7. GENE EXPRESSION ANALYSIS
MICROARRAYS

BIOINFORMATICS COURSE
MTAT.03.239

24.10.2012
GENE EXPRESSION ANALYSIS
MICROARRAYS

Slides adapted from Konstantin Tretyakov’s 2011/2012 and Priit Adlers 2010/2011 year slides
SCIENTIFIC PROCESS

Observation → Analysis → Generalization
SCIENTIFIC PROCESS

Observation → Analysis → Generalization

“Gene expression analysis - microarrays”
Bioinformatics Course
CONTEMPORARY SCIENCE

http://www.sciencephoto.com
CONTEMPORARY SCIENCE

Observation → Analysis → Generalization

“Gene expression analysis - microarrays"
Bioinformatics Course
CONTEMPORARY BIOINFORMATICS

Observation → Analysis → Generalization

Nucleotide sequences
Molecular interactions
Gene expression
Heredity
Drug effects
...

“Gene expression analysis - microarrays"
Bioinformatics Course
CONTEMPORARY BIOINFORMATICS

Observation → Analysis → Generalization

- Nucleotide sequences
- Molecular interactions
- Gene expression
- Heredity
- Drug effects
- ...

“Gene expression analysis - microarrays”
Bioinformatics Course
FLOW OF GENETIC INFORMATION

http://www.nature.com/scitable/topicpage/gene-expression-14121669

"Gene expression analysis - microarrays"
Bioinformatics Course
If the genetic information in the cells is always the same, what makes the difference between kidney cells and brain cells???
GENE EXPRESSION
is the presence of the gene’s product in the cell in the form of a protein or mRNA
FLOW OF GENETIC INFORMATION

Gene expression analysis - microarrays
Bioinformatics Course
MODULATION OF TRANSCRIPTION

Transcription factor (TF)

Is a protein that binds to specific DNA sequence, thereby controlling the flow (or transcription) of genetic information from DNA to mRNA.

http://www.nature.com/scitable/topicpage/gene-expression-14121669
MODULATION OF TRANSCRIPTION

Transcription factor (TF)

Is a protein that binds to specific DNA sequence, thereby controlling the flow (or transcription) of genetic information from DNA to mRNA.

http://www.nature.com/scitable/topicpage/gene-expression-14121669
TRANSCRIPTIONAL REGULATION DETERMINES CELL FATE

http://www.nature.com/scitable/topicpage/gene-expression-14121669
WHY MEASURE GENE EXPRESSION?

DIFFERENCES BETWEEN TISSUES

BRAIN CELL  STEM CELL  BLOOD CELL

“Gene expression analysis - microarrays"
Bioinformatics Course
WHY MEASURE GENE EXPRESSION?

UNDERSTANDING DISEASES


"Gene expression analysis - microarrays"
Bioinformatics Course
WHY MEASURE GENE EXPRESSION?

RESPONSE TO DIFFERENT CONDITIONS


WHY MEASURE GENE EXPRESSION?

RESPONSE TO DIFFERENT CONDITIONS

http://www.guardian.co.uk/science/blog/2011/apr/21/real-water-added-electrons
The Nobel Prize in Physiology or Medicine 2012
Sir John B. Gurdon, Shinya Yamanaka

The Nobel Prize in Physiology or Medicine 2012 was awarded jointly to Sir John B. Gurdon and Shinya Yamanaka "for the discovery that mature cells can be reprogrammed to become pluripotent"
QUESTIONS FOR GENE EXPRESSION

- 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?
- How gene expression changes when a cell is treated by a drug?

TECHNIQUES

- Polymerase chain reaction (PCR)
- Real - Time PCR (qPCR)
- Serial Analysis of Gene Expression (SAGE)
- Northern Blot
- Microarrays
- RNA - seq
TECHNIQUES – SINGLE GENE

- Polymerase chain reaction (PCR)
- Real – Time PCR (qPCR)
- Serial Analysis of Gene Expression (SAGE)
- Northern Blot
- Microarrays
- RNA – seq
TECHNIQUES – LARGE SCALE

- Polymerase chain reaction (PCR)
- Real – Time PCR (qPCR)
- Serial Analysis of Gene Expression (SAGE)
- Northern Blot
- Microarrays
- RNA – seq
TECHNIQUES – LARGE SCALE
HOW TO MEASURE GENE EXPRESSION?

