Universität Rostock Traditio et Innovatio

Universitätsmedizin Rostock

Integrative Bioinformatics and Systems Biology

Dr. Mohamed Hamed

LECTURE PLANNING

Lecture: 12 lectures, 3 hrs each.

Total workload: 42 hrs : 36 hrs of lectures and tutorials and 6 hrs of self studies.

Entrance requirements: basic knowledge of biology and computer science.

Literature: Lecture slides, tutorial handouts and scripts will be provided.

INTEGRATIVE BIOINFORMATICS

COURSE ABBREVIATION: INT-BIO

LANGUAGE: ENGLISH
USED MEDIA: POWERPOINT PRESENTATION

Module: Lecture and tutorial.

By: Dr. Mohamed Hamed

Head of the Integrative OMICs Analysis Group in Rostock University medical center, Rostock University, Germany.

MOTIVATION AND COURSE OBJECTIVES

The main challenge of modern systems biology is unraveling the holistic picture of the complex molecular interactions that occur on different molecular levels (genomic, transcriptomic, epigenomic, proteomic, etc). Therefore, the needs to integrate/jointly analyze biological data from different high-throughput technologies emerged in order to identify biomarkers for early diagnosis and prognosis of complex diseases and facilitating the development of novel treatment approaches.

This course aims at teaching students how to perform data-specific computational analyses as well as integrative analysis approaches, combining knowledge from different OMICs-based datasets. Both, theoretical and practical aspects will be covered. Students will have the opportunity to work both independently during tutorials and in teams during the research project.

COMPETENCES TO BE DEVELOPED

Students will get practical and extensive hands-on experience on:

-R scripting language and bioconductor packages.

-Data-specific computaional analysis and pipelines for the vast amounts of biological data produced using high-throughput technologies.

-Developing and applying integrative bioinformatics methods that could be utilized in all biology-related areas of interest.

-Basics of machine learning methods as tools for integrating biological features from heterogeneous Omics data.

-Students will be developing their own research projects, interpreting the obtained results, writing a manuscript, scientifically discussing the results.

ASSESSMENT

-Students need to finalize a research project applying all/ most of the learned methods and skills during the course. Novelty and extending the learned methods is highly encouraged and will be well graded.

-The outcomes of each research project should be compiled in a high scientific quality research article that is ready for submission in a peer-review journal.

-All projects will be presented, discussed and scientifically reviewed in the last lecture.

R language mini-course 1	R language mini-course 2	R language mini-course 3	R language mini-course 4
-Introduction to the course -Basics of R language and -R studio IDE	-R statistics -Advanced R statistics	-Bio-conductor packages	-Case study: Microarray analysis using R
Transcriptomic analysis I	**Transcriptomic analysis I I**	**Non-coding RNAs**	**Introduction to integrative bioinformatics**
-From microarrays to RNA-seq -RNA-seq analysis	-Linking to Ontologies and pathways -Data visualization	-Small and long non-coding RNAs -miRNA sequencing analysis -miRNA databases	-Importance of data integration -Different methods for biological data integration -OMICs data types, and TCGA repository -Databases of diseases-related genes and miRNAs
Network- based integrative methods	**Epigenetics**	**Integrative analysis based on machine learning**	**PROJECTS DISCUSSION AND CLOSURE**
-TFmiR analysis -Network motif analysis -Central hubs identifications	-Introduction to the epigenetic landscapes of normal and tumor cells -DNA methylation, Co-methylation analysis -DMRs identifications	-Introduction to machine learning in bioinformatics. -Unsupervised methods: Clustering biological data -PCA analysis, MOFA method -Outlook at deep neural networks applications in bioinformatics	-Projects presentation. -Reviews of the potential manuscripts

LECTURE PLANNING

Lecture: 12 lectures, 3 hrs each.

Total workload: 42 hrs : 36 hrs of lectures and tutorials and 6 hrs of self studies.

Entrance requirements: basic knowledge of biology and computer science.

Literature: Lecture slides, tutorial handouts and scripts will be provided.

INTEGRATIVE BIOINFORMATICS

COURSE ABBREVIATION: INT-BIO

LANGUAGE: ENGLISH
USED MEDIA: POWERPOINT PRESENTATION

Module: Lecture and tutorial.

By: Dr. Mohamed Hamed

Head of the Integrative OMICs Analysis Group in Rostock University medical center, Rostock University, Germany.

MOTIVATION AND COURSE OBJECTIVES

The main challenge of modern systems biology is unraveling the holistic picture of the complex molecular interactions that occur on different molecular levels (genomic, transcriptomic, epigenomic, proteomic, etc). Therefore, the needs to integrate/jointly analyze biological data from different high-throughput technologies emerged in order to identify biomarkers for early diagnosis and prognosis of complex diseases and facilitating the development of novel treatment approaches.

This course aims at teaching students how to perform data-specific computational analyses as well as integrative analysis approaches, combining knowledge from different OMICs-based datasets. Both, theoretical and practical aspects will be covered. Students will have the opportunity to work both independently during tutorials and in teams during the research project.

COMPETENCES TO BE DEVELOPED

Students will get practical and extensive hands-on experience on:

-R scripting language and bioconductor packages.

-Data-specific computaional analysis and pipelines for the vast amounts of biological data produced using high-throughput technologies.

-Developing and applying integrative bioinformatics methods that could be utilized in all biology-related areas of interest.

-Basics of machine learning methods as tools for integrating biological features from heterogeneous Omics data.

-Students will be developing their own research projects, interpreting the obtained results, writing a manuscript, scientifically discussing the results.

ASSESSMENT

-Students need to finalize a research project applying all/ most of the learned methods and skills during the course. Novelty and extending the learned methods is highly encouraged and will be well graded.

-The outcomes of each research project should be compiled in a high scientific quality research article that is ready for submission in a peer-review journal.

-All projects will be presented, discussed and scientifically reviewed in the last lecture.

R language mini-course 1	R language mini-course 2	R language mini-course 3	R language mini-course 4
-Introduction to the course -Basics of R language and -R studio IDE	-R statistics -Advanced R statistics	-Bio-conductor packages	-Case study: Microarray analysis using R
Transcriptomic analysis I	Transcriptomic analysis I I	Non-coding RNAs	Introduction to integrative bioinformatics
-From microarrays to RNA-seq -RNA-seq analysis	-Linking to Ontologies and pathways -Data visualization	-Small and long non-coding RNAs -miRNA sequencing analysis -miRNA databases	-Importance of data integration -Different methods for biological data integration -OMICs data types, and TCGA repository -Databases of diseases-related genes and miRNAs
Network- based integrative methods	Epigenetics	Integrative analysis based on machine learning	PROJECTS DISCUSSION AND CLOSURE
-TFmiR analysis -Network motif analysis -Central hubs identifications	-Introduction to the epigenetic landscapes of normal and tumor cells -DNA methylation, Co-methylation analysis -DMRs identifications	-Introduction to machine learning in bioinformatics. -Unsupervised methods: Clustering biological data -PCA analysis, MOFA method -Outlook at deep neural networks applications in bioinformatics	-Projects presentation. -Reviews of the potential manuscripts

Lecture 4
Microarray analysis using R

Preparation for the tutorials

- Installing bioconductor itself and other bioconductor packages
- *Downloading the dataset*

Installing bioconductor

Type : sessionInfo()
In RStudio Console

- If R version <3.5.0

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

Install specific packages, e.g., "GenomicFeatures" and "AnnotationDbi", with

```
BiocInstaller::biocLite(c("GenomicFeatures", "AnnotationDbi"))
```

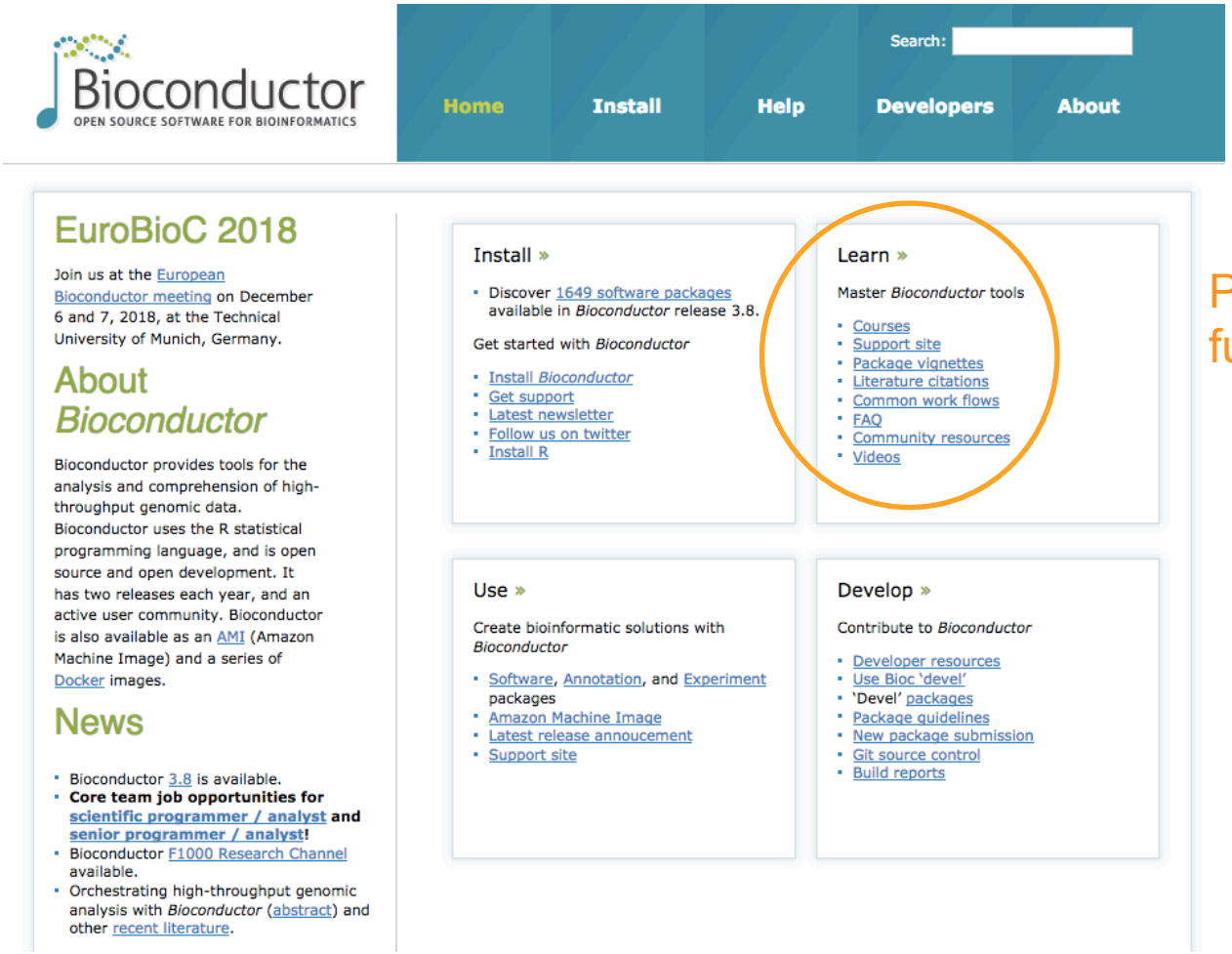
- R version 3.5.0 or higher

```
if (!requireNamespace("BiocManager"))
    install.packages("BiocManager")
BiocManager::install()
```

