR","DM"))
stripchart(exprs(mainz)[row,] ~ mainz.fac, method="jitter")
```



Plot all the genes

Here is an MA plot of all 22,000+ probes in the first of the 200 experiments compared against the average expression in all the experiments.

```
plotMA(mainz)
```



Fitting a linear model

Suppose we want to find genes that change expression between the cases that have distant metastases in under 5 years from those that don't. Instead of running t-tests on 22,000+ genes, we can fit a model to the data.

The design matrix we use uses the factor we constructed above to divide the arrays into two groups. We can then construct a contrast matrix to compare expression in the patients with distant metastases to those that did not develop metastases.

The `limma` User's Guide contains many case studies with different types of microarray experiments, it is very helpful when designing your own analyses.

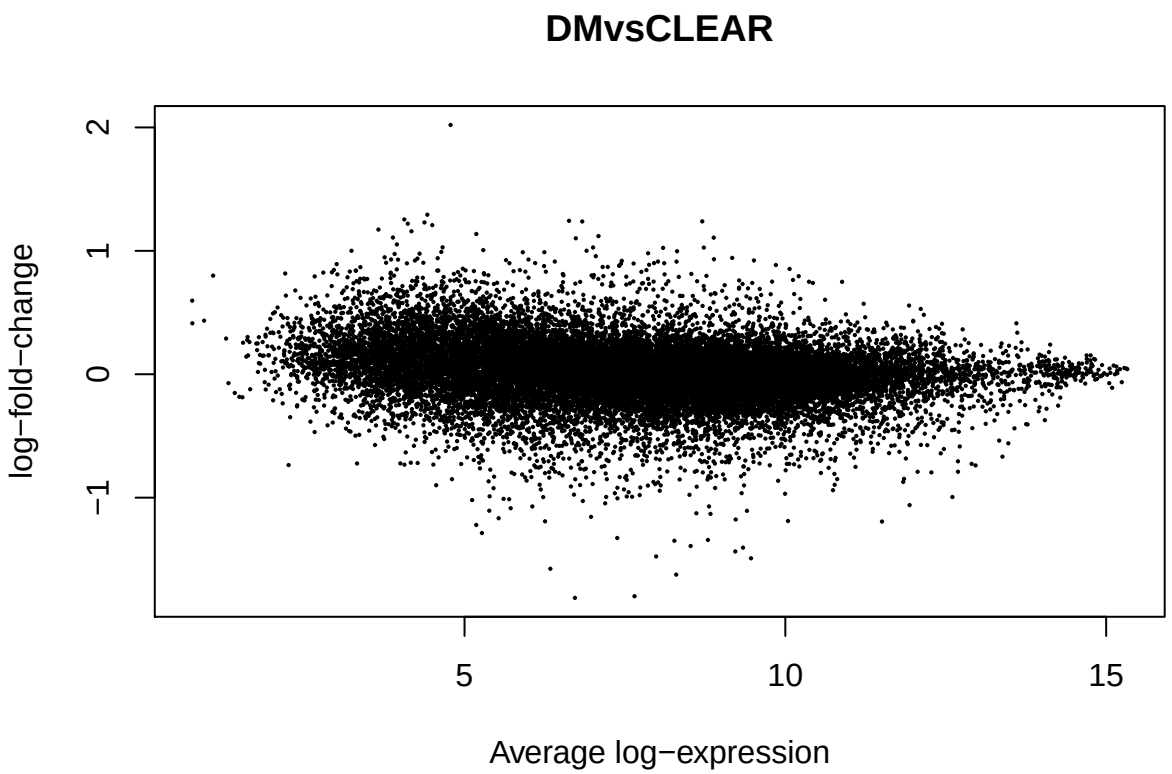
```
design <- model.matrix(~0 + mainz.fac)
colnames(design) <- c("CLEAR", "DM")
fit <- lmFit(exprs(mainz), design)
cont.matrix <- makeContrasts(DMvsCLEAR = DM - CLEAR, levels = design)
fit2 <- contrasts.fit(fit, cont.matrix)
fit.b <- eBayes(fit2)
```

Reporting the results

The resulting `fit.b` object has the results of the linear model fit, and we can produce plots and table summarizing the evidence for differential expression between the conditions.

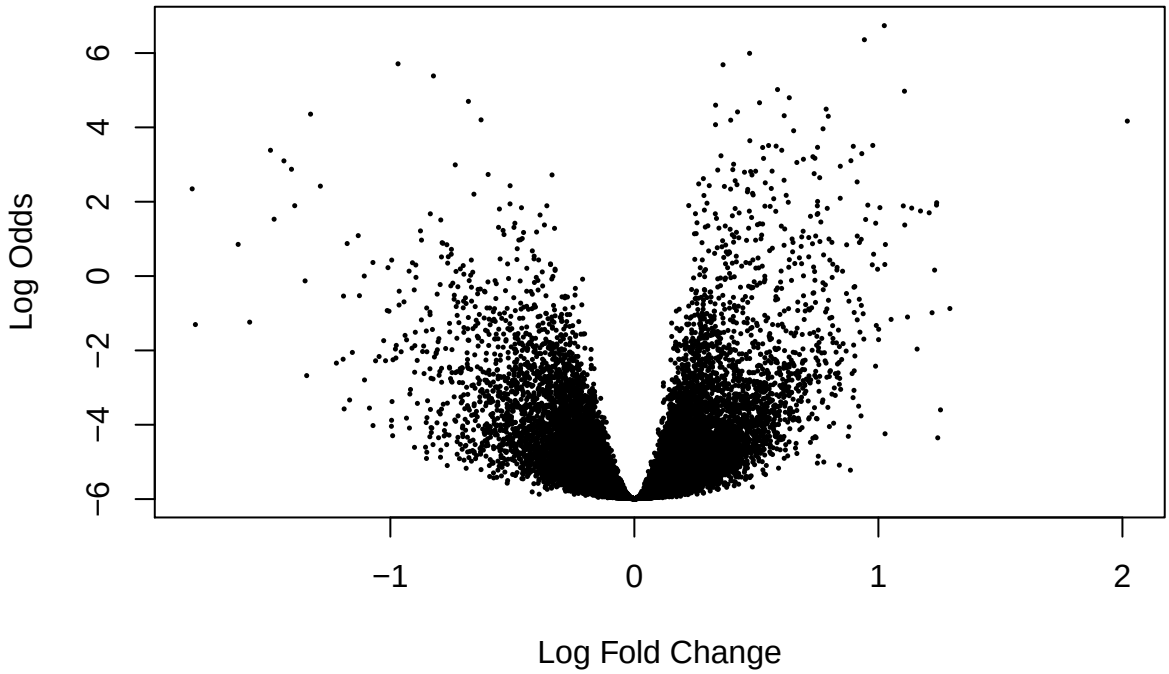
The plot of the fit object shows the size of the predicted effect on each gene.