DNA

RNA

PROTEIN
RNA EXPRESSION

DNA

RNA

PROTEIN

"Gene expression analysis - microarrays"
Bioinformatics Course
RNA EXPRESSION

DNA → RNA → PROTEIN

"Gene expression analysis - microarrays"
Bioinformatics Course
MICROARRAYS

RNA
APPLICATION OF MICROARRAYS

- Gene discovery
  
  Identification of new genes and their function and level in different conditions.

- Molecular classification of complex diseases
  
  To classify the different cancer types based on the patterns of gene activity in the tumor cells, to develop drugs specifically designed for these cancers.

- Drug discovery
  
  Comparative analysis of the genes from a diseased and a normal cell help the identification of the biochemical constitution of the proteins synthesized by the diseased genes. This information can be used to synthesize drugs that combat with these proteins and reduce their effect.

- Toxicological research
  
  Microarray technology provides a robust platform to study the effect of different toxins on the cells and it’s progenitors.

http://estbioinfo.stat.ub.es/docs/jmosquera/teaching/BIOINF/MDALab/Lecture1-IntroductionToMicrarrayDataAnalysis.pdf
APPLICATION OF MICROARRAYS

PERSONALIZED MEDICINE

http://tristar3research.wordpress.com/2009/04/20/the-promise-of-personalized-medicine-hype-or-hope/

"Gene expression analysis - microarrays"
Bioinformatics Course
HYBRIDIZATION

- The process of establishing sequence-specific interaction between two complementary strands of nucleic acid into a single complex.

...TGGGGATGCCCCGAATAGCATGGACTAGCTGGA...
...ACCCCTACGGGCTTATCGTACCTGATCGACCT...

- By knowing the target gene sequence we can build a oligonucleotide probe that is complementary to the mRNA of the target gene sequence.
HYBRIDIZATION ON MICROARRAYS

RNA

AGCUGAUUAUUA
UUACUGUGAGAGAUU
AGCUGAUUAUUA
CGUGAAAAGUGAUGUA

cDNA

ACUAGCAGUGCAUGGU
ACTAGCATGCATGCT
AGCTGATTATTA
CGTGAAAGTGTGATGTA

Microarray

TGTGAAAAATGCC
CAGTCTT
TACTAGCATGCC
TGATCGT
CACAATT

probe 1
probe 2
probe 3
HYBRIDIZATION ON MICROARRAYS

fixed probes

different features
(e.g. bind different genes)

Fully complementary strands bind strongly

Partially complementary strands bind weakly

labelled target (sample)
GOOD MATCH STICKS, BAD ONE DOESN’T

http://www.csus.edu/indiv/r/rogersa/Bio181/GenechipHO.pdf
Fabrication of in-situ GeneChip

➢ Oligonucleotides are synthesized directly on silicon chip one base at a time.
➢ Uses a process called photolithography.

http://www.bcm.edu/garp/index.cfm?pmid=2904
Fabrication of in-situ GeneChip

http://cmr.asm.org/content/22/4/611/F3.expansion.html

“Gene expression analysis - microarrays”
Bioinformatics Course
DUAL-CHANNEL DETECTION

- cDNA prepared two samples to be compared (e.g. Diseased tissue versus healthy tissue) labeled with two different fluorophores.

- Although absolute levels of gene expression may be determined in the two-color array in rare instances, the relative differences in expression among different spots within a sample and between samples is the preferred method of data analysis for the two-color system.
cDNA prepared two samples to be compared (e.g. Diseased tissue versus healthy tissue) labeled with two different fluorophores.

Although absolute levels of gene expression may be determined in the two-color array in rare instances, the relative differences in expression among different spots within a sample and between samples is the preferred method of data analysis for the two-color system.

http://www.people.vcu.edu/~mreimers/OGMDA/image.html

"Gene expression analysis - microarrays"
Bioinformatics Course
cDNA prepared two samples to be compared (e.g. Diseased tissue versus healthy tissue) labeled with two different fluorophores.

