Microarray data analysis using BioConductor

Guiyuan Lei

Centre for Integrated Systems Biology of Ageing and Nutrition (CISBAN)
School of Mathematics & Statistics
Newcastle University
http://www.mas.ncl.ac.uk/~ngl9/

6 June, 2008

Outline

Bioconductor `http://www.bioconductor.org`

- Pre-process data
- Identify differential expression
- Network inference

Pre-process of data

- Entering data into Bioconductor
- Extraction of Cerevisiae probesets
- Exploratory data analysis
- Normalising Microarray data
- Probeset level expression to gene level expression
- Principal Component Analysis

Entering data into Bioconductor

```
library(affy)
fns2 = list.celfiles(path="data2", full.names=TRUE)
rawdata = ReadAffy(filenames=fns2)
print(rawdata)
```

Strain	0 hours	1 hour	2 hours	3 hours	4 hours
Mutant 1	yeast01.cel	yeast02.cel	yeast03.cel	yeast04.cel	yeast05.cel
Wild type 1	yeast06.cel	yeast07.cel	yeast08.cel	yeast09.cel	yeast10.cel
Mutant 2	yeast11.cel	yeast12.cel	yeast13.cel	yeast14.cel	yeast15.cel
Wild type 2	yeast16.cel	yeast17.cel	yeast18.cel	yeast19.cel	yeast20.cel
Mutant 3	yeast21.cel	yeast22.cel	yeast23.cel	yeast24.cel	yeast25.cel
Wild type 3	yeast26.cel	yeast27.cel	yeast28.cel	yeast29.cel	yeast30.cel

Mask file for Cerevisiae probesets

- Mask file to filter out pombe probesets
  `http://www.affymetrix.com/Auth/support/downloads/mask_files/s_cerevisiae.zip`

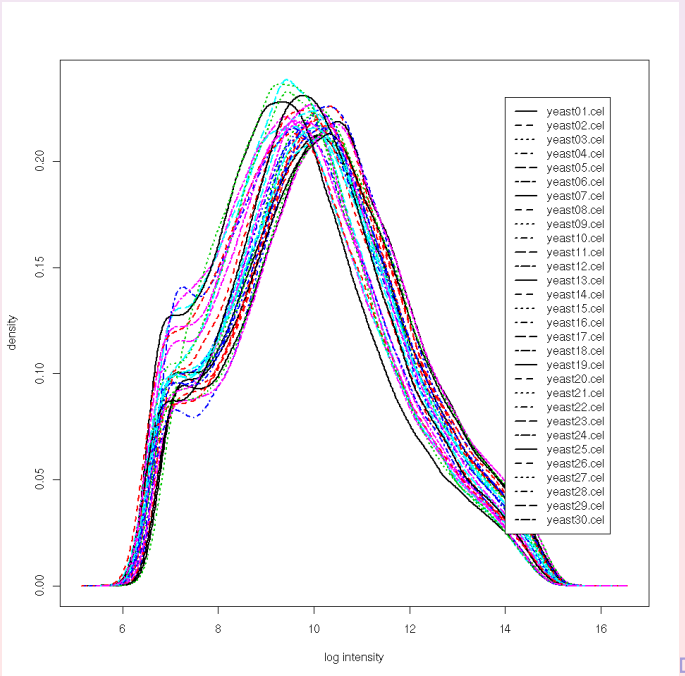
```
s_cerevisiae<-scan("s_cerevisiae.msk", skip=2, list("", ""))
pombe_filter_out<-s_cerevisiae[[1]]
```

- RemoveProbe [1]