```
plotMA(fit.b)
```



The volcanoplot shows the log odds of differential expression vs the log ratio of expression.

```
volcanoplot(fit.b)
```



The topTable lists the genes ranked by the evidence of differential expression.

```r
topTable(fit.b, adjust = "BH")
```

```
##                 logFC   AveExpr         t      P.Value   adj.P.Val
## 222039_at    1.0241907  8.094413  5.387081 1.988696e-07 0.002835900
## 218009_s_at  0.9425805  9.172245  5.300403 3.023342e-07 0.002835900
## 57703_at     0.4721307  7.167348  5.215717 4.531487e-07 0.002835900
## 221760_at   -0.9685860  9.995876 -5.149831 6.188802e-07 0.002835900
## 200830_at    0.3632375 11.382618  5.143923 6.363370e-07 0.002835900
## 204156_at   -0.8234357  7.522004 -5.072697 8.882609e-07 0.003298853
## 202412_s_at  0.5866401  8.633824  4.984981 1.333395e-06 0.003808274
## 204825_at    1.1068133  8.883956  4.974329 1.400342e-06 0.003808274
## 213007_at    0.6347787  8.397143  4.931885 1.700912e-06 0.003808274
## 203799_at   -0.6801316 10.016417 -4.907973 1.896838e-06 0.003808274
##                   B
## 222039_at   6.737055
## 218009_s_at 6.358004
## 57703_at    5.992056
## 221760_at   5.710387
## 200830_at   5.685258
## 204156_at   5.384055
## 202412_s_at 5.017503
## 204825_at   4.973321
## 213007_at   4.797999
## 203799_at   4.699734
```

Mapping probe identifiers to external databases.

We can use bioconductor annotation packages to find information for the probes in our dataset.