Update your R version to the newest one :
https://www.r-project.org/

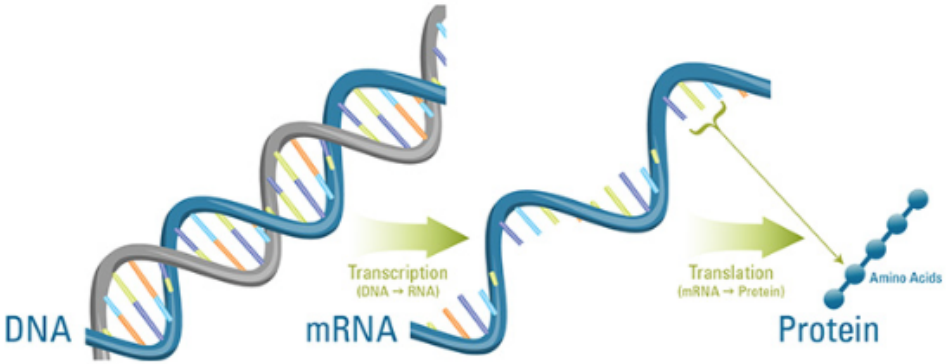
```
BiocManager::install(c("GenomicFeatures", "AnnotationDbi"))
```

Bioconductor – Collection of R packages for bioinformatics

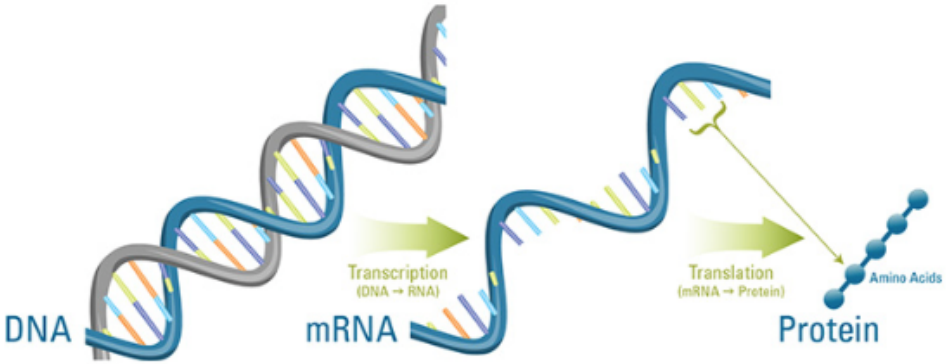


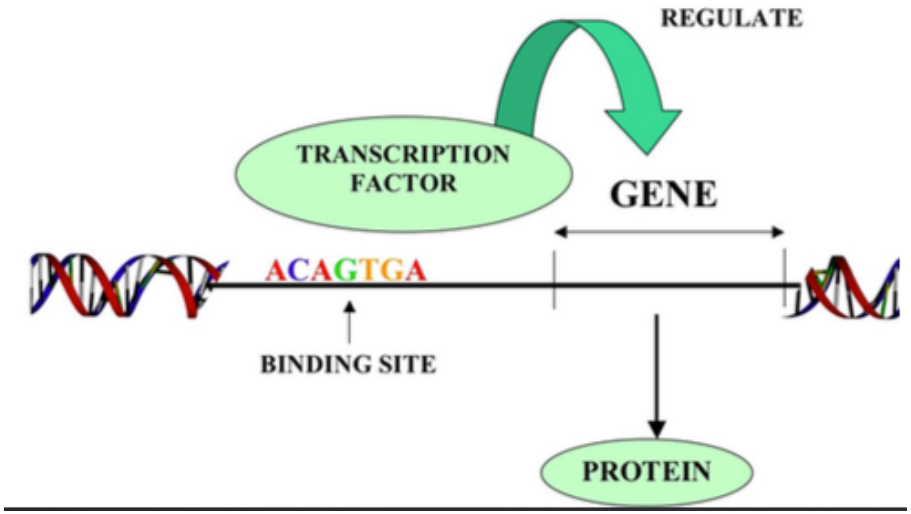
Provide Tutorial / Links to further learning material

