Micorarray data analysis / Bioconductor

Priit Adler (adler@ut.ee)

MTAT.03.239 Bioinformatics **27.10**.2010

Bioconductor

www.bioconductor.org/

– is an **open source**, open development software project to provide tools for the **analysis and comprehension of high-throughput** genomic data. It is based primarily on the **R** programming language.

- /install
- /help
- /developers
- /about

www.bioconductor.org/install

Install Bioconductor / Packages

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

>biocLite("lumi")    #illumina pre-processing package
>biocLite("expresso") #affy pre-processing wrapping
```

www.bioconductor.org/help

- Release notes
  - BiocViews – *package discovery*
- Workflows
  - Oligonucleotide Arrays
- Vignettes – *task-oriented descriptions of package functionality*

  >openVignette()

  >browseVignettes()
- Mailing Lists
- …

BiocViews

Bioconductor version 2.7 (Release)

- AnnotationData (510)
- ExperimentData (67)
- Software (408)
  - Annotation (55)
  - AssayDomains (161)
  - AssayTechnologies (254)
    - FlowCytometry (13)
    - HighThroughputSequencing (25)
    - MassSpectrometry (10)
    - Microarray (206)
      - MultiChannel (1)
      - OneChannel (55)
      - TwoChannel (38)
    - MicrotitrePlateAssay (1)
    - SAGE (6)
    - Sequencing (8)
  - Bioinformatics (240)
  - BiologicalDomains (39)
  - Infrastructure (176)

Packages

Software > AssayTechnologies > Microarray > OneChannel

ABarray • affy • affycomp • AffyCompatible • affycoretools • AffyExpress • affyILM • affylmGUI
affypdnn • affyPLM • affyQCReport • Agi4x44PreProcess • AgiMicroRna • altcdfenvs • annaffy