```r
library(hgu133a.db)
```

```
## Loading required package: AnnotationDbi
## Loading required package: stats4
## Loading required package: GenomeInfoDb
## Loading required package: S4Vectors
## Loading required package: IRanges
##
## Attaching package: 'AnnotationDbi'
##
## The following object is masked from 'package:GenomeInfoDb':
##
##     species
##
## Loading required package: org.Hs.eg.db
## Loading required package: DBI
```

```r
ls("package:hgu133a.db")
```

```
##  [1] "hgu133a"              "hgu133a_dbconn"       "hgu133a_dbfile"
##  [4] "hgu133a_dbInfo"       "hgu133a_dbschema"     "hgu133a.db"
##  [7] "hgu133aACCNUM"        "hgu133aALIAS2PROBE"   "hgu133aCHR"
```

```
## [10] "hgu133aCHRLENGTHS"    "hgu133aCHRLOC"        "hgu133aCHRLOCEND"
## [13] "hgu133aENSEMBL"       "hgu133aENSEMBL2PROBE" "hgu133aENTREZID"
## [16] "hgu133aENZYME"        "hgu133aENZYME2PROBE"  "hgu133aGENENAME"
## [19] "hgu133aGO"            "hgu133aGO2ALLPROBES"  "hgu133aGO2PROBE"
## [22] "hgu133aMAP"           "hgu133aMAPCOUNTS"     "hgu133aOMIM"
## [25] "hgu133aORGANISM"      "hgu133aORGPKG"        "hgu133aPATH"
## [28] "hgu133aPATH2PROBE"    "hgu133aPFAM"          "hgu133aPMID"
## [31] "hgu133aPMID2PROBE"    "hgu133aPROSITE"       "hgu133aREFSEQ"
## [34] "hgu133aSYMBOL"        "hgu133aUNIGENE"       "hgu133aUNIPROT"
```

The package contains mappings from affy probe ids to many different databases. We're interested in the KEGG pathways:

```
x <- hgu133aPATH2PROBE
# Get the probe identifiers that are mapped to a KEGG pathway
mapped_probes <- mappedkeys(x)
# Convert to a list
xx <- as.list(x[mapped_probes])

indices <- ids2indices(xx, rownames(mainz))
```

The `mroast` function tests sets of genes for coordinated changes in expression.

We can run `mroast` to find KEGG pathways with differential expression.

```
res <- mroast(mainz, indices, design)
head(res)
```

```
##       NGenes PropDown PropUp Direction PValue   FDR PValue.Mixed FDR.Mixed
## 01100   1457        0      1        Up  0.001 5e-04        0.001     5e-04
## 05200    619        0      1        Up  0.001 5e-04        0.001     5e-04
## 04010    475        0      1        Up  0.001 5e-04        0.001     5e-04
## 04510    402        0      1        Up  0.001 5e-04        0.001     5e-04
## 04080    380        0      1        Up  0.001 5e-04        0.001     5e-04
## 04810    379        0      1        Up  0.001 5e-04        0.001     5e-04
```

To see the pathways we need a data matrix with ENTREZ Gene ID as the rowname

```
x <- hgu133aENTREZID
# Convert to a list
xx <- as.list(x)
entrezid <- sapply(rownames(fit.b), function(x) xx[x], USE.NAMES=FALSE)
```

Now we can plot the Kegg pathway and color the nodes by their fitted expression estimates.