Although absolute levels of gene expression may be determined in the two-color array in rare instances, the relative differences in expression among different spots within a sample and between samples is the preferred method of data analysis for the two-color system.
SINGLE-CHANNEL DETECTION

- Each array chip is exposed to only one sample.
- Data is more easily compared to arrays from different experiments.
- Comparison between hybridization intensities of same oligonucleotides need to made on different chips.
DUAL VS SINGLE CHANNEL DETECTION

Type 1 - single channel (expensive)

Type 2 - dual channel (cheaper)

AFFYMETRIX GeneChip PROPERTIES

$100-850

500,000 cells per chip

Millions of probes per cell

Each probe = 25 bp

(Courtesy of Affymetrix)
AFFYMETRIX GeneChip HYBRIDIZATION

(Courtesy of Affymetrix)
AFFYMETRIX GeneChip HYBRIDIZATION

After staining, RNA (purple UAGUAC) bound to the DNA probe built on the array will fluoresce.

"Gene expression analysis - microarrays"
Bioinformatics Course
AFFYMETRIX GeneChip SCANNING

Shining a laser light at GeneChip® array causes tagged DNA fragments that hybridized to glow

Non-hybridized DNA

Hybridized DNA

(Courtesy of Affymetrix)
## AFFYMETRIX GeneChip DESIGN

### Exon 1.0 ST Array design statistics summary

<table>
<thead>
<tr>
<th></th>
<th>Human</th>
<th>Mouse</th>
<th>Rat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nr of distinct probes</td>
<td>&gt;5 million</td>
<td>&gt;5 million</td>
<td>&gt;5 million</td>
</tr>
<tr>
<td>Probe sets</td>
<td>1.4 million</td>
<td>1.2 million</td>
<td>1.0 million</td>
</tr>
<tr>
<td>Exon clusters</td>
<td>&gt;1 million</td>
<td>~1 million</td>
<td>850,000</td>
</tr>
<tr>
<td>Supported by putative full-length mRNA</td>
<td>289,961 probe sets</td>
<td>266,200 probe sets</td>
<td>92,038 probe sets</td>
</tr>
<tr>
<td>Supported by Ensembl transcripts</td>
<td>306,583 probe sets</td>
<td>266,791 probe sets</td>
<td>195,943 probe sets</td>
</tr>
<tr>
<td>Supported by EST</td>
<td>665,175 probe sets</td>
<td>554,003 probe sets</td>
<td>211,451 probe sets</td>
</tr>
<tr>
<td>Supported by syntenic mRNA</td>
<td>220,262 probe sets</td>
<td>214,763 probe sets</td>
<td>272,061 probe sets</td>
</tr>
<tr>
<td>Supported by gene prediction</td>
<td>883,105 probe sets</td>
<td>835,897 probe sets</td>
<td>875,666 probe sets</td>
</tr>
<tr>
<td>Probe selection region</td>
<td>Along the entire length of the transcripts</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Probes/probe selection region</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Background subtraction strategy</td>
<td>Median fluorescence intensity of up to 1,000 background probes with the same GC content</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total features per array</td>
<td>&gt;5,500,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interrogated strand</td>
<td>Sense²</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### GeneChip Gene 1.0 ST Array design statistics summary

<table>
<thead>
<tr>
<th></th>
<th>Human</th>
<th>Mouse</th>
<th>Rat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nr of distinct probes</td>
<td>764,885</td>
<td>770,317</td>
<td>722,254</td>
</tr>
<tr>
<td>Estimated number of gene</td>
<td>28,869</td>
<td>28,853</td>
<td>27,342</td>
</tr>
<tr>
<td>Gene-level probe sets with Ensembl support</td>
<td>28,132</td>
<td>27,543</td>
<td>26,008</td>
</tr>
<tr>
<td>Gene-level probe sets with putative full-length transcript support</td>
<td>19,734</td>
<td>19,434</td>
<td>9,916</td>
</tr>
<tr>
<td>Probes per probe set</td>
<td>11</td>
<td>11</td>
<td>11</td>
</tr>
</tbody>
</table>