aroma.light • ArrayExpress • arrayQualityMetrics • ArrayTools • beadarray • bridge • clippda
codelink • copa • exonmap • gage • gcrma • GeneRegionScan • GEOquery • GlobalAncova • globaltest
lapmix • limma • lumi • LVSmiRNA • maDB • makecdfenv • makePlatformDesign • oligo
oneChannelGUI • plw • PROMISE • puma • RefPlus • RLMM • RTools4TB • SAGx • simpleaffy • Starr
tilingArray • vsn • webbioc • xmapcore • xps • yaqcaffy

Bioconductor Workflows

OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Using Bioconductor for Microarray Analysis

Bioconductor has advanced facilities for analysis of microarray platforms including Affymetrix, Illumina, Nimblegen, Agilent, and other one- and two-color technologies.

Bioconductor includes extensive support for analysis of expression arrays, and well-developed support for exon, copy number, SNP, methylation, and other assays.

Major workflows in Bioconductor include pre-processing, quality assessment, differential expression, clustering and classification, gene set enrichment analysis, and genetical genomics.

Bioconductor offers extensive interfaces to community resources, including GEO, ArrayExpress, Biomart, genome browsers, GO, KEGG, and diverse annotation sources.

Sample Workflow

The following psuedo-code illustrates a typical R / Bioconductor session. It uses RMA from the `affy` package to pre-process Affymetrix arrays, and the `limma` package for assessing differential expression.

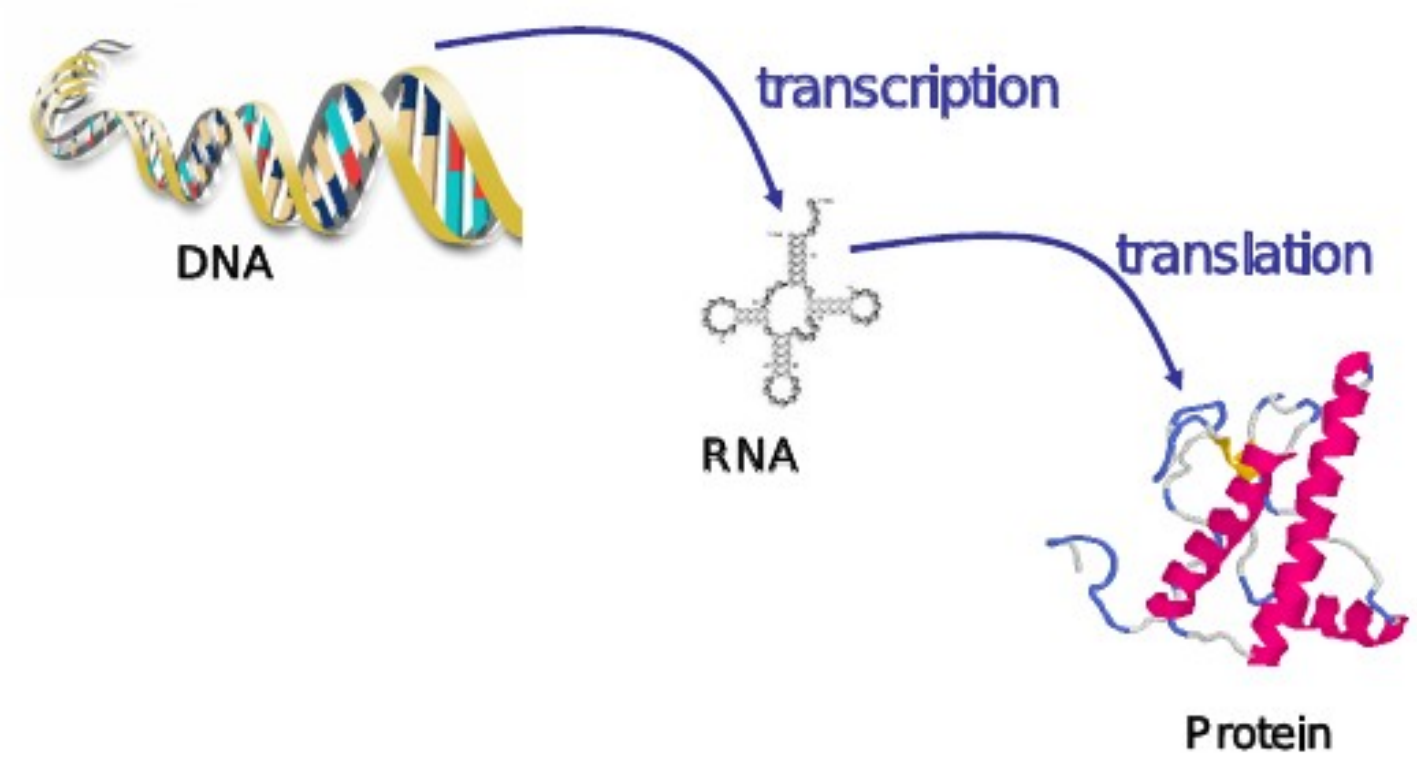
```
## Load packages
> library(affy)    # Affymetrix pre-processing
> library(limma)   # two-color pre-processing; differential
                   # expression