```
library(pathview)
```

```
## Loading required package: KEGGgraph
## Loading required package: XML
## Loading required package: graph
##
## Attaching package: 'graph'
```

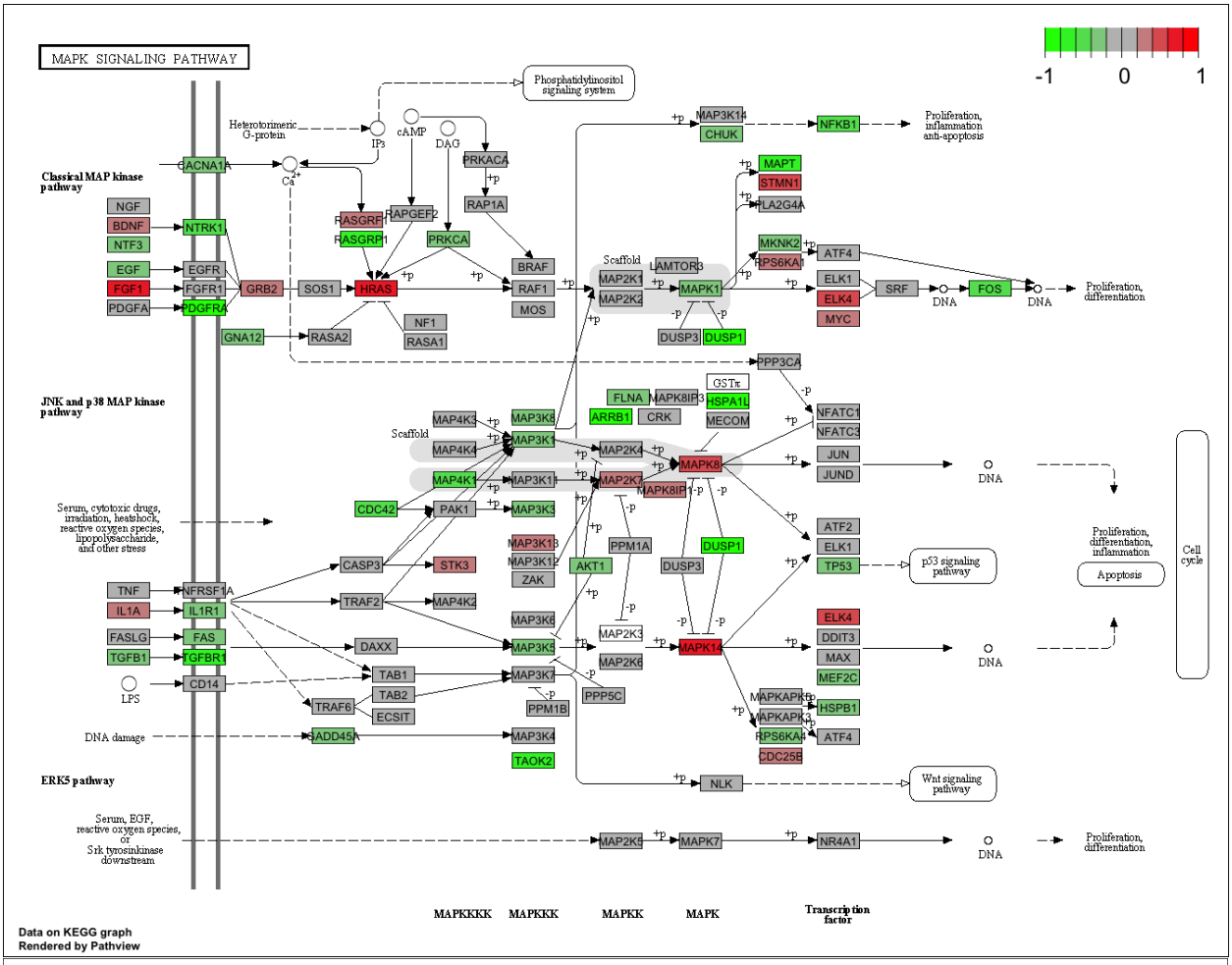
```
##
## The following object is masked from 'package:XML':
##
##     addNode
##
## ###############################################################################
## Pathview is an open source software package distributed under GNU General
## Public License version 3 (GPLv3). Details of GPLv3 is available at
## http://www.gnu.org/licenses/gpl-3.0.html.
##
## The pathview downloads and uses KEGG data. Academic users may freely use the
## KEGG website at