Central Dogma of life

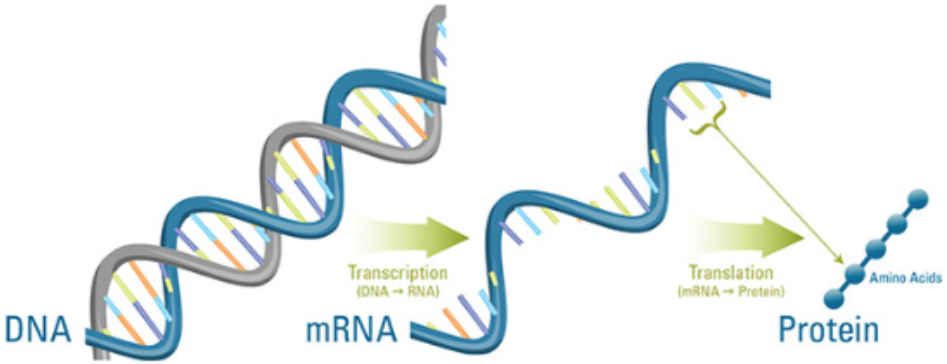


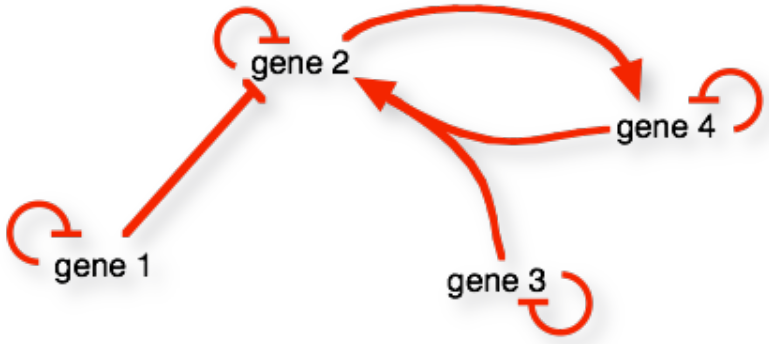
Central Dogma of life



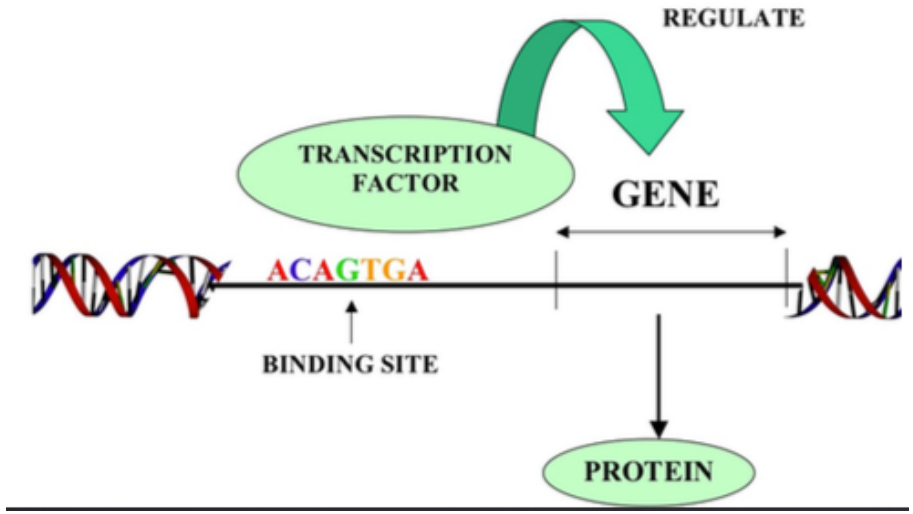


Central Dogma of life





Gene Regulatory Network



https://sbi4u2013.wordpress.com/author/viceteacher/

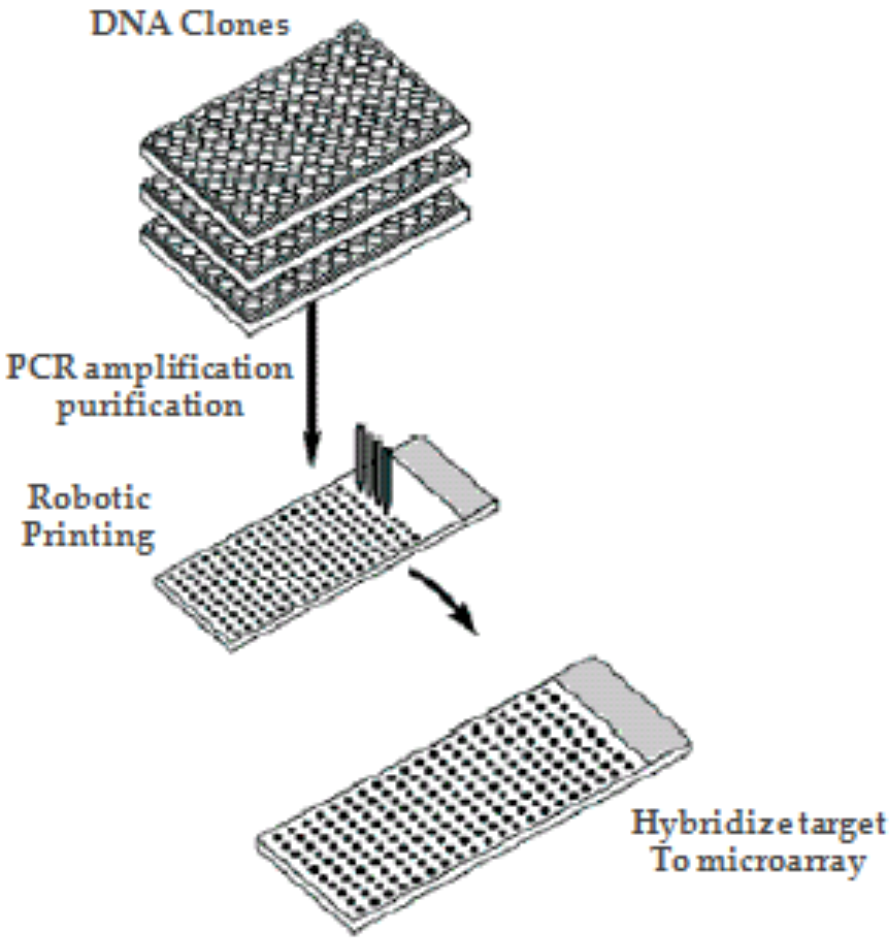
Quantifications of Gene expressions

- Microarrays
- RNA-Seq
- PCR

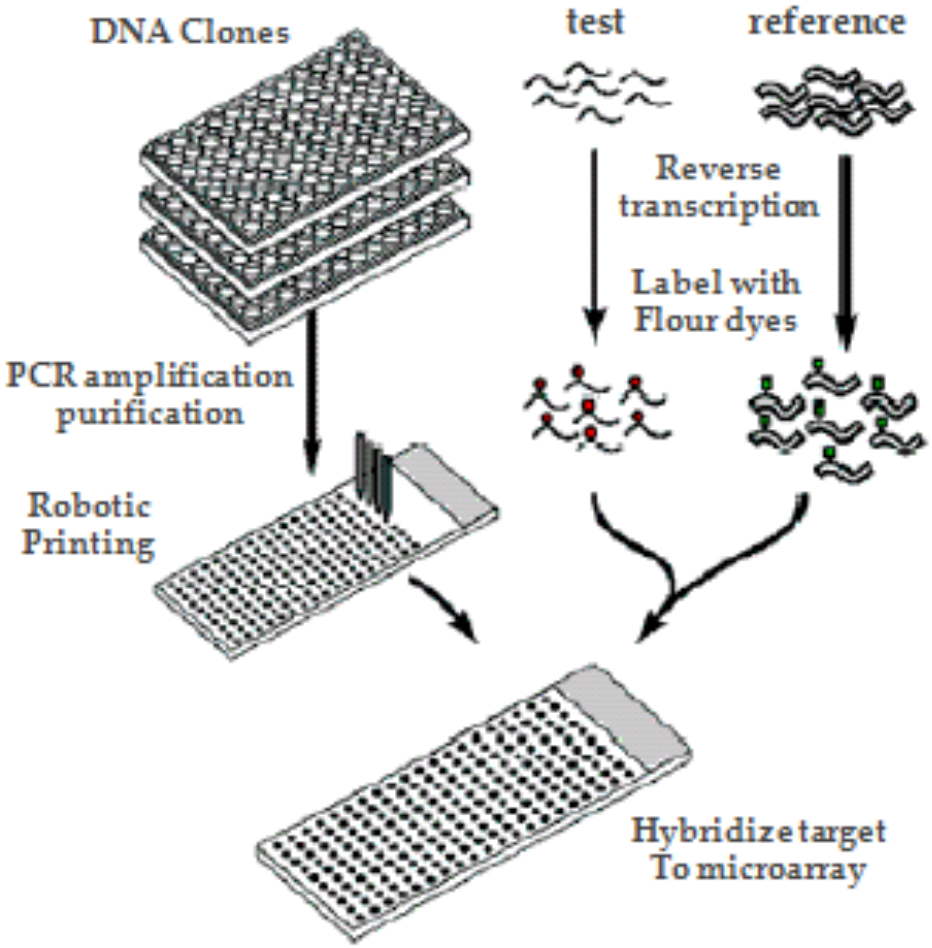
Microarrays

- Microarrays provide a powerful tool for relative quantification of a cell's gene expression.
- Using two samples, ratios of gene-expression levels can be used to detect meaningfully different expression levels between the samples for a given gene.
- With multiple tissue samples, microarray data can be used to cluster genes based on expression profiles to characterize and classify disease based on the expression levels of gene sets.
- Thousands of tests can be done in parallel.