## import "phenotype" data, describing the experimental design
> phenoData <- read.AnnotatedDataFrame("sample-description.csv")

## RMA normalization
> eset <- justRMA("/celfile-directory", phenoData=phenoData)
```

Molecular biology

Central dogma of molecular biology



http://www.cs.helsinki.fi/bioinformatiikka/mbi/courses/06-07/pcmda/slides/Microarrays_Brazma_lecture1.pdf

Gene Expression

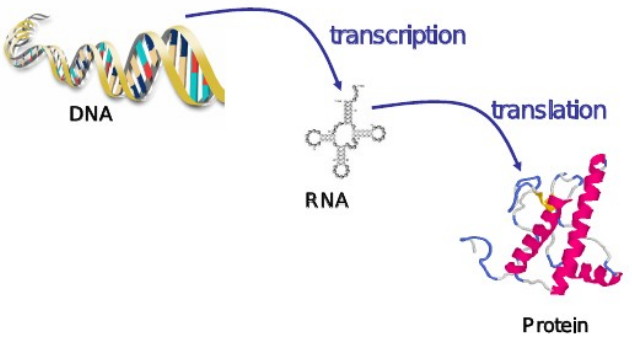
A human organism has over 250 different cell types (e.g. skin, bone, blood, etc)

- most have identical genomes, yet quite different look and function

It is believed that less than 20% of the genes are "expressed" in a typical cell type

Does the gene expression make the cells different ?

– apparently **yes**



http://www.cs.helsinki.fi/bioinformatiikka/mbi/courses/06-07/pcmda/slides/Microarrays_Brazma_lecture1.pdf

Some questions for the golden age of genomics

- How gene expression differs in different cell types ?
- How gene expression differs in normal vs. diseased cell ? (cancer)
- How gene expression changes occur during organisms life span