http://www.kegg.jp/ or its mirror site at GenomeNet
## http://www.genome.jp/kegg/. Academic users may also freely link to the KEGG
## website. Non-academic users may use the KEGG website as end users for
## non-commercial purposes, but any other use requires a license agreement
## (details at http://www.kegg.jp/kegg/legal.html).
## ###############################################################################
```

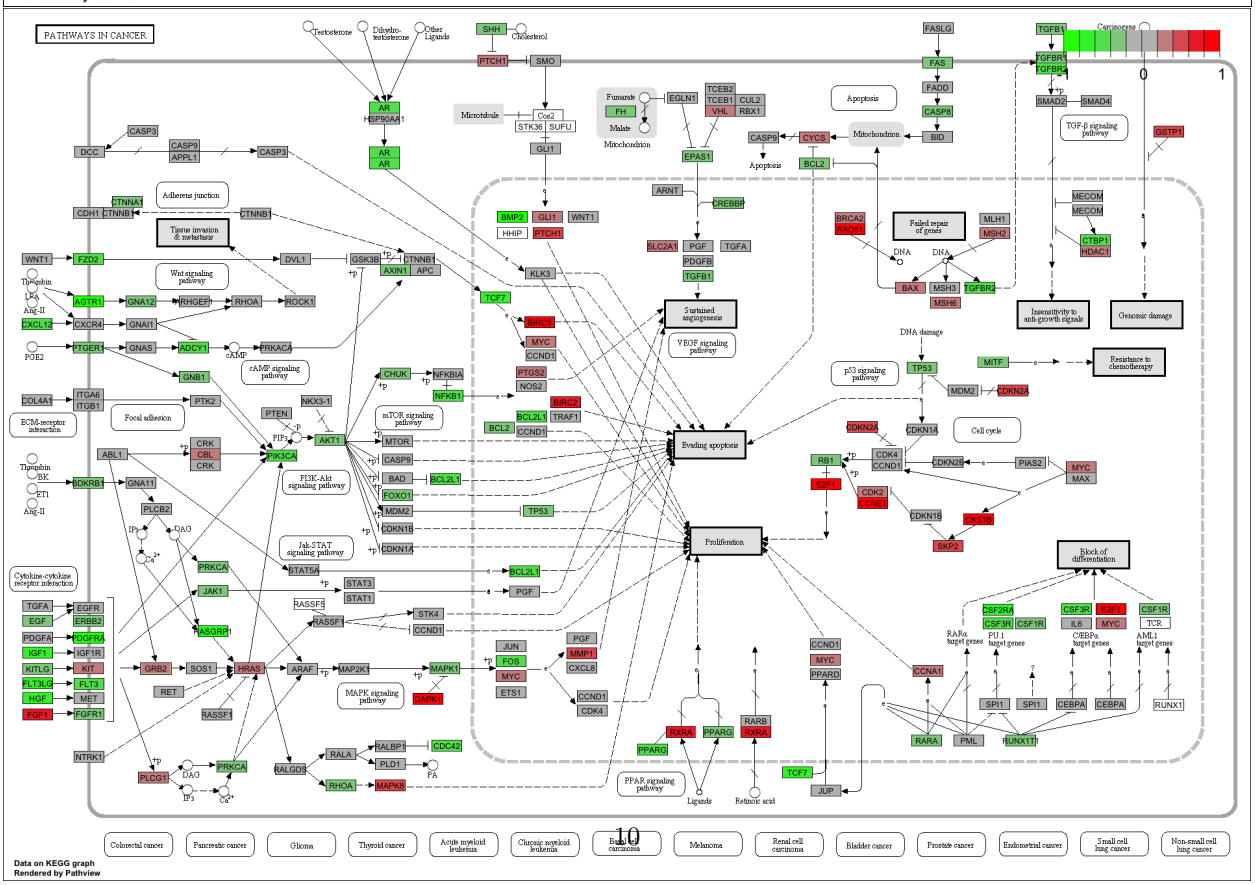
```
gene.data <- fit.b$coefficients
rownames(gene.data) <- entrezid
pv.out <- pathview(gene.data = gene.data, pathway.id = "04010", species = "hsa",
                   out.suffix = "dmvsclear", kegg.native = T, same.layer = F)
```