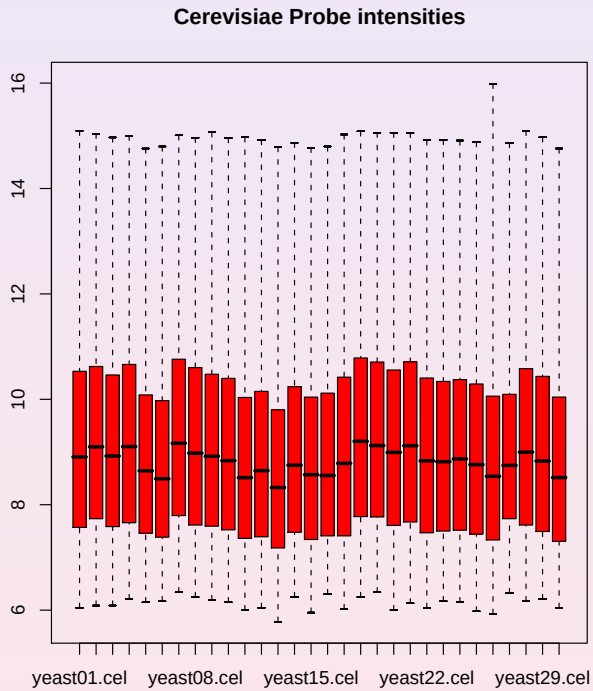
```
source("RemoveProbes.r")
library(yeast2probe)
cleancdf = cleancdfname("yeast2")
RemoveProbes(listOutProbes=NULL, pombe_filter_out,
                          "yeast2cdf","yeast2probe")
```