- How gene expression is regulated – which genes regulate which and how ?

http://www.cs.helsinki.fi/bioinformatiikka/mbi/courses/06-07/pcmda/slides/Microarrays_Brazma_lecture1.pdf

So how do we answer those questions ?

- measure gene expression !

How can we measure gene expression in a living cell ?

- we don't ! but we can do something really similar using expression mircoarrays !

Micorarrays

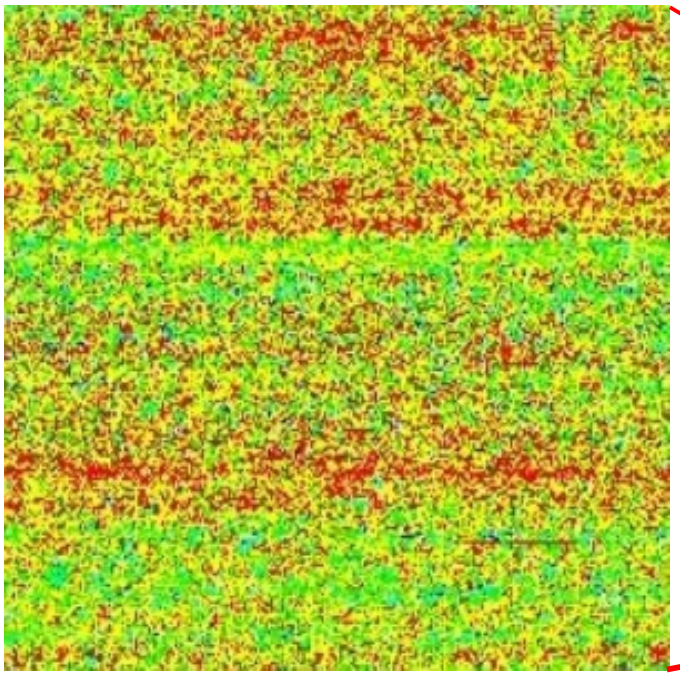
Microarray Types

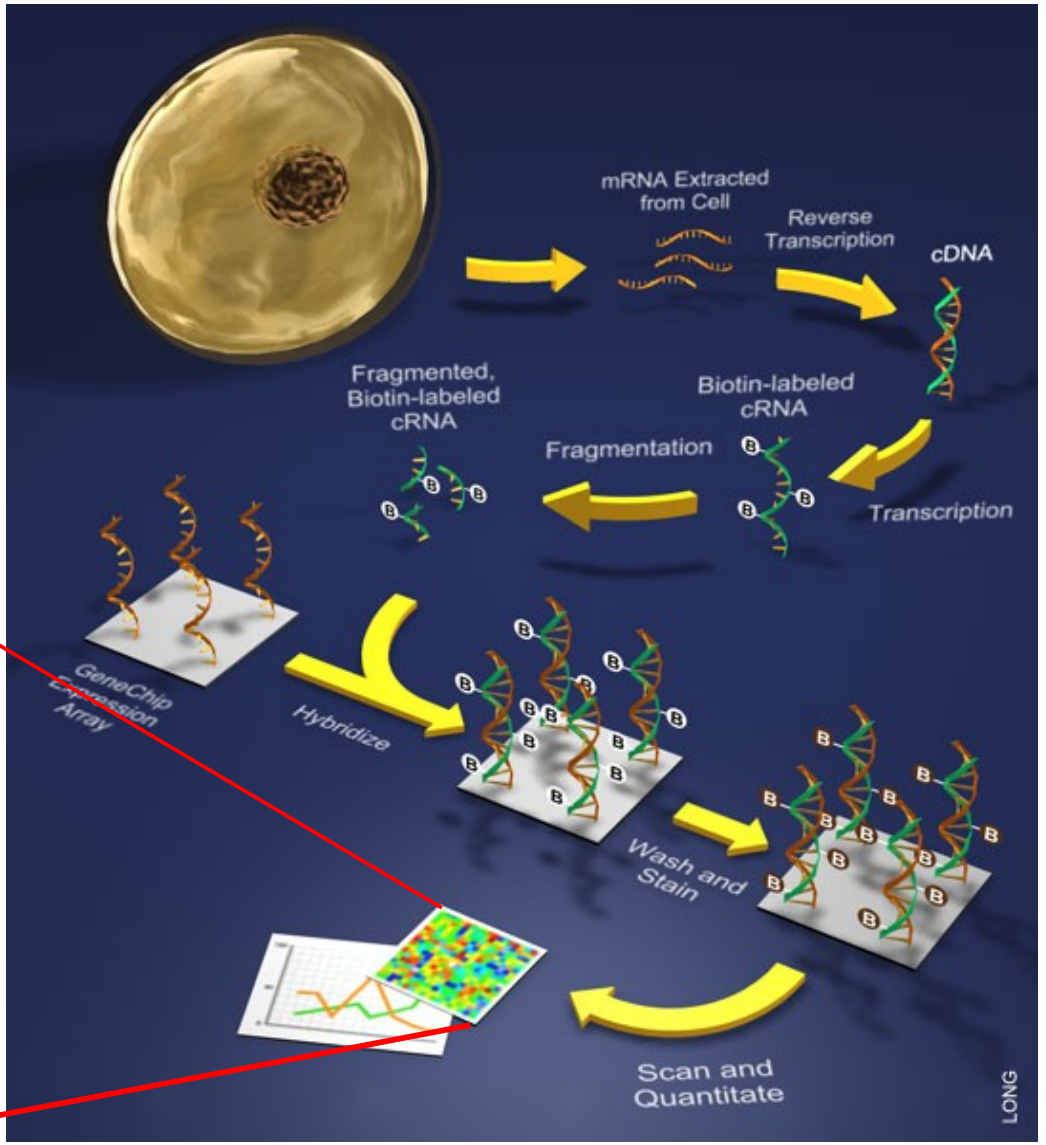
- Expression
  - **Single channel** (e.g., Affymetrix) – *affy, etc*
  - Two channel (e.g., Aligent / Genepix) – *limma*
  - **Bead array** (e.g., Illumina) – *lumi, beadarray*
  - Long oligo (Nimblegen) – *oligo*
- Exon
- Array CGH
- Genotyping, e.g., SNP
- Methylation

http://www.bioconductor.org/help/course-materials/2010/SeattleJan10/day2/PreProcessing.pdf

Expression microarrays : single channel

Affymetrix GeneChip





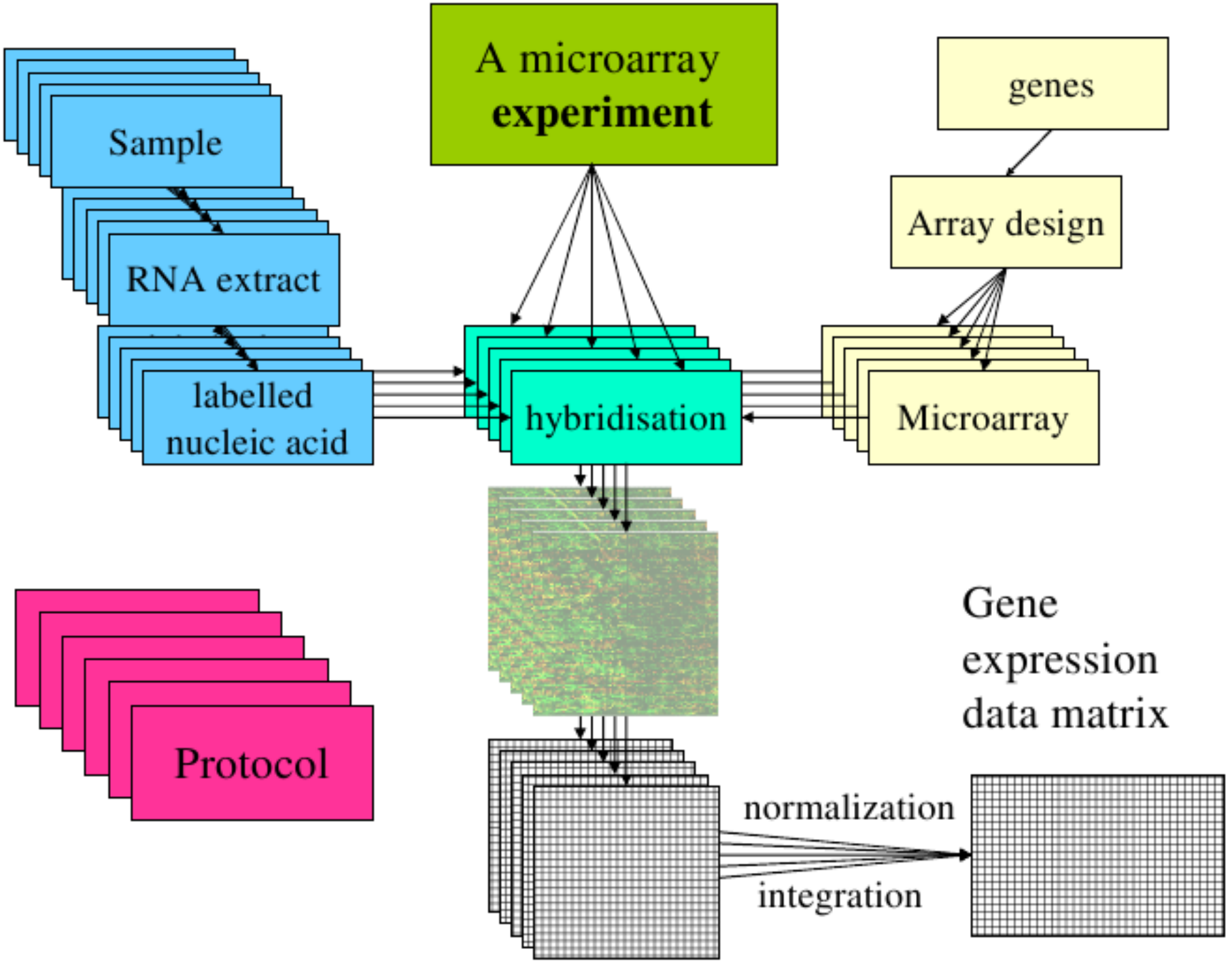
A microarray experiment

Normally it will be more than one array per "experiment"

- More than 2 conditions can be compared
- The same condition can be used on array many times (replicate experiments) to fin out what is the 'noise level' or natural gene expression variability within the same experiment

http://www.cs.helsinki.fi/bioinformatiikka/mbi/courses/06-07/pcmda/slides/Microarrays_Brazma_lecture1.pdf

http://www.cs.helsinki.fi/bioinformatiikka/mbi/courses/06-07/pcmda/slides/Microarrays_Brazma_lecture1.pdf



Pre-processing (single channel)

Background correction

- PM / MM probes

Normalization

- Key assumption: most probes are not differentially expressed; distribution of intensities is approximately equal across arrays

Summarization

- from probes to probesets (approximately, genes)

http://www.bioconductor.org/help/course-materials/2010/SeattleJan10/day2/PreProcessing.pdf

Pre-processing in Bioconductor : Affymetrix GeneChip

- Quality control
  - **ArrayQualityMetrics**
- Normalization
  - MAS 5.0
  - **RMA**
  - gcRMA
  - farms
  - etc ...

Pre-processing in Bioconductor : Affymetrix GeneChip

In practice:

```
>setwd("/path/to/the/cel/files/")
>library(affy)
>eset = just.rma()
>expData = exprs(eset)
```

- also: *just.gcrma*
- *expresso* – for more flexible control;
- *http://www.bioconductor.org/help/workflows/* for common analyses methods

http://www.bioconductor.org/help/course-materials/2010/SeattleJan10/day2/PreProcessing.pdf

Pre-processing in Bioconductor : Example: RMA (robust multi-chip average)

Background correction

- Observation: using MM probes is problematic when MM >PM
- Model PM probes as exponentially distributed signal, plus normal noise, $\exp(\alpha) + N(\mu,\sigma^2)$.

Normalization

- Quantile normalization – force the distribution of background-corrected expression values of each array to have exactly the same distribution.

Summarization

- Estimate probeset effect by fitting a linear model to all probes in each probe set, across array.

http://www.bioconductor.org/help/course-materials/2010/SeattleJan10/day2/PreProcessing.pdf

Pre-processing in Bioconductor : Quality assessment

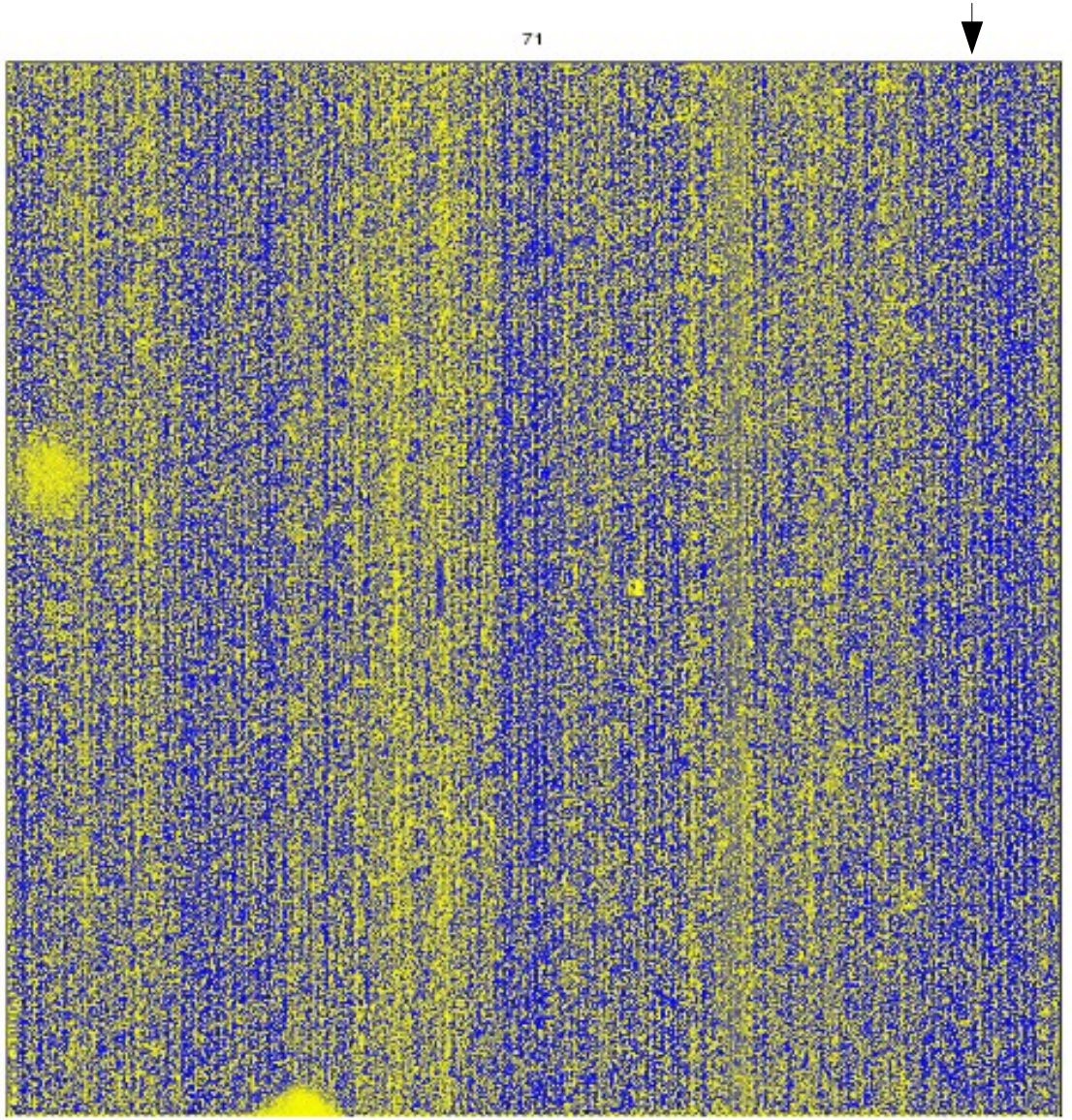
In practice:

```
>library(arrayQualityMetrics)
>rpt = arrayQualityMetrics(abatch)
>## or, as appropriate,
>## rpt = arrayQualityMetrics(eset / rg)
>browseURL(rpt)
```

- QC summary statistics: acceptable ranges for 'control' probes
- Between-array distances: no unintended associations with experimental conditions, e.g., run date.
- NUSE (normalized unscaled standard error) and RLE (relative log expression) plots: consistent expression and variability

http://www.bioconductor.org/help/course-materials/2010/SeattleJan10/day2/PreProcessing.pdf

Array #	Array Name	MA-plot	Spatial distribution	Boxplots/Density plots	Heatmap	RLE	NUSE
70	DNA9042-070_20471201.CEL.gz						
71	DNA9042-071_20461201.CEL.gz			*	*		
72	DNA9042-072_20351201.CEL.gz						*



Pre-processing in Bioconductor : Illumina Expression BeadChip

- Quality assessment & Normalization
  - lumi
  - beadarray

Pre-processing in Bioconductor : Illumina Expression BeadChip

In practice:

```
>setwd("/path/to/the/BeadStudio/output/files/")
>library(lumi)
>x.lumi = lumiR.batch(dir())
>lumi.N.Q = lumiExpresso(x.lumi,
QC.evaluation=TRUE)
>expData = exprs(lumi.N.Q)
```

- *http://www.bioconductor.org/help/workflows/* for common analyses methods

Pre-processing
final touch (optional)

- Filter out genes with
  - low expression
  - low expression variance
- Centering the data
- Averaging expression values (for technical replicates)

Differential expression

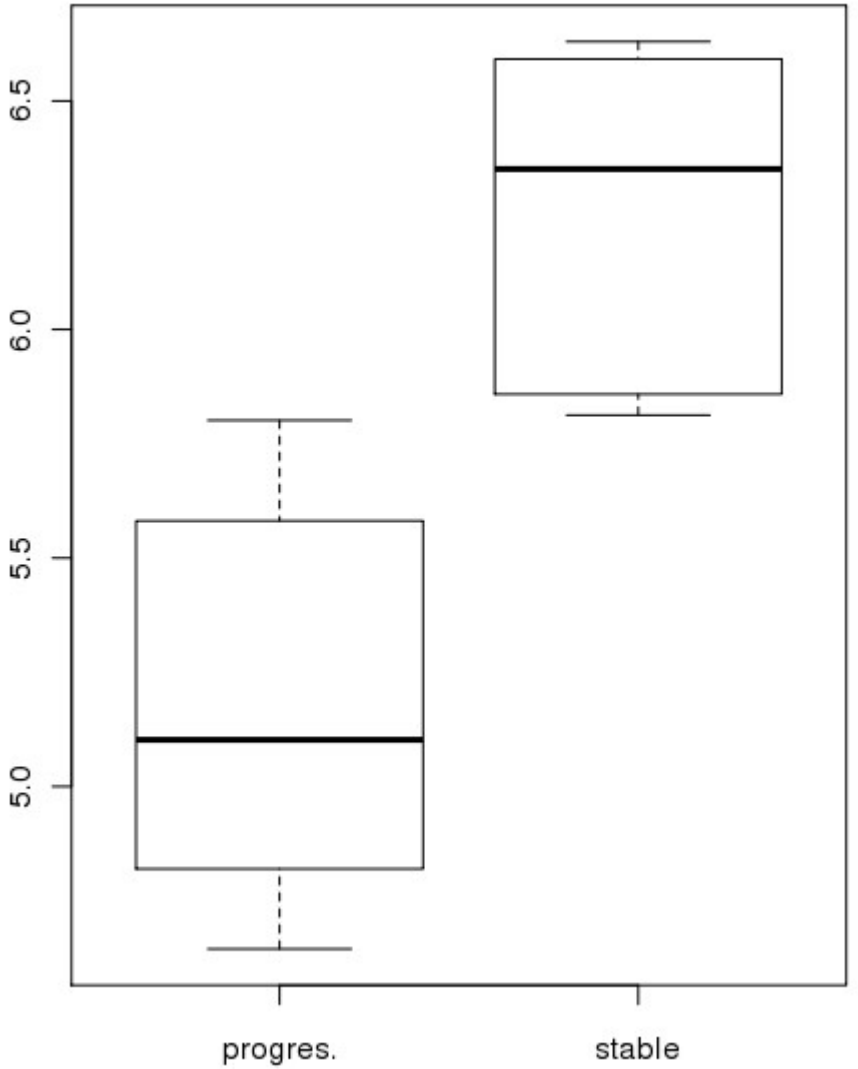
Find genes with different expression between conditions

- t-test and derivates

$$t = \frac{\mu_1 - \mu_2}{\sigma}$$

- ANOVA

$$e = \alpha + b_1\mu_1 + b_2\mu_2 + b_3\mu_3 + \epsilon$$



Differential expression

In practice:

```
>##standard and simple
>library(genefilter)
>groups = factor(sample_classification_vector)
>tt.res = rowttests(expData, groups)
>tt.res$p.value = p.adjust(tt.res$p.value, “bonf”)
>#use t.test and apply function
```

Differential expression

- Using *limma* package
  - Use design matrix to establish parameters of the model: *model.matrix*
  - Use linear model to fit the contrast parameters: *lmFit()*
  - Use function *eBayes* to get moderate *t*-statistics and relevant statistics

http://www.bioconductor.org/help/course-materials/2010/SeattleJan10/day2/DifferentialExpression.pdf

Differential expression

In practice:

```
>library(limma)
>mm = model.matrix(~ sample_classification_vector - 1)
>colnames(mm) = c("normal", "bad")
>fit1 = lmFit(expData, mm)
>contr = makeContrasts(normal - bad, levels=colnames(mm))
>fit = contrasts.fit(fit,contr)
>fit = eBayes(fit)
>dT = decideTests(fit, adjust.method="fdr", p.value=0.05)
>tT = topTable(fit, coef = "normal - bad", …)
```

Co-expression

Find similarly behaving genes using correlation or distance metrics

use *dist()* for distance measures in R

use *cor()* for correlation measures in R

Unsupervised data exploration – clustering

use *hclust()* for hierarchical clustering– requires as.dist object

use *kmeans()* for k-means clustering

Functional analysis

- Gene Ontology enrichment analysis
  - In WEB – http://biit.cs.ut.ee/gprofiler
    - automatic gene-id conversion
    - submit a list and You'll have the result in seconds
  - In Bioconductor – GOstats
    - figure out the correct data structure
    - find Your correct gene id mappings
    - query one ontology at a time
    - on other hand GOstats is more flexible than gprofiler
      - e.g. in addition to over representation also under representation

BIIT web resources

biit.cs.ut.ee

/gprofiler
/mem
/kegganim
/vishic
/expressview
/diffexp
/graphweb

Lab

Use bioconductor website and vignettes as Your guide and help to tackle to problems.

1. follow pre-processing lab exercises 1-4 by Martin Morgan

http://www.bioconductor.org/help/course-materials/2010/SeattleJan10/day2/PreProcessingLab.pdf

running arrayQualityMetric would probably kill Your computer - the report is available at
http://biit.cs.ut.ee/~adler/bioinf.03.239/CLL/QMreport.html

2. Calculate differential expression between “progressive” and “stable” disease variants. How many differentially expressed probesets did You get ? which pipeline and parameters did You used to get them ? play a little with the parameters (alternative p.value corrections)

3. Find 50 most similarly expressed probesets to a differentially expressed probeset of Your choice. which probeset did You use ? which distance / correlation measure did You use ? try different measures! are they stable ?

4. Perform GO enrichment analysis using GOstats and gprofiler with the list of Your 51 genes. which categories are enriched ? are the found categories helping to understand the disease variants ? try different queries (with other gene as the seed / all diff exp genes)