---

1. About 10 percent of the exon probe sets have fewer than four probes due to the probe selection region length and sequence constraints.
2. The probes tiled on the array are designed in the anti-sense orientation, requiring sense strand-labeled targets to be hybridized to the array.
3. By convention, the array is called ST Array, representing the necessity of using sense targets (the labeled sample to be hybridized to the array).
4. Supported by mouse or rat mRNA.
5. Supported by human or mouse mRNA.
AFFYMETRIX GeneChip DESIGN

Gene

10-20 probes, each 25nuc long

Microarray

probe set X

probe 4 4 ATGGACCGATGACGCGGATCGGAGCT
probe 2 2 CCGGTGACGCGGTACGTGAAAGTGAC
probe 1 1 TGTGTCATGCAACCGACCGAGCT
probe 3 3 GATGAGACGACGACGTCGAGAGGCT

“Gene expression analysis - microarrays”
Bioinformatics Course
### Bioconductor version 2.11 (Release)

<table>
<thead>
<tr>
<th>Software (608)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annotation (82)</td>
</tr>
<tr>
<td>AssayDomains (236)</td>
</tr>
<tr>
<td>AssayTechnologies (357)</td>
</tr>
<tr>
<td>Bioinformatics (394)</td>
</tr>
<tr>
<td>BiologicalDomains (80)</td>
</tr>
<tr>
<td>Infrastructure (167)</td>
</tr>
<tr>
<td>AnnotationData (667)</td>
</tr>
<tr>
<td>ExperimentData (137)</td>
</tr>
</tbody>
</table>

### Packages

<table>
<thead>
<tr>
<th>Package</th>
<th>Maintainer</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>a4</td>
<td>Tobias Verbeke</td>
<td>Automated Affymetrix Array Analysis Umbrella Package</td>
</tr>
<tr>
<td>a4Base</td>
<td>Tobias Verbeke</td>
<td>Automated Affymetrix Array Analysis Base Package</td>
</tr>
<tr>
<td>a4Classif</td>
<td>Tobias Verbeke</td>
<td>Automated Affymetrix Array Analysis Classification Package</td>
</tr>
<tr>
<td>a4Core</td>
<td>Tobias Verbeke</td>
<td>Automated Affymetrix Array Analysis Core Package</td>
</tr>
<tr>
<td>a4Preproc</td>
<td>Tobias Verbeke</td>
<td>Automated Affymetrix Array Analysis Preprocessing Package</td>
</tr>
<tr>
<td>a4Reporting</td>
<td>Tobias Verbeke</td>
<td>Automated Affymetrix Array Analysis Reporting Package</td>
</tr>
<tr>
<td>ABarray</td>
<td>Yongming Andrew Sun</td>
<td>Microarray QA and statistical data analysis for Applied Biosystems Genome Survey Microrarray (AB1700) gene expression data.</td>
</tr>
<tr>
<td>aCGH</td>
<td>Peter Dimitrov</td>
<td>Classes and functions for Array Comparative Genomic Hybridization data.</td>
</tr>
<tr>
<td>ACME</td>
<td>Sean Davis</td>
<td>Algorithms for Calculating Microarray Enrichment (ACME)</td>
</tr>
<tr>
<td>ADaCGH2</td>
<td>Ramon Diaz-Ularte</td>
<td>Analysis of data from aCGH experiments using parallel computing and ff objects</td>
</tr>
<tr>
<td>adSplit</td>
<td>Claudio Lottaz</td>
<td>Annotation-Driven Clustering</td>
</tr>
<tr>
<td>affxparser</td>
<td>Kasper Daniel Hansen</td>
<td>Affymetrix File Parsing SDK</td>
</tr>
<tr>
<td>affy</td>
<td>Rafael A. Irizarry</td>
<td>Methods for Affymetrix Oligonucleotide Arrays</td>
</tr>
<tr>
<td>affycomp</td>
<td>Rafael A. Irizarry</td>
<td>Graphics Toolbox for Assessment of Affymetrix Expression Measures</td>
</tr>
<tr>
<td>AffyCompatible</td>
<td>Martin Morgan</td>
<td>Affymetrix GeneChip software compatibility</td>
</tr>
<tr>
<td>affyContam</td>
<td>V. Carey</td>
<td>structured corruption of affymetrix cel file data</td>
</tr>
</tbody>
</table>
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
- Installation and Use
- Exploring Package Content
- Pre-Processing Resources