```
## Info: Working in directory /Users/humberto/Documents/class/2014/expression-workshop
## Info: Writing image file hsa04010.dmvsclear.png
```

```
pv.out <- pathview(gene.data = gene.data, pathway.id = "05200", species = "hsa",
                   out.suffix = "dmvsclear", kegg.native = T, same.layer = F)
```

```
## Info: Working in directory /Users/humberto/Documents/class/2014/expression-workshop
## Info: Writing image file hsa05200.dmvsclear.png
## Info: some node width is different from others, and hence adjusted!
```


MAPK SIGNALING PATHWAY
-1
0
1
Phosphatidylinositol signaling system
Heterotorimeric G-protein
IP3
cAMP
DAG
PRKACA
Classical MAP kinase pathway
CACNA1A
Ca2+
NGF
BDNF
NTF3
NTRK1
EGF
FGF1
PDGFA
EGFR
FGFR1
PDGFRA
GRB2
SOS1
HRAS
RASGRP
RAPGEF2
RAP1A
PRKCA
BRAF
RAF1
MOS
GNA12
RASA2
NF1
RASA1
MAP3K14
CHUK
NFKB1
Proliferation, inflammation anti-apoptosis
MAPT
STMN1
HLA2G4A
MKNK2
RPS6KA1
ATF4
ELK1
ELK4
MYC
SRF
FOS
DNA
Proliferation, differentiation
Scaffold
MAP2K1
MAP2K2
MTOR
MAPK1
DUSP3
DUSP1
PPP3CA
JNK and p38 MAP kinase pathway
GSTπ
FLNA
MAPK8IP3
HSPA1L
ARRB1
CRK
MECOM
NFATC1
NFATC3
JUN
JUND
MAP4K3
MAP4K4
MAP4K1
MAP3K8
MAP3K1
MAP3K3
MAP2K4
MAP2K7
MAPK8
MAPK8IP1
CDC42
PAK1
Serum, cytotoxic drugs, irradiation, heatshock, reactive oxygen species, lipopolysaccharide, and other stress
MAP3K13
MAP3K12
ZAK
PPM1A
AKT1
DUSP1
ATF2
ELK1
TP53
p53 signaling pathway
Proliferation, differentiation, inflammation
Apoptosis
Cell cycle
CASP3
STK3
TNF
IL1A
FASLG
TGFB1
NFRSF1A
IL1R1
FAS
TGFBR1
TRAF2
MAP4K2
DAXX
TAB1
TAB2
ECSIT
TRAF6
CD14
LPS
MAP3K6
MAP3K5
MAP3K7
MAP2K3
MAP2K6
MAPK14
PPP5C
PPM1B
ELK4
DDIT3
MAX
MEF2C
MAPKAPK5
MAPKAPK3
RPS6KA5
HSPB1
ATF4
CDC25B
DNA damage
GADD45A
MAP3K4
TAOK2
ERK5 pathway
NLK
Wnt signaling pathway
Serum, EGF, reactive oxygen species, or Src tyrosinkinase downstream
MAP2K5
MAPK7
NR4A1
Proliferation, differentiation
MAPKKKK
MAPKKK
MAPKK
MAPK
Transcription factor
Data on KEGG graph
Rendered by Pathview


PATHWAYS IN CANCER
Carcinogens
-1
0
1
Testosterone
Dihydro-testosterone
Other Ligands
AR
HSP90AA1
SHH
Cholesterol
PTCH1
SMO
Microtubule
Cos2
STK36
SUFU
GLI1
Fumarate
FH
Malate
Mitochondrion
EGLN1
TCEB2
TCEB1
CUL2
VHL
RBX1