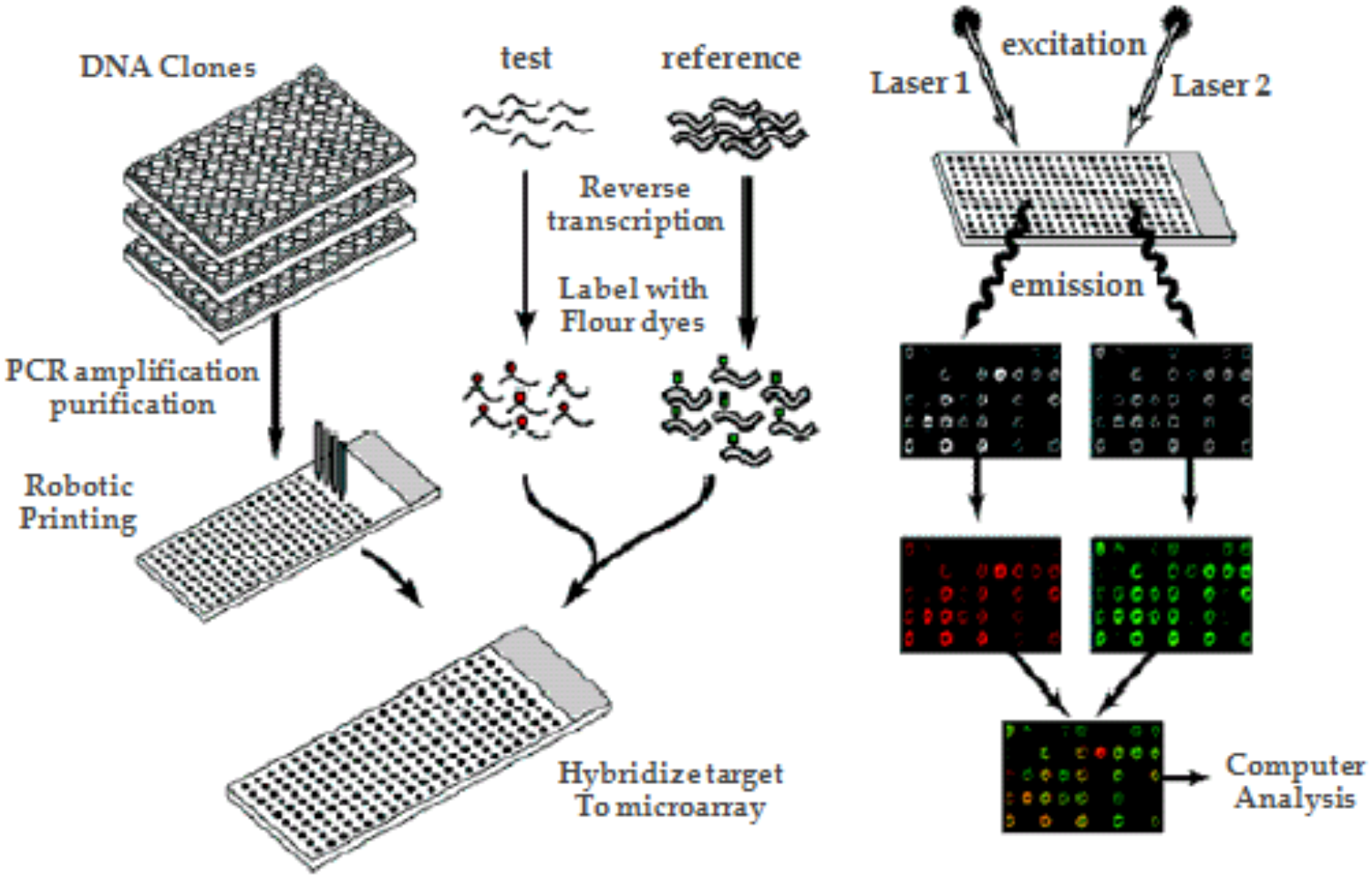
Microarray design and *experiment*



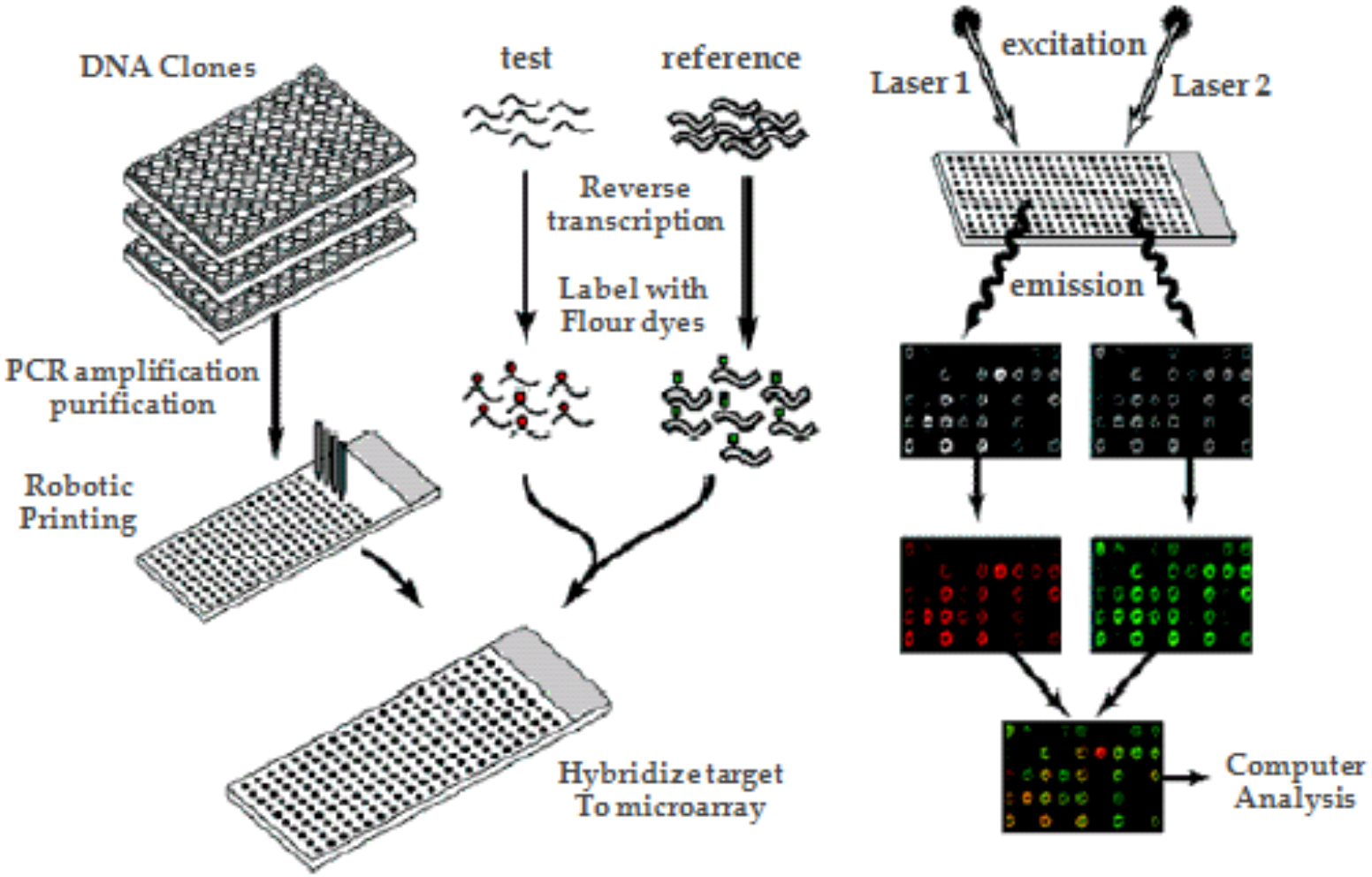
Microarray design and *experiment*



Microarray design and *experiment*

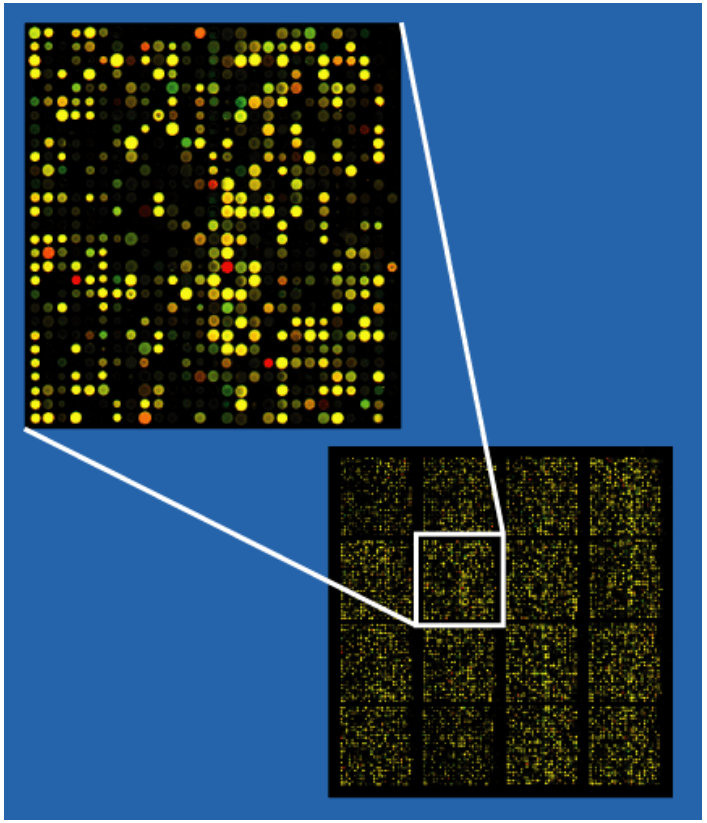


Microarray design and *experiment*

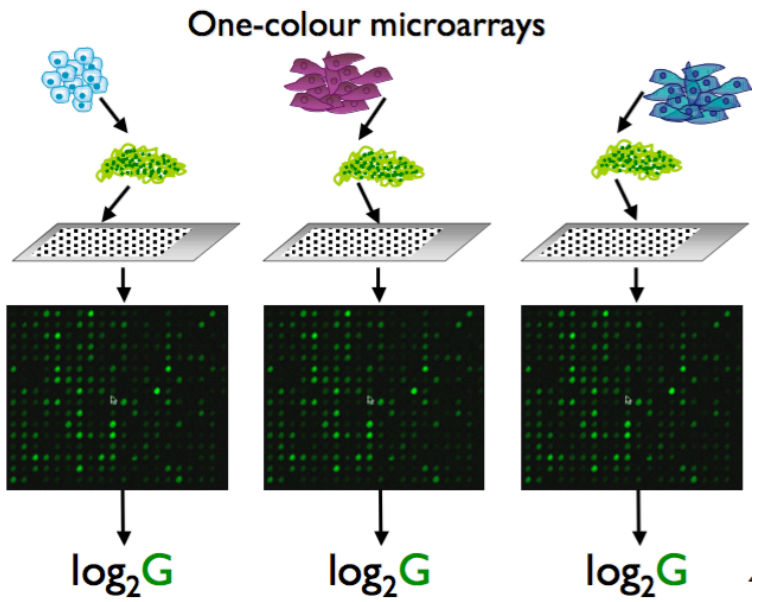


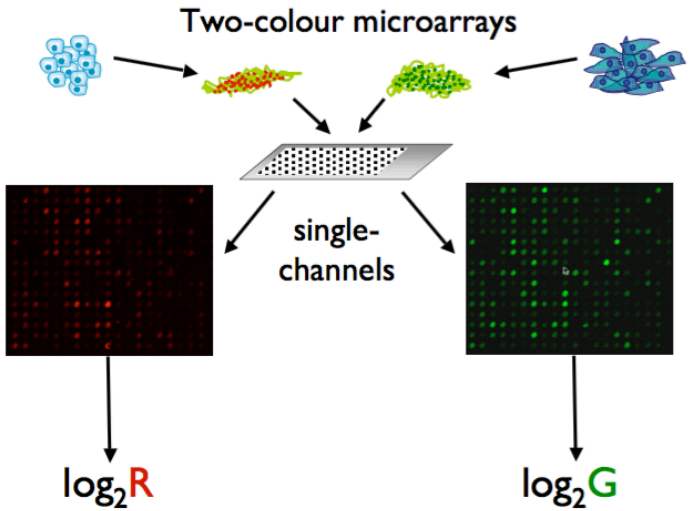
Transcription quantification

- Intense red spots indicates a high level of expression of that gene in the *test* sample with little expression in the *control* sample.
- Intense green spots indicates the opposite.
- When both *test* and *control* samples express a gene at similar levels, the observed array spot is yellow.



Transcription quantification





	c1	c2	c_m
g1			
g2			
g_n			

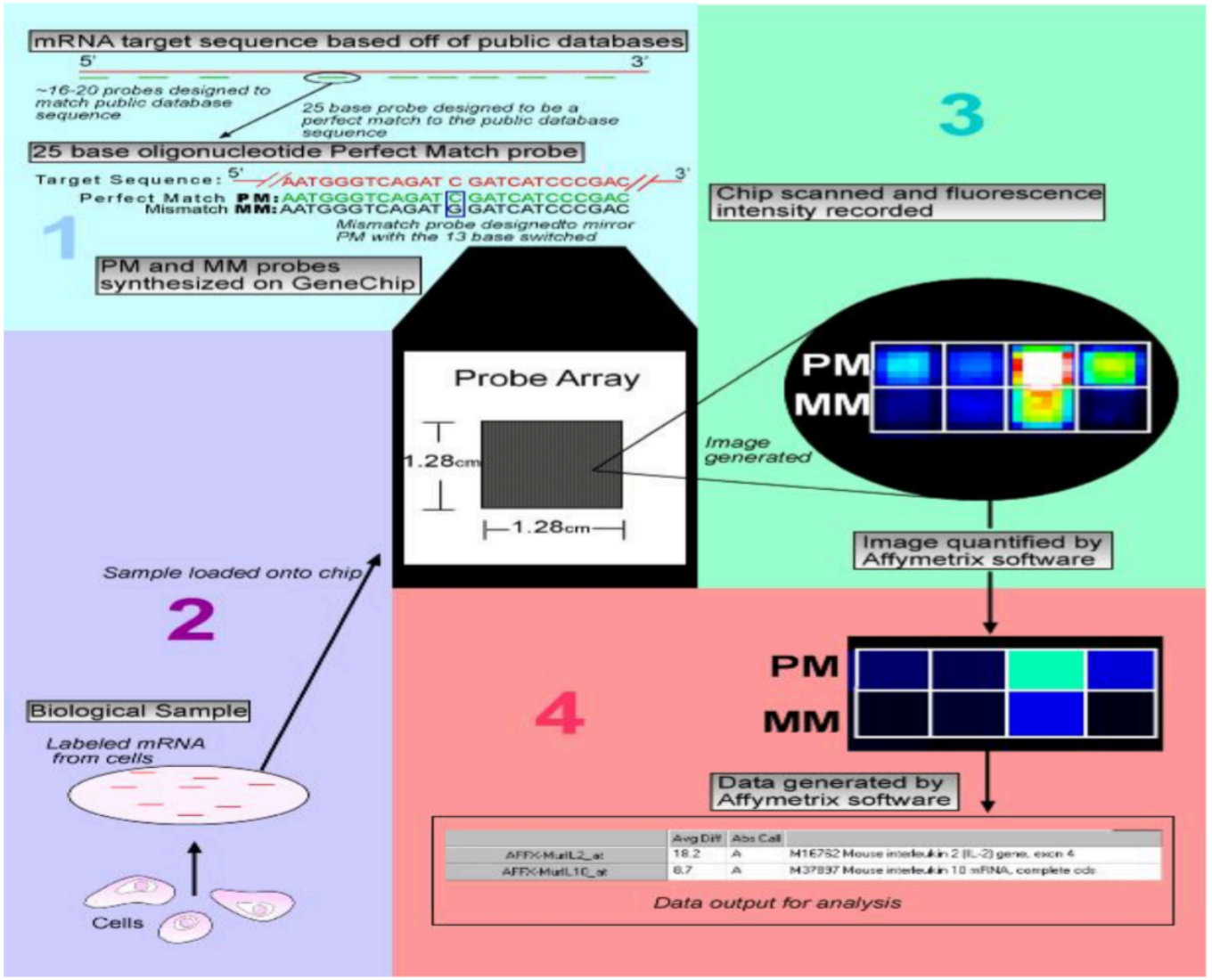
Output is
Data Matrix
In both cases

Two-colour microarray statistics

$\log_2 R$ $\log_2 G$

$$M = \log_2 R - \log_2 G$$
$$= \log_2 (R / G)$$

Affymatrix GeneChip technology



Affymatrix GeneChip technology

- Each gene or portion of a gene is represented by 1 to 20 oligonucleotides of 25 base-pairs.
- **Probe**: an oligonucleotide of 25 base-pairs, i.e., a 25-mer.
- **Perfect match (PM)**: A 25-mer complementary to a reference sequence of interest (e.g., part of a gene).
- **Mismatch (MM)**: same as PM but with a single base change for the middle (13th) base (transversion purine <-> pyrimidine, G <->C, A <->T). Used to measure non-specific binding and background noise.
- **Probe-pair**: a (PM,MM) pair.
- **Probe-pair set**: a collection of probe-pairs (1 to 20) related to a common gene or fraction of a gene.
- **Affy ID**: an identifier for a probe-pair set.