Sample Workflow

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

```r
## Load packages
> library(effy) # 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)

## differential expression
> design <- # describe model to be fit
    model.matrix(~ Disease, pData(eset))
> fit <- lmFit(eset, design) # fit each probeset to model
> efit <- eBayes(Fit) # empirical Bayes adjustment
> topTable(efit, coef=2) # table of differentially expressed probesets
```
```
# source('http://www.bioconductor.org/biocLite.R')

# biocLite('ygs98probe')
library(ygs98probe)
as.data.frame(ygs98probe)[1:10, 1:4]
```

<table>
<thead>
<tr>
<th></th>
<th>sequence</th>
<th>x</th>
<th>y</th>
<th>Probe.Set.Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CGATATGGAGGCAATGCTGGTAGAC</td>
<td>90</td>
<td>121</td>
<td>10000_at</td>
</tr>
<tr>
<td>2</td>
<td>GGAGGCAATGCTGGTAGACGAACCTC</td>
<td>91</td>
<td>121</td>
<td>10000_at</td>
</tr>
<tr>
<td>3</td>
<td>AATGCTGGTAGACGAACTCGATATGT</td>
<td>92</td>
<td>121</td>
<td>10000_at</td>
</tr>
<tr>
<td>4</td>
<td>GGTAGACGAACTCGATATGTCAGACG</td>
<td>93</td>
<td>121</td>
<td>10000_at</td>
</tr>
<tr>
<td>5</td>
<td>CGAACTCGATATGTCAGACGAGGGAC</td>
<td>94</td>
<td>121</td>
<td>10000_at</td>
</tr>
<tr>
<td>6</td>
<td>CGTATGTGACACGAGGACCTACTT</td>
<td>95</td>
<td>121</td>
<td>10000_at</td>
</tr>
<tr>
<td>7</td>
<td>GAGGGACCTACTTGATGTCAGATGAA</td>
<td>96</td>
<td>121</td>
<td>10000_at</td>
</tr>
<tr>
<td>8</td>
<td>TGAGTCAGACCTTGAGGAAGAAACC</td>
<td>97</td>
<td>121</td>
<td>10000_at</td>
</tr>
<tr>
<td>9</td>
<td>TGAAGAACTAACCTCCTACTTACGGAA</td>
<td>98</td>
<td>121</td>
<td>10000_at</td>
</tr>
<tr>
<td>10</td>
<td>TCTACTTTACGGAAGACGTGCAGTA</td>
<td>99</td>
<td>121</td>
<td>10000_at</td>
</tr>
</tbody>
</table>
```R
# biocLite('ygs98.db')
library(ygs98.db)
ygs98ORF[['10000_at']]

## [1] "YLR331C"

ygs98CHRLOC[['10000_at']]

##    12
## -790669

ygs98CHRLOCEND[['10000_at']]

##    12
## -791046
```

"Gene expression analysis - microarrays"
Bioinformatics Course
```r
# biocLite('BSgenome.Scerevisiae.UCSC.sacCer1')
library(BSgenome.Scerevisiae.UCSC.sacCer1)
library(Biostrings)
Scerevisiae$chr12[790669:791046]

## 378-letter "DNASTRING" instance
## seq: TCATAGTATGTGTCTTTTCACAACCAAGAATAGT...TTCAGATTCTTCTCATCCTCCTCATCCTTTTCAT

seq = Scerevisiae$chr12[790669:791046]
probe = ygs98probe[1, 1]
library(Biostrings)
matchPattern(probe, seq)

## Views on a 378-letter DNAString subject
## subject: TCATAGTATGTGTCTTTTCACAACCAAGAATA...CAGATTCTTCTCATCCTCCTCATTCTTTTCAT
## views: NONE
```