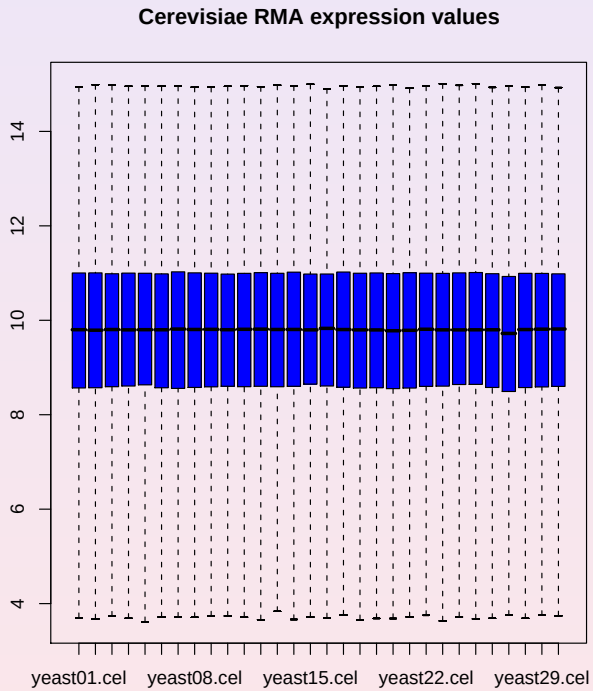
Exploratory data analysis

- Examining raw images
- Probe intensities
- MA plots
- RNA degradation

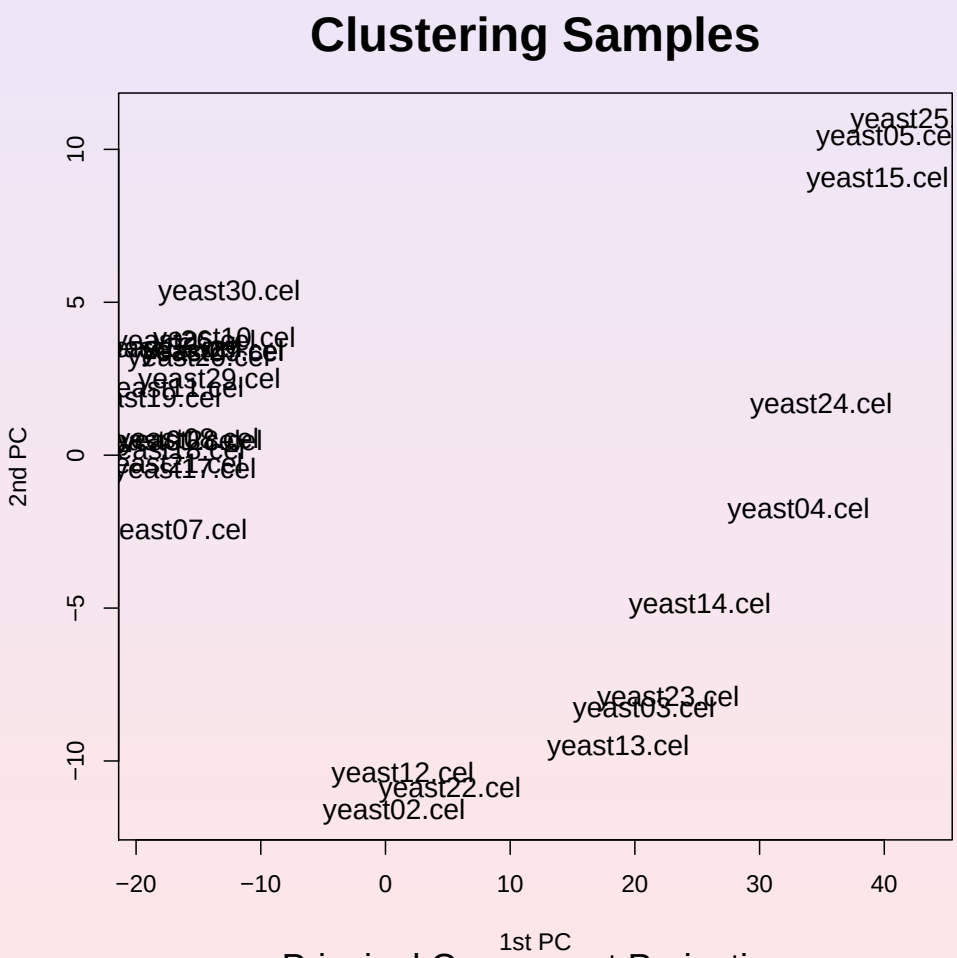


Pre-process of data: Normalisation





Principal Component Analysis [5]



Principal Component Projection

Model Fitting for Identifying Differential Expression

- Limma model [4]
  - Construct design matrix
  - Construct constrasts
- Up-regulated and down-regulated list
- Plot time course for top differential expression
- Volcano Plot for viewing p-value and fold-change
- Heatmap

Up and down regulated list

For identifying differential expression, combine the contrasts by comparing mutant type and wild type at time point 1,2,3 and 4.

```
#Model Fitting
fit<-lmFit(CerevisiaeProbeData, design)
mc<-makeContrasts('m1-w1','m2-w2','m3-w3','m4-w4',levels=design)
fit2<-contrasts.fit(fit, mc)
eb<-eBayes(fit2)
```

Probeset ID	Gene Symbol	T1	T2	T3	T4
ProbesetID 1	Gene 1	-1	-1	-1	-1
ProbesetID 2	Gene 2	1	1	1	1
ProbesetID 3	Gene 3	-1	-1	-1	-1

Table: Up and down regulated list[1]