Affymatrix GeneChip technology

Related Files:

- **DAT file**: Image file, 107 pixels, 50 MB.
- **CEL file**: Cell intensity file, probe level PM and MM values.
- **CDF file**: Chip Description File. Describes which probes go in which probe sets and the location of probe-pair sets (genes, gene fragments, ESTs). (**Annotation**)

Affymatrix GeneChip technology

Expression Measures :

- 10-20K genes represented by 11-20 pairs of probe intensities (PM & MM).

- Obtain expression measure for each gene on each array by summarizing these pairs.

- Traditional Preprocessing steps are:
  - Exploratory analysis
  - Background adjustment
  - Normalization
  - Transformation
  - Summarization

Microarray data repositories

1- GEO: **G**ene **E**xpression **O**mnibus

NCBI

DATASET BROWSER (CURATED)

GEO Gene Expression Omnibus

Search for | Search | Clear | Show All | **Advanced Search** | Page size 20

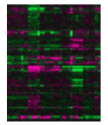
4348 DataSet records | Page 1 of 218 > >>

DataSet	Title	Organism(s)	Platform	Series	Samples
GDS6248	Diet-induced obesity model: liver	*Mus musculus*	GPL6887	GSE39549	51
GDS6247	Diet-induced obesity model: white adipose tissue	*Mus musculus*	GPL6887	GSE39549	40
GDS6177	Acute alcohol consumption effect on whole blood (control group): time course	*Homo sapiens*	GPL570	GSE20489	25
GDS6176	Caspase-1 deficiency effect on lipid-loaded intestines	*Mus musculus*	GPL11533	GSE32515	18
GDS6100	MicroRNA-135b overexpression effect on prostate cancer cell line: time course	*Homo sapiens*	GPL10558	GSE57820	12
GDS6083	Chronic lymphocytic leukemia cells response to the neutralization of inhibitor of apoptosis proteins	*Homo sapiens*	GPL570	GSE62533	12
GDS6082	Sendai virus infection effect on monocytic cell line: dose response	*Homo sapiens*	GPL10558	GSE67198	11
GDS6064	Arthritic tarsal joints induced by collagen: time course	*Mus musculus*	GPL6246	GSE61140	15
GDS6063	Influenza A effect on plasmacytoid dendritic cells	*Homo sapiens*	GPL10558	GSE68849	10
GDS6016	Transcription factor engrailed-2 loss-of-function model of autism spectrum disorder: hippocampus	*Mus musculus*	GPL7202	GSE51612	6

DataSet Record GDS6248: Expression Profiles | Data Analysis Tools | Sample Subsets

Title:	Diet-induced obesity model: liver		
Summary:	Analysis of livers of C57BL/6J mice fed a high fat diet for up to 24 weeks. Significant body weight gain was observed after 4 weeks. Results provide insight into the effect of high fat diets on metabolism in the liver.		
Organism:	*Mus musculus*		
Platform:	GPL6887: Illumina MouseWG-6 v2.0 expression beadchip		
Citations:	Kwon EY, Shin SK, Cho YY, Jung UJ et al. Time-course microarrays reveal early activation of the immune transcriptome and adipokine dysregulation leads to fibrosis in visceral adipose depots during diet-induced obesity. *BMC Genomics* 2012 Sep 4;13:450. PMID: 22947075 Do GM, Oh HY, Kwon EY, Cho YY et al. Long-term adaptation of global transcription and metabolism in the liver of high-fat diet-fed C57BL/6J mice. *Mol Nutr Food Res* 2011 Sep;55 Suppl 2:S173-85. PMID: 21618427		
Reference Series:	GSE39549	**Sample count:**	51
Value type:	transformed count	**Series published:**	2014/03/01

Cluster Analysis

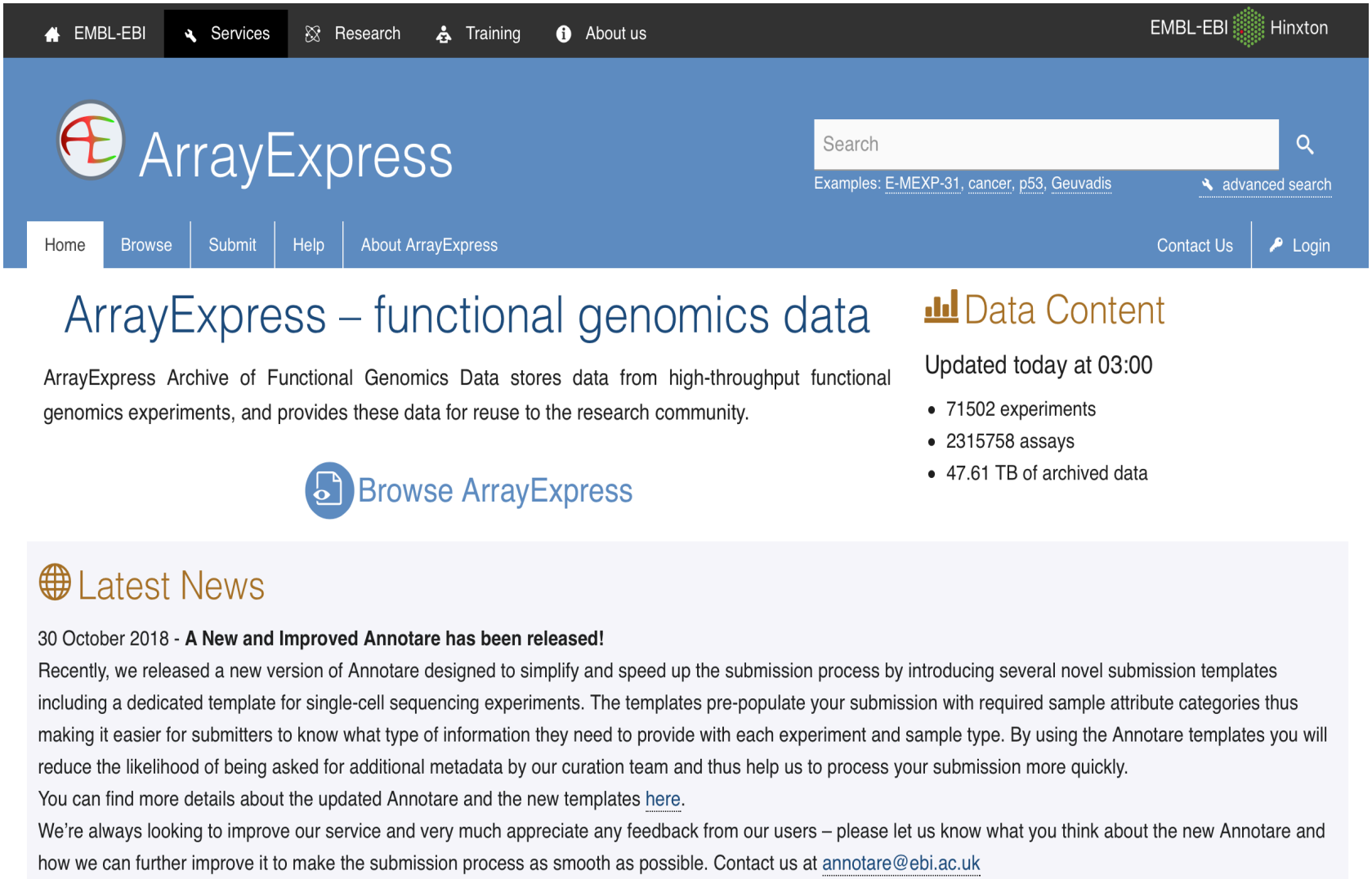


Download

- DataSet full SOFT file
- DataSet SOFT file
- Series family SOFT file
- Series family MINiML file
- Annotation SOFT file

NLM NIH GEO Help Disclaimer Accessibility

2- ArrayExpress



Bioinformatics workflow for Microarray data

Data loading

- Download the dataset of interest from GEO /ArrayExpress as CEL files

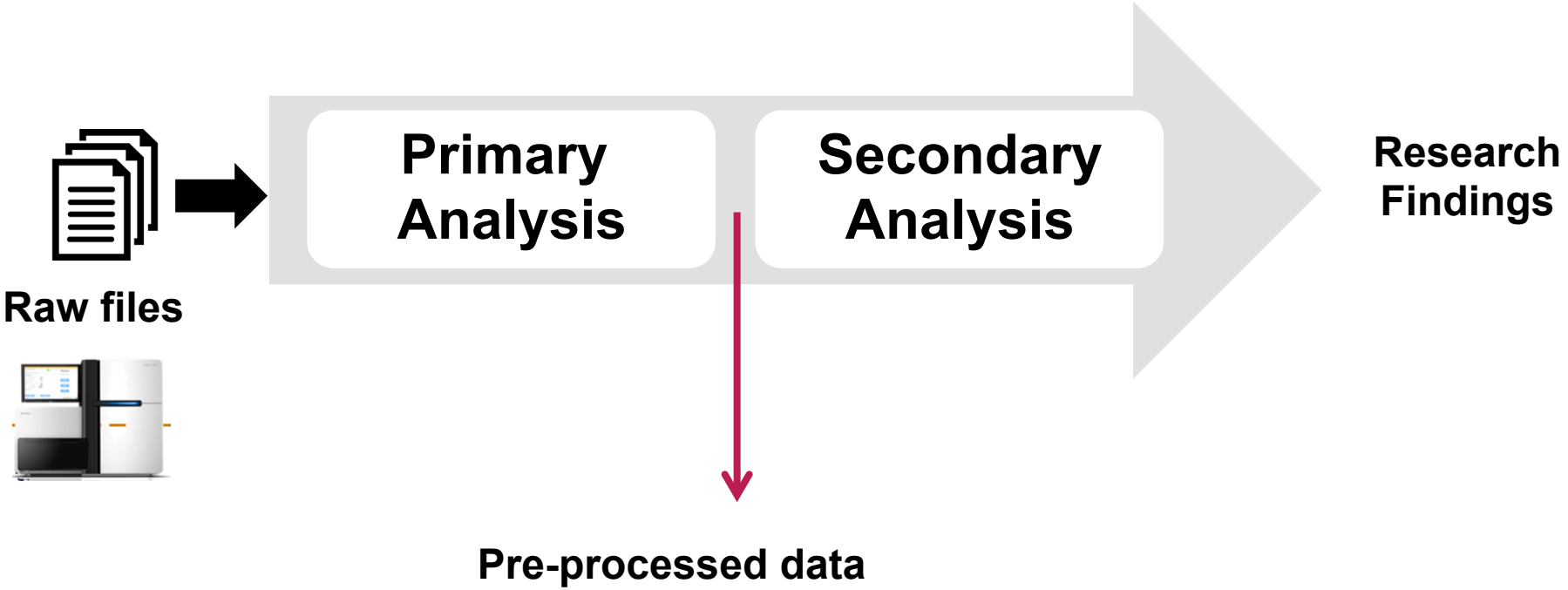
Loading the data into your R Environment