EPAS1
ARNT
CREBBP
Apoptosis
CASP9
CYCS
BCL2
Mitochondrion
FASLG
FAS
FADD
CASP8
BID
TGFB1
TGFBR1
TGFBR2
SMAD2
SMAD4
TGF-β signaling pathway
GSTP1
MECOM
CTBP1
HDAC1
DCC
CASP3
CASP9
APPL1
Adherens junction
CTNNA1
CDH1
CTNNB1
Tight junction & metastasis
WNT1
FZD2
DVL1
GSK3B
AXIN1
APC
Wnt signaling pathway
Thrombin
LPA
Ang-II
AGTR1
GNA12
RHOA
ROCK1
CXCL12
CXCR4
GNAI1
PGE2
PTGER1
GNAS
ADCY1
cAMP
PRKACA
cAMP signaling pathway
GNB1
COL4A1
ITGA6
ITGB1
ECM-receptor interaction
PTK2
Focal adhesion
PTEN
NKX3-1
PIP3
AKT1
PIK3CA
ABL1
CRK
CBL
CRK
EDKRB1
ET1
GNA11
PLCB2
PIP2
DAG
IP3
Ca2+
PRKCA
PI3K-Akt signaling pathway
Jak-STAT signaling pathway
Cytokine-cytokine receptor interaction
JAK1
STAT5A
STAT3
STAT1
TGFA
EGF
PDGFA
IGF1
KITLG
FLT3LG
HGF
FGF1
EGFR
ERBB2
PDGFRA
IGF1R
KIT
FLT3
MET
FGFR1
GRB2
SOS1
HRAS
RET
RASSF1
ARAF
MAP2K1
MAPK1
MAPK signaling pathway
DAPK1
NTRK1
PLCG1
DAG
IP3
PRKCA
RALGDS
RALA
RALBP1
CDC42
PLD1
PA
MAPK8
NFKBIA
NFKB1
mTOR signaling pathway
MTOR
CASP9
BAD
BCL2L1
FOXO1
MDM2
TP53
CDKN1B
CDKN1A
STK4
CCND1
BMP2
GLI1
WNT1
PTCH1
HHIP
KLK3
TCF7
BIRC5
MYC
CCND1
PTGS2
NOS2
BIRC2
BCL2L1
TRAF1
BCL2
CCND1
Evading apoptosis
Sustained angiogenesis
VEGF signaling pathway
SLC2A1
PGF
TGFA
PDGFB
TGFB1
BRCA2
RAD51
Failed repair of genes
MLH1
MSH2
DNA
DNA
BAX
MSH3
TGFBR2
MSH6
Insensitivity to anti-growth signals
Genomic damage
DNA damage
TP53
p53 signaling pathway
MDM2
CDKN2A
MITF
Resistance to chemotherapy
CDKN2A
CDKN1A
Cell cycle
RB1
E2F1
CDK4
CCND1
CDKN2B
PIAS2
MYC
MAX
CDK2
CCNE1
CDKN1B
CKS1B
SKP2
Proliferation
Block of differentiation
RARα target genes
CSF2RA
CSF3R
CSF1R
PU.1 target genes
CSF3R
IL6
CEBPε target genes
E2A1
MYC
CSF1R
TCR
AML1 target genes
SPI1
SPI1
CEBPA
CEBPA
RUNX1
RARA
PML
RUNX1T1
CCNA1
CCND1
MYC
PPARD
RXRA
PPARG
RARB
RXRA
PPARG
PPAR signaling pathway
Ligands
Retinoic acid
TCF7
JUP
JUN
FOS
MYC
ETS1
PGF
MMP1
CXCL8
ECND1
CDK4
BCL2L1
PGF
Colorectal cancer
Pancreatic cancer
Glioma
Thyroid cancer
Acute myeloid leukemia
Chronic myeloid leukemia
Basal cell carcinoma
Melanoma
Renal cell carcinoma
Bladder cancer
Prostate cancer
Endometrial cancer
Small cell lung cancer
Non-small cell lung cancer
Data on KEGG graph
Rendered by Pathview

Bibliography