[1] Not use real gene names here

Plot time course for top differential expression

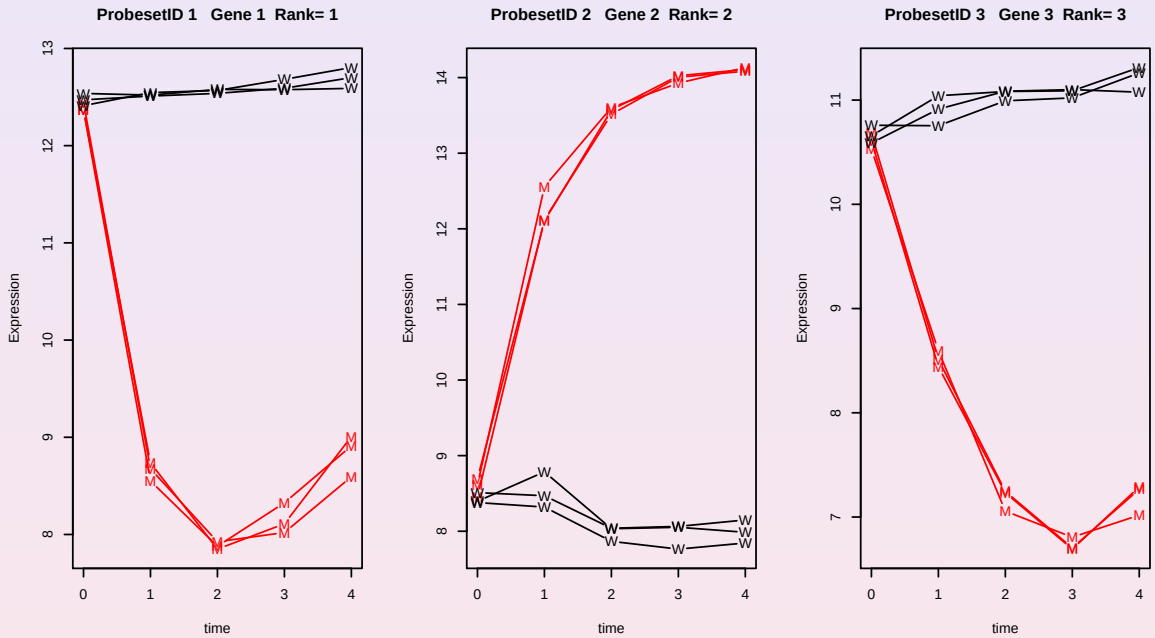


Figure: Time course expression for top 3 differentially expressed Yeast genes[2]

[2]Not use real gene names here

Volcano Plot for viewing p-value and fold-change

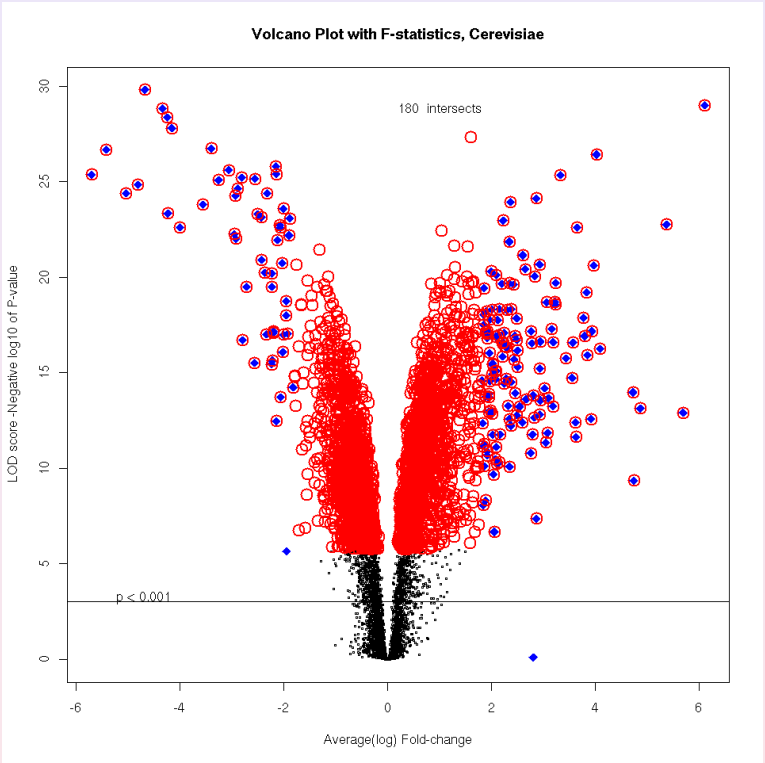
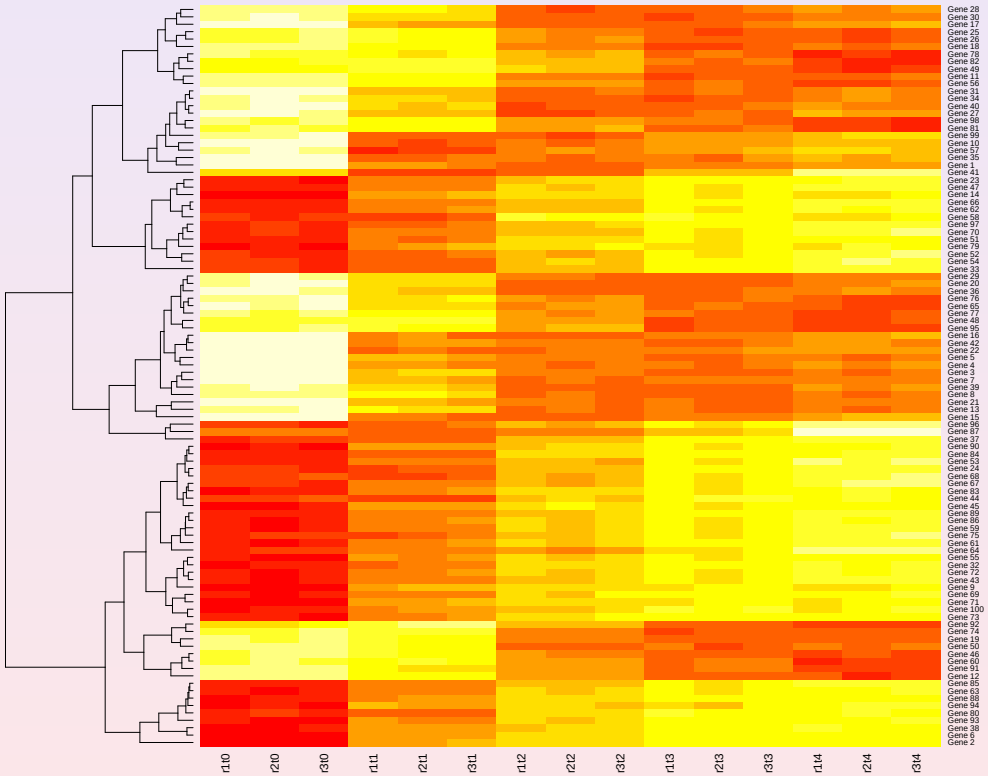