- Affymetrix GeneChip CEL files are imported using ReadAffy

from the affy package.

```
library("affy")
myAB1 <- ReadAffy()
myAB2 <- ReadAffy(filenames = c("a1.cel", "a2.cel", "a3.cel"))
```

Bioinformatics Workflow



Bioinformatics Workflow

Raw files (CEL,fastq)

Primary Analysis

Secondary Analysis

Research Findings

	Control samples				Study samples			
	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7	Sample8
Gene 1	12	11	15	13	1	2	4	5
Gene 2	7	6	4	0.5	23	21	23	22
Gene 3	5	2	5	6	12.4	14	12	15.5
.....	18	15	15	20	3	2	3	4
.....	19	18	11	17	11	14	12	17
Gene n	5	13	14	22	12	11	23	10

Gene Expression Matrix

Bioinformatics Workflow

Raw files (*.CEL)

Primary Analysis

- Exploratory analysis
- Background correction
- Transformation
- Normalization
- Summarization

Secondary Analysis

- GSEA
- Differential Expression Analysis
- Biomarker Identification
- Gene Regulatory Network

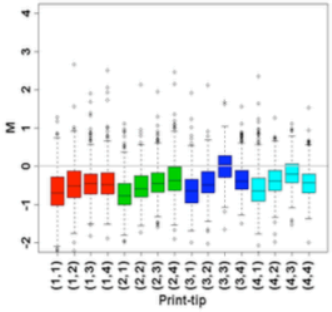
Research Finding

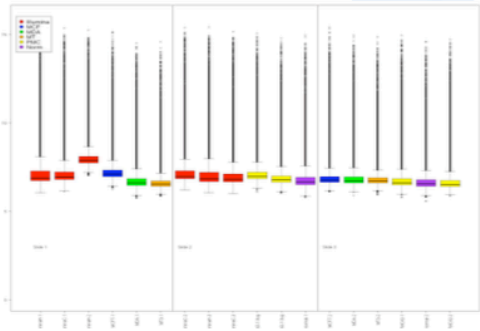
Preprocessing Steps

- Exploratory Analysis
- Background correction
- Transformation
- Normalization
- Summarization

1- Exploratory analysis

Boxplots



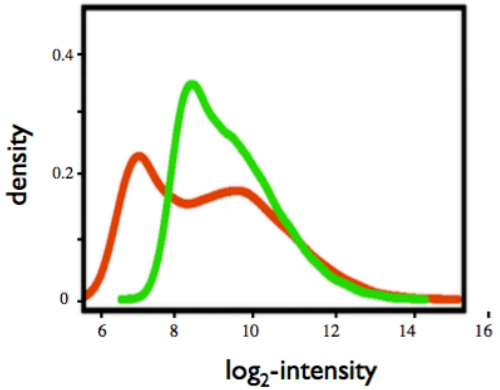


Print-tip / spatial Effects

3 Illumina bead chips

9

Density plots



10

2-What is background correction?

- Background correction involves an attempt to remove any portion of a raw fluorescence intensity measurement that is not attributable to fluorescence from target nucleic acid molecules hybridized to their complementary probe.

- Example sources of fluorescence other than hybridized target nucleic acid molecules include fluorescence in the microarray slide itself, fluorescence from neighboring probe spots, or fluorescence from unbound labeled nucleic acid sequences or other stray particles not washed from the slide.

3-What is Normalization?

- Normalization describes the process of removing (or minimizing) non-biological variation in measured signal intensity levels so that biological differences in gene expression can be appropriately detected.

- Normalization does not necessarily have anything to do with the normal distribution that plays a prominent role in statistics.

Sources of Non-Biological Variation

- Variation across replicate microarray slides resulting from the manufacturing process
- Dye variation: differences in heat and light sensitivity of dyes and differences in the efficiency of dye incorporation
- Variation in the preparation of target samples

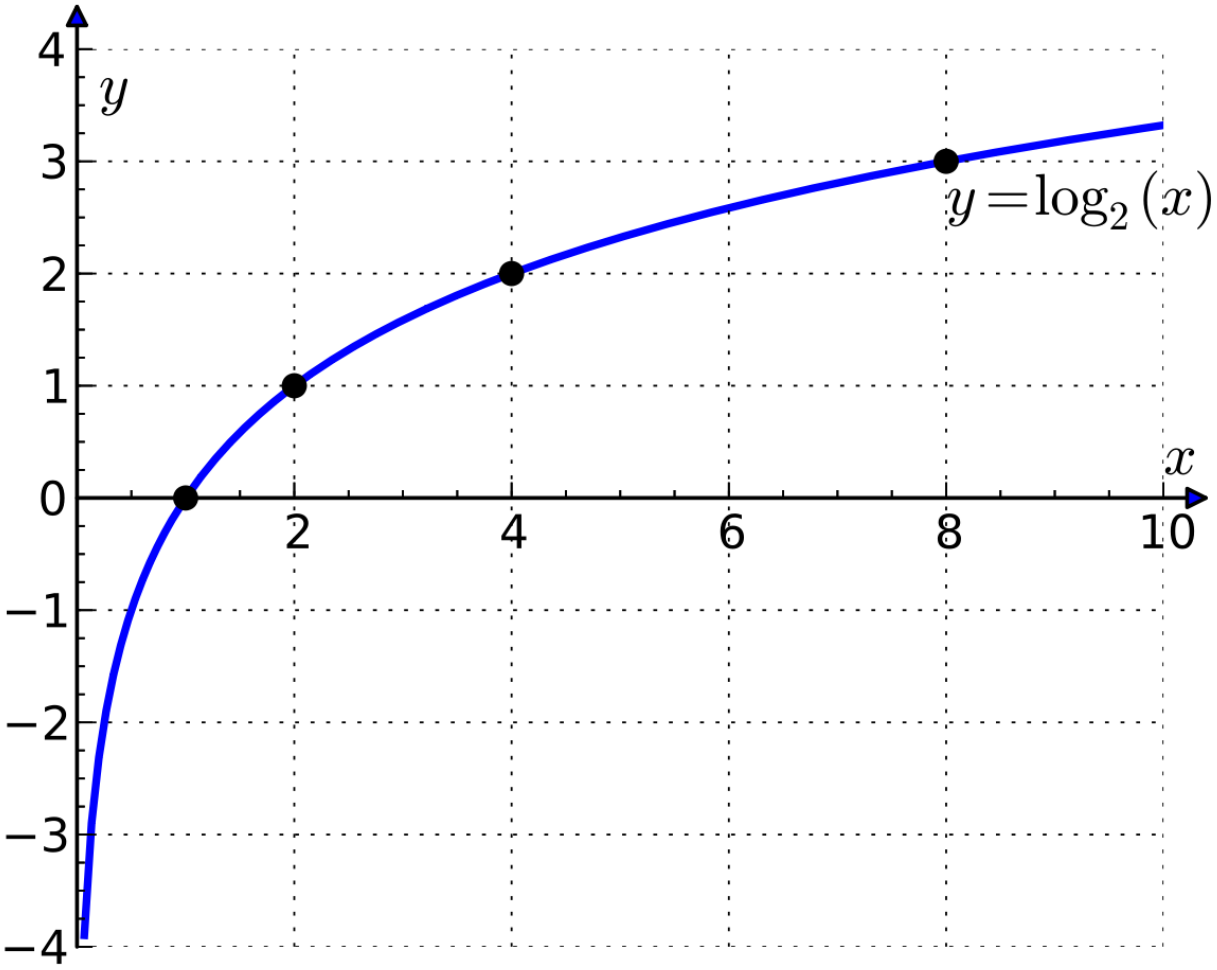
Sources of Non-Biological Variation (continued)

- Variation across various steps in the measurement process, such as hybridization, washing, and microarray image acquisition
- Variation in laboratory conditions from day to day
- Variation among technicians doing the lab work
- etc.

4-What is transformation?

- Transformation refers to transforming the gene expression measures (usually after background correction).
- The most commonly used transformation is the log transformation.
- The base is irrelevant, but log base 2 is popular for microarray data.
- More complex transformations have been proposed that are linear for low values and logarithmic for high values.

4-Log2 transformation



5-What is summarization?

- If a gene is represented by multiple probes on a microarray, it may be desirable to combine the measures from multiple probes to obtain a single measure of the gene's expression level.
- Simply computing the mean or median is often reasonable.

Pre-processing

- Background adjustment
- normalization
- Summarization
- Log transformtion

*eset = **threestep**(affyData_GDS596, background.method = "IdealMM", normalize.method = "quantile", summary.method = "median.polish")*

Eset.log=Log2(eset)

Bioinformatics Workflow (Transcriptomics)

Raw files (CEL,fastq)

Primary Analysis

Secondary Analysis

Research Findings

	Control samples				Study samples			
	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7	Sample8
Gene 1	12	11	15	13	1	2	4	5
Gene 2	7	6	4	0.5	23	21	23	22
Gene 3	5	2	5	6	12.4	14	12	15.5
.....	18	15	15	20	3	2	3	4
.....	19	18	11	17	11	14	12	17
Gene n	5	13	14	22	12	11	23	10

Gene Expression Matrix

- **Differential Expression Analysis**
- **Scoring Methods for the DEGs**
- **Gene set enrichment analysis**
- **Biological Network Construction**

. . . . .

Bioinformatics Workflow

Raw files (*.CEL)

Primary Analysis

Secondary Analysis

Research Findings

	Control samples				Study samples			
	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7	Sample8
Gene 1	12	11	15	13	1	2	4	5
Gene 2	7	6	4	0.5	23	21	23	22
Gene 3	5	2	5	6	12.4	14	12	15.5
.....	18	15	15	20	3	2	3	4
.....	19	18	11	17	11	14	12	17
Gene n	5	13	14	22	12	11	23	10

Gene Expression Matrix

Differential Expression Analysis

Scoring Methods for the DEGs

Gene Expression Matrix

	Control samples				Study samples			
	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7	Sample8
Gene 1	12	11	15	13	1	2	4	5
Gene 2	7	6	4	0.5	23	21	23	22
Gene 3	5	2	5	6	12.4	14	12	15.5
.....	18	15	15	20	3	2	3	4
.....	19	18	11	17	11	14	12	17
Gene n	5	13	14	22	12	11	23	10

Gene Expression Matrix

	Control samples				Study samples			
	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7	Sample8
Gene 1	12	11	15	13	1	2	4	5
Gene 2	7	6	4	0.5	23	21	23	22
Gene 3	5	2	5	6	12.4	14	12	15.5
.....	18	15	15	20	3	2	3	4
.....	19	18	11	17	11	14	12	17
Gene n	5	13	14	22	12	11	23	10

- Each **Row** represents the gene expression profile of a **gene** along all samples/experiments.