"Gene expression analysis - microarrays"
Bioinformatics Course
```
matchPattern(probe, reverseComplement(seq))
```

```
# Views on a 378-letter DNAString subject
# subject: ATGGAAAAAGGATGAGGAGGATGAAGATCTGA...TTTCGTTGTGAAGACAAACATACATATGA
# views:
#   start   end    width
# [1]    90   114    25 [CGATATGGGAGCAATGCTGGTAGAC]
```

```
as.data.frame(ygs98probe[1, 4:5])
```

```
# Probe.Set.Name  Probe.Interrogation.Position
# 1   10000_at     102
```

"""Gene expression analysis - microarrays"
Bioinformatics Course"""
AFFYMETRIX FILE TYPES

➢ .DAT (Data File)
  Raw (TIFF) optical image of the hybridized chip.

➢ .CEL (Cell Intensity File)
  Processed DAT file (intensity/position values). A .CEL file contains a single value for each feature. Each value in a .CEL file is a statistical “summary” of the fluorescence from a single feature.

➢ .CDF (Chip Description File)
  Provided by Affymetrix, describes layout of chip – identity probe sets.
A .DAT File

http://www.ucdmc.ucdavis.edu/medmicro/microarray/

"Gene expression analysis - microarrays"
Bioinformatics Course
QUALITY CONTROL (QC) – VISUAL INSPECTION

➢ Look at the image file

➢ Scratches? Spots? Fingerprints?
➢ Array name is written out in oligos.

source("http://bioconductor.org/biocLite.R")
biocLite("arrayQualityMetrics")
Minimum Information About a Microarray Experiment - MIAME

MIAME describes the Minimum Information About a Microarray Experiment that is needed to enable the interpretation of the results of the experiment unambiguously and potentially to reproduce the experiment. [Brazma et al., Nature Genetics]

The six most critical elements contributing towards MIAME are:

1. The raw data for each hybridisation (e.g., CEL or GPR files)
2. The final processed (normalised) data for the set of hybridisations in the experiment (study) (e.g., the gene expression data matrix used to draw the conclusions from the study)
3. The essential sample annotation including experimental factors and their values (e.g., compound and dose in a dose response experiment)
4. The experimental design including sample data relationships (e.g., which raw data file relates to which sample, which hybridisations are technical, which are biological replicates)
5. Sufficient annotation of the array (e.g., gene identifiers, genomic coordinates, probe oligonucleotide sequences or reference commercial array catalog number)
6. The essential laboratory and data processing protocols (e.g., what normalisation method has been used to obtain the final processed data)

For more details, see MIAME 2.0.

http://www.mged.org/Workgroups/MIAME/miame.html
GENE EXPRESSION DATABASES

- Gene Expression Omnibus
- ArrayExpress
  http://www.ebi.ac.uk/arrayexpress/
Gene expression analysis - microarrays

Bioinformatics Course
ARRAYEXPRESS & AFFY

> library(ArrayExpress)
> library(affy)
> affydata = ArrayExpress("E-GEOD-18037")

> affydata
AffyBatch object
size of arrays=534x534 features (18 kb)
cdf=YG_S98 (9335 affyids)
number of samples=4
number of genes=9335
annotation=ygs98
notes=E-GEOD-18037
    E-GEOD-18037
    NA
c("unknown_experiment_design_type"
### Gene expression analysis - microarrays

**Bioinformatics Course**

```R
> pm(affydata)[1:5,]
   GSM451051.CEL GSM451049.CEL GSM451050.CEL GSM451048.CEL
64704  10066.3   10329.3    4719.0    6084.0
64705    112.3      105.8     81.3      96.0
64706    112.8      112.3    106.3    105.5
64707    85.8     112.5      70.0    90.5
64708    69.5      80.3      58.0    76.0
> mmm(affydata)[1:5,]
   GSM451051.CEL GSM451049.CEL GSM451050.CEL GSM451048.CEL
65238    7597.0    8203.3    2949.0    4438.3
65239    113.3     144.5     100.0    131.0
65240     84.0      80.8      60.0     76.0
65241     73.3      91.5      55.0     80.5
65242     84.3      92.5      61.3     72.0
> dim(pm(affydata))
[1] 138412     4
> dim(ygs98probe)
[1] 138412     6
```
PROBLEMS