Figure: Bonferroni adjusted p-value less than 0.01 and fold-change larger than 3.5

Heatmap


r1t0
r2t0
r3t0
r1t1
r2t1
r3t1
r1t2
r2t2
r3t2
r1t3
r2t3
r3t3
r1t4
r2t4
r3t4

Network Inference

- GeneNet [2]: simple, quick, but not robust
- Metropolis Hastings for decomposable graphs (MH-d) [3]. Computationaly intensive but more stable
- Reversiable Jump MCMC for time course data

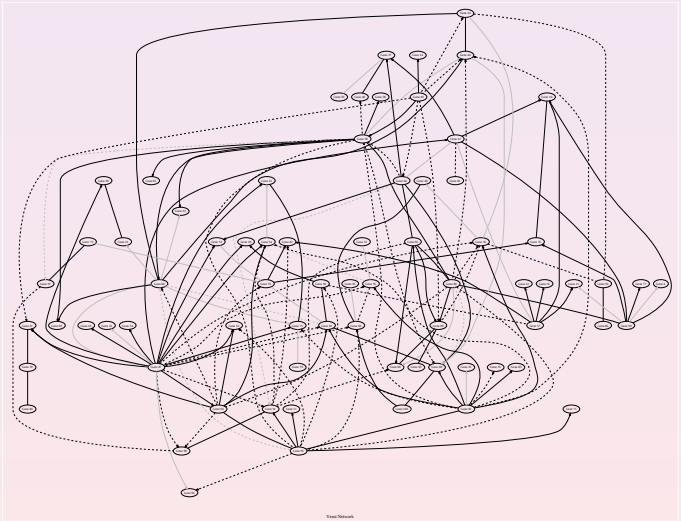


Figure: Inferred network by GeneNet package for top 100 differentially expressed Yeast genes

References


W.G. Alvord, J.A. Roayaei, O.A. Quinones, and K.T. Schneider.
A microarray analysis for differential gene expression in the soybean genome using Bioconductor and R.
Briefings in Bioinformatics, September 29, 2007.


Schafer J. and K. Strimmer.
An empirical bayes approach to inferring large-scale gene association networks.
Bioinformatics, 21:754–764, 2005.


Beatrix Jones, Carlos Carvalho, Adrian Dobra, Chris Hans, Chris Carter, and Mike West.
Experiments in stochastic computation for high-dimensional graphical models.
Statistical Science, 20(4):388–400, 2005.


G.K. Smyth.
Linear models and empirical Bayes methods for assessing differential expression in microarray experiments.
Statistical Applications in Genetics and Molecular Biology, 3, 2004.


E. Wit and J. Mcclure.
Statistics for Microarrays: Design, Analysis and Inference.
Wiley, 2004.