Gene Expression Matrix

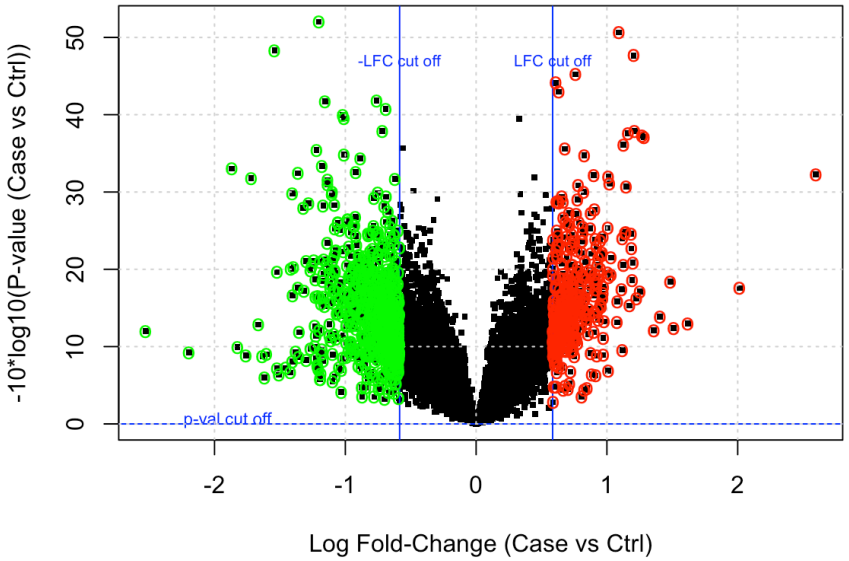
	Control samples				Study samples			
	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7	Sample8
Gene 1	12	11	15	13	1	2	4	5
Gene 2	7	6	4	0.5	23	21	23	22
Gene 3	5	2	5	6	12.4	14	12	15.5
.....	18	15	15	20	3	2	3	4
.....	19	18	11	17	11	14	12	17
Gene n	5	13	14	22	12	11	23	10

- Each **Row** represents the gene expression profile of a **gene** along all samples/experiments.
- Each **Column** represents all the gene expression levels from a single **sample**/experiment.

Differential Expression Analysis (Volcano plot)

Differentially Expressed Genes (DEGs)

Are those genes whose expressions changed ***significantly*** *between the two conditions*

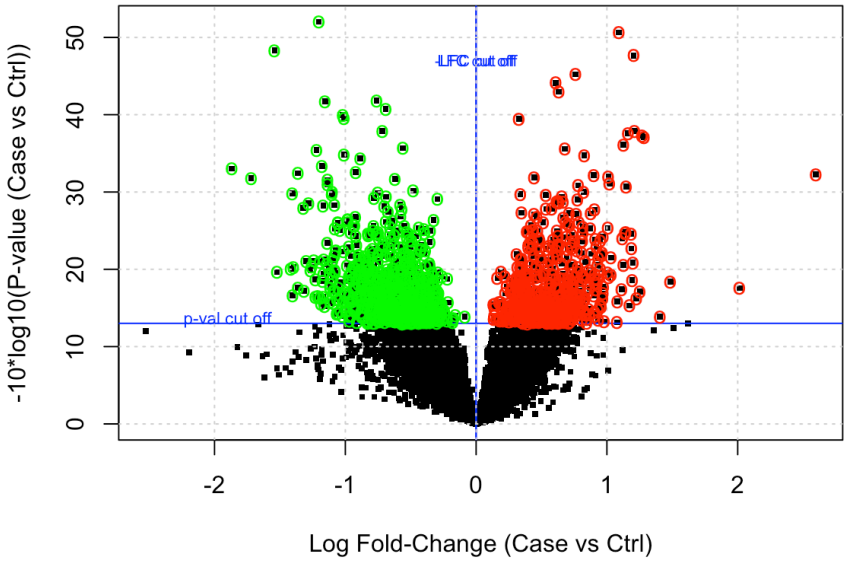


Selection Criteria:

1-Log Fold Change (LFC)= $\log_2 (X_1 / X_2)$

Differential Expression Analysis (Volcano plot)

Differentially Expressed Genes *(**DEGs**)*
Are those genes whose expressions changed ***significantly*** *between the two conditions*



Selection Criteria:

1-Log Fold Change (LFC)= $\log_2 (X_1 / X_2)$

2-Statistical significance (p-value)

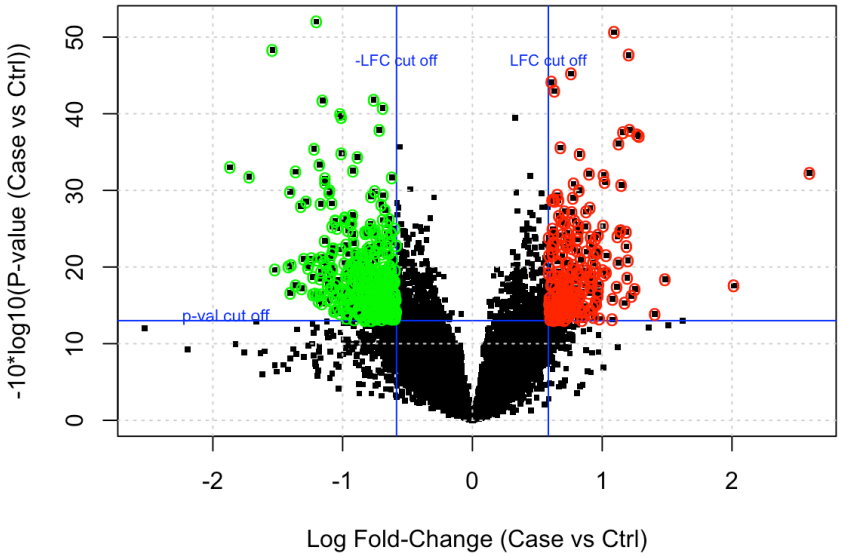
FDR is corrected by BH or Bonferroni

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\left(\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}\right)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$

Differential Expression Analysis (Volcano plot)

Differentially Expressed Genes (DEGs)

Are those genes whose expressions changed ***significantly*** *between the two conditions*



Selection Criteria:

1-Log Fold Change (LFC)= log2 (X1 / X2)

2-Statistical significance (p-value)

FDR is corrected by BH or Bonferroni

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{(N_1-1)s_1^2 + (N_2-1)s_2^2}{N_1+N_2-2}\right)\left(\frac{1}{N_1}+\frac{1}{N_2}\right)}}$$

3- 1 and 2 together

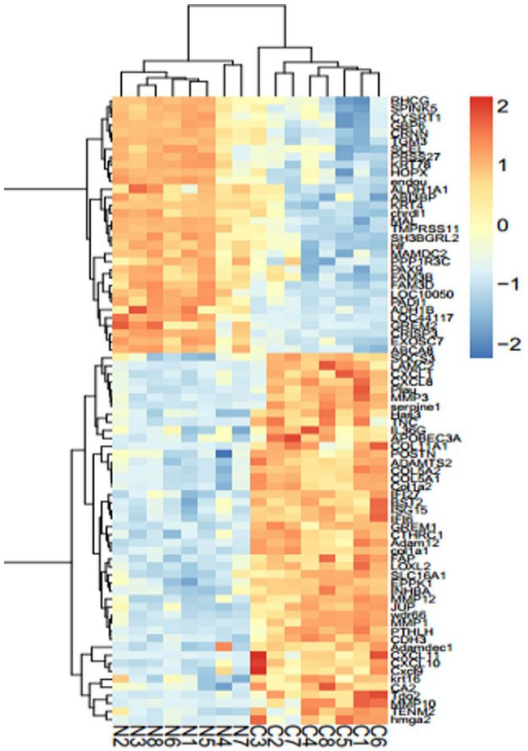
Differential Expression Analysis (Heatmap)

Differentially Expressed Genes *(**DEGs**)*

Are those genes whose expressions

changed ***significantly***

between the two conditions



Selection Criteria:

1-Log Fold Change (LFC)= log2 (X1 / X2)

2-Statistical significance (p-value)

FDR is corrected by BH or Bonferroni

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{(N_1-1)s_1^2 + (N_2-1)s_2^2}{N_1 + N_2 - 2}\right)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$

3- 1 and 2 together

Bioinformatics Workflow

Raw files (*.CEL)

Primary Analysis

Secondary Analysis

Research Findings

	Control samples				Study samples			
	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7	Sample8
Gene 1	12	11	15	13	1	2	4	5
Gene 2	7	6	4	0.5	23	21	23	22
Gene 3	5	2	5	6	12.4	14	12	15.5
.....	18	15	15	20	3	2	3	4
.....	19	18	11	17	11	14	12	17
Gene n	5	13	14	22	12	11	23	10

Gene Expression Matrix

Differential Expression Analysis

Scoring Methods for the DEGs

Biomarker Detection from DEGs

- **Biomarker**: A characteristic (molecule, gene, miRNA, or a mutational signature)…. that is measured and evaluated as an indicator of a biologica process or a disease state.
- **Ranking and scoring the DEGs**
  - How many DEGs are confirmed by other external datasets.
  - Ranking Criteria : based on tissue expression level (Gene Expression Atlas database)
  - DEGs which are absent/non-detected (A) in all tissues except the tissue of interest (Prostate), get **high score**
  - DEGs which are present in all tissues, get **low score**
  - Which DEGs are known to be prostate-specific genes.
  - Involved in disease-related pathways. (mTOR pathway)
  - Confirmed by other assays/ protocols (ex: PCR)

MS Arredouani et al 2009

Bioinformatics Workflow

Raw files (CEL,fastq)

Primary Analysis

Secondary Analysis

Research Findings

	Control samples				Study samples			
	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7	Sample8
Gene 1	12	11	15	13	1	2	4	5
Gene 2	7	6	4	0.5	23	21	23	22
Gene 3	5	2	5	6	12.4	14	12	15.5
.....	18	15	15	20	3	2	3	4
.....	19	18	11	17	11	14	12	17
Gene n	5	13	14	22	12	11	23	10

Gene Expression Matrix

- **Differential Expression Analysis**
- **Scoring Methods for the DEGs**
- **Gene set enrichment analysis**
- **Biological Network Construction**
- .
- .
- .
- .
- .