```
citation()
```

```
##
## To cite R in publications use:
##
##   R Core Team (2014). R: A language and environment for
##   statistical computing. R Foundation for Statistical Computing,
##   Vienna, Austria. URL http://www.R-project.org/.
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {R: A Language and Environment for Statistical Computing},
##     author = {{R Core Team}},
##     organization = {R Foundation for Statistical Computing},
##     address = {Vienna, Austria},
##     year = {2014},
##     url = {http://www.R-project.org/},
##   }
##
## We have invested a lot of time and effort in creating R, please
## cite it when using it for data analysis. See also
## 'citation("pkgname")' for citing R packages.
```

```
citation("breastCancerMAINZ")
```

```
##
## To cite package 'breastCancerMAINZ' in publications use:
##
##   Markus Schroeder, Benjamin Haibe-Kains, Aedin Culhane, Christos
##   Sotiriou, Gianluca Bontempi and John Quackenbush (2011).
##   breastCancerMAINZ: Gene expression dataset published by Schmidt
##   et al. [2008] (MAINZ).. R package version 1.3.1.
##   http://compbio.dfci.harvard.edu/
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {breastCancerMAINZ: Gene expression dataset published by Schmidt et al. [2008]
## (MAINZ).},
##     author = {Markus Schroeder and Benjamin Haibe-Kains and Aedin Culhane and Christos Sotiriou and
##     year = {2011},
##     note = {R package version 1.3.1},
##     url = {http://compbio.dfci.harvard.edu/},
##   }
##
## ATTENTION: This citation information has been auto-generated from
## the package DESCRIPTION file and may need manual editing, see
## 'help("citation")'.
```

```
citation("hgu133a.db")
```

```
## Warning in citation("hgu133a.db"): no date field in DESCRIPTION file of
## package 'hgu133a.db'

##
## To cite package 'hgu133a.db' in publications use:
##
##   Marc Carlson (). hgu133a.db: Affymetrix Human Genome U133 Set
##   annotation data (chip hgu133a). R package version 3.0.0.
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {hgu133a.db: Affymetrix Human Genome U133 Set annotation data (chip hgu133a)},
##     author = {Marc Carlson},
##     note = {R package version 3.0.0},
##   }
##
## ATTENTION: This citation information has been auto-generated from
## the package DESCRIPTION file and may need manual editing, see
## 'help("citation")'.
```

```
citation("Biobase")
```

```
##
##   Bioconductor: Open software development for computational
##   biology and bioinformatics R. Gentleman, V. J. Carey, D. M.
##   Bates, B.Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier,
##   Y. Ge, and others 2004, Genome Biology, Vol. 5, R80
##
## A BibTeX entry for LaTeX users is
##
##   @Article{,
##     author = {Robert C Gentleman and Vincent J. Carey and Douglas M. Bates and {others}},
##     title = {Bioconductor: Open software development for computational biology and bioinformatics},
##     journal = {Genome Biology},
##     volume = {5},
##     year = {2004},
##     pages = {R80},
##     url = {http://genomebiology.com/2004/5/10/R80},
##   }
```

```
citation("limma")
```

```
##
## Please cite the paper below for the limma software itself.  Please
## also try to cite the appropriate methodology articles that
## describe the statistical methods implemented in limma, depending
## on which limma functions you are using.  The methodology articles
## are listed in Section 2.1 of the limma User's Guide.
```

```
##
##   Smyth, GK (2005). Limma: linear models for microarray data. In:
##   'Bioinformatics and Computational Biology Solutions using R and
##   Bioconductor'. R. Gentleman, V. Carey, S. Dudoit, R. Irizarry,
##   W. Huber (eds), Springer, New York, pages 397-420.
##
## A BibTeX entry for LaTeX users is
##
##   @InCollection{,
##     title = {Limma: linear models for microarray data},
##     author = {Gordon K Smyth},
##     booktitle = {Bioinformatics and Computational Biology Solutions Using {R} and Bioconductor},
##     editor = {R. Gentleman and V. Carey and S. Dudoit and R. Irizarry and W. Huber},
##     publisher = {Springer},
##     address = {New York},
##     year = {2005},
##     pages = {397--420},
##   }
```

```
citation("pathview")
```

```
##
## To cite pathview:
##
##   Luo, W. and Brouwer C., Pathview: an R/Bioconductor package for
##   pathway-based data integration and visualization.
##   Bioinformatics, 2013, 29(14): 1830-1831, doi:
##   10.1093/bioinformatics/btt285
##
## A BibTeX entry for LaTeX users is
##
##   @Article{,
##     author = {{Luo} and {Weijun} and {Brouwer} and {Cory}},
##     title = {Pathview: an R/Bioconductor package for pathway-based data integration and visualizatio
##     journal = {Bioinformatics},
##     year = {2013},
##     doi = {10.1093/bioinformatics/btt285},
##     volume = {29},
##     number = {14},
##     pages = {1830-1831},
##   }
##
## This free open-source software implements academic research by the
## authors. Its development took a large amount of extra time and
## effort. If you use it, please support the project by citing the
## listed journal articles.
```