- Do we trust the measured data?
- How to make several arrays comparable?
- What to do with PM/MM if they exist?
- How to match Probeset to a Gene?
- How to visualize the results?
PREPROCESSING (SINGLE CHANNEL)

- Background correction
  - PM/MM probes, against GC content

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

- Summarization
  - from probes to probesets (approximately, genes)

PREPROCESSING

```r
library(affy)
library(ArrayExpress)
affydata = ArrayExpress("E-GEOID-18037")
expdata <- rma(affydata)
```

## Background correcting
## Normalizing
## Calculating Expression

```r
head(exprs(expdata))
```

<table>
<thead>
<tr>
<th></th>
<th>GSM451051.CEL</th>
<th>GSM451049.CEL</th>
<th>GSM451050.CEL</th>
<th>GSM451048.CEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>10000_at</td>
<td>4.351</td>
<td>4.065</td>
<td>4.005</td>
<td>4.307</td>
</tr>
<tr>
<td>10001_at</td>
<td>10.148</td>
<td>9.505</td>
<td>10.223</td>
<td>9.879</td>
</tr>
<tr>
<td>10002_i_at</td>
<td>12.196</td>
<td>12.207</td>
<td>11.537</td>
<td>11.700</td>
</tr>
<tr>
<td>10003_f_at</td>
<td>12.900</td>
<td>13.011</td>
<td>12.780</td>
<td>12.938</td>
</tr>
<tr>
<td>10004_at</td>
<td>4.174</td>
<td>4.497</td>
<td>4.460</td>
<td>4.596</td>
</tr>
<tr>
<td>10005_at</td>
<td>8.615</td>
<td>9.149</td>
<td>8.345</td>
<td>8.743</td>
</tr>
</tbody>
</table>
LOG TRANSFORMATION

- Makes variation of intensities and ratios of intensities more independent of absolute magnitude.
- Evens out highly skewed distributions; gives more realistic sense of variation.
- Approximate normal distribution; treats up- and down-regulated genes symmetrically; helps visualize variation in both directions.

http://odin.mdacc.tmc.edu/~kdo/TeachBioinf/Week%201/Lecture1-Jan7-08.pdf
NORMALIZATION

Before the quantification of individual probesets, we need to check whether the intensities of all microarrays are comparable. Adding twice as much sample may make the resultant image brighter, but it doesn’t tell us anything new about the underlying biology. **Removes non-biological variation.**

```r
library(affy)

> normdata = mas5(affydata)  # After calling mas5

> normdata = rma(affydata)  # After calling rma

> head( exprs( normdata ) )
```

"Gene expression analysis - microarrays"
Bioinformatics Course
SOURCES OF NON-BIOLOGICAL VARIATION

- Dye bias: differences in heat and light sensitivity; efficiency of dye incorporation.
- Differences in amount of labeled cDNA hybridized.
- Different amounts of mRNA.
- Different scanning parameters.
- Different technicians producing arrays.
- Any process that produces systematic error.
NORMALIZATION METHODS

• Robust Multi-array Average (RMA)

• Factor Analysis for Robust Microarray Summarization (FARMS)

• Distribution Free Weighted method (DFW)

• Probe Logarithmic Intensity Error (PLIER)
QUANTILE NORMALIZATION

Before normalization

After normalization

GENE EXPRESSION DATA

"Gene expression analysis - microarrays"
Bioinformatics Course
DIFFERENTIAL EXPRESSION

To understand the effect of a drug we might be interested to know what genes are up-regulated (increased in expression) or down-regulated (decreased in expression) between treatment and control groups?

Find genes with different expression between conditions
DIFFERENTIAL EXPRESSION METHODS

• use a t-test or it’s derivates

• Limma R package
  > library(limma)

• RankProd R package
  > library(RankProd)

• Fold change
LIMMA

- use design matrix to establish the 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

```r
> 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”, …)
```