5- Functional enrichment analysis

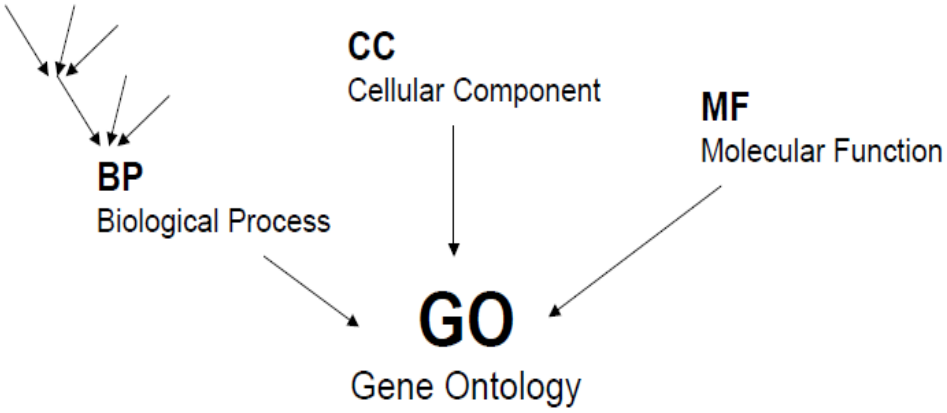
Enrichment analysis of:

- GO terms
- KEGG Pathways
- miRNA targets
- TFBS
- Gene family (MSigDB)
- ……..
- ……..

5- Functional enrichment analysis

Enrichment analysis of:

- **GO terms**
- KEGG Pathways
- miRNA targets
- TFBS
- Gene family (MSigDB)



Ex: GO:0042149 cellular response to glucose starvation

g1, g2, g3, g4, g5,g6,
g7,g8,g9,g10,
g11,g12,g13,g14..g_N

$$P\text{-value} = 1 - \sum_{i=0}^{x} \frac{\binom{k}{i}\binom{M-k}{N-i}}{\binom{M}{N}}$$

DEGs

g1, g2, g3, g4,
g5, g6,.
g…,g10,…,gk

DAVID: Functional Enrichment Analysis

DAVID Bioinformatics Resources 6.8
Laboratory of Human Retrovirology and Immunoinformatics (LHRI)

Home | Start Analysis | Shortcut to DAVID Tools | Technical Center | Downloads & APIs | Term of Service | Why DAVID? | About Us

*** Welcome to DAVID 6.8 ***
*** If you are looking for DAVID 6.7, please visit our development site. ***

Shortcut to DAVID Tools

Functional Annotation
Gene-annotation enrichment analysis, functional annotation clustering , BioCarta & KEGG pathway mapping, gene-disease association, homologue match, ID translation, literature match and more

Gene Functional Classification
Provide a rapid means to reduce large lists of genes into functionally related groups of genes to help unravel the biological content captured by high throughput technologies. More

Gene ID Conversion
Convert list of gene ID/accessions to others of your choice with the most comprehensive gene ID mapping repository. The ambiguous accessions in the list can also be determined semi-automatically. More

Gene Name Batch Viewer
Display gene names for a given gene list; Search functionally related genes within your list or not in your list; Deep links to enriched detailed information. More

Recommending: A paper published in ***Nature Protocols*** describes step-by-step procedure to use DAVID!

Welcome to DAVID 6.8

2003 - 2018

The **D**atabase for **A**nnotation, **V**isualization and **I**ntegrated **D**iscovery (**DAVID**) v6.8 comprises a full Knowledgebase update to the sixth version of our original web-accessible programs. DAVID now provides a comprehensive set of functional annotation tools for investigators to understand biological meaning behind large list of genes. For any given gene list, DAVID tools are able to:

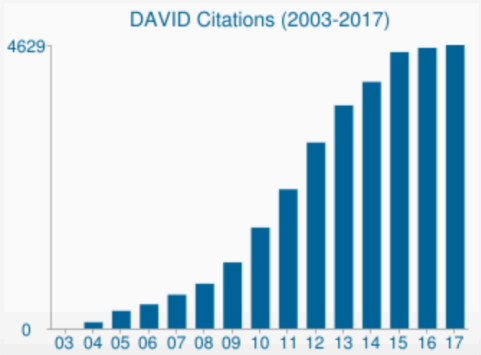
- Identify enriched biological themes, particularly GO terms
- Discover enriched functional-related gene groups
- Cluster redundant annotation terms
- Visualize genes on BioCarta & KEGG pathway maps
- Display related many-genes-to-many-terms on 2-D view.
- Search for other functionally related genes not in the list
- List interacting proteins
- Explore gene names in batch
- Link gene-disease associations
- Highlight protein functional domains and motifs

Search

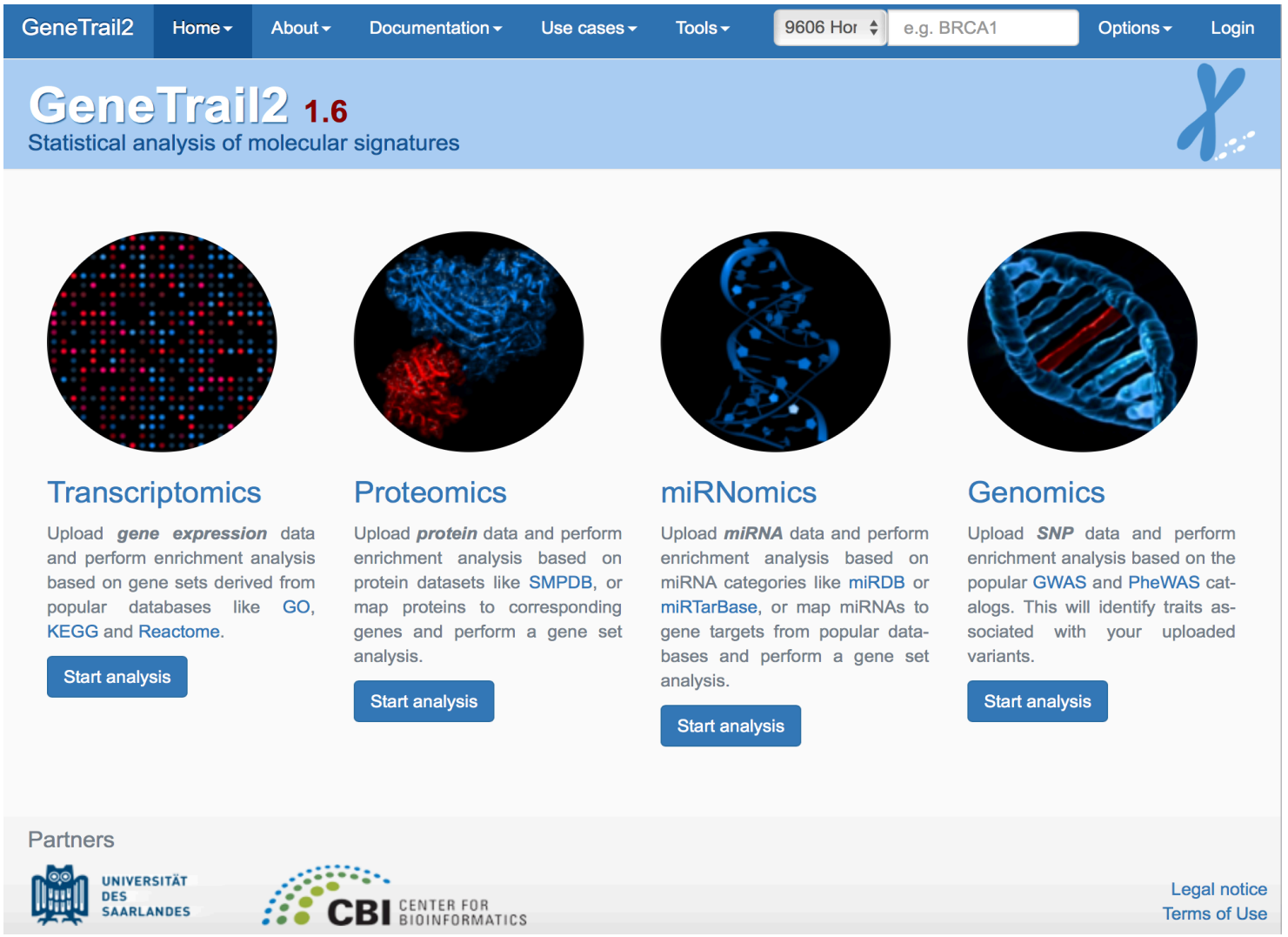
What's Important in DAVID?

- Cite DAVID
- IDs of Affy Exon and Gene arrays supported
- Novel Classification Algorithms
- Pre-built Affymetrix and Illumina backgrounds
- User's customized gene background
- Enhanced calculating speed

Statistics of DAVID



GeneTrail: Functional Enrichment Analysis



YOUR TURN
START WITH THE TUTORIALS

Quick Start Tutorial (clicks)

Task 1

- **Search for the prostate cancer expression dataset with accession number: “GSE55945”.** It is located on the NCBI Gene Expression Omnibus (GEO)

- What is the objective of this study ?

- How many samples (Tumour and Normal) are included in this study?

Task 1

Series GSE55945 Query DataSets for GSE55945

Status	Public on Mar 18, 2014
Title	Gene Expression Profiling of Prostate Benign and Malignant Tissue
Organism	Homo sapiens
Experiment type	Expression profiling by array
Summary	We profiled genome-wide gene expression of human prostate benign and malignant tissue to identify potential biomarkers and immunotherapy targets. We stratified malignant specimens according to their TMPRSS2:ERG gene fusion status.
Overall design	Radical prostatectomy tissue samples were obtained from the Hershey Foundation Prostate Cancer Serum and Tumor Bank at our institution. Morphologic diagnosis was done by a pathologist. OCT blocks containing >30% of PCa tissue (with Gleason score of 6 or 7) were selected for RNA purification. A biopsy punch was used to select the PCa tissues from the OCT sample blocks. Benign or PCa tissues were homogenized using a TissueLyser (Qiagen) at 28 Hz for 5 min. Total RNA was isolated using Trizol reagent. RNA was quantified by NanoDrop ND-1000 spectrophotometer, and quality was evaluated with Agilent RNA 6000 NanoChip and the 2100 Bioanalyzer, with 28S/18S ratios and RIN determined by 2100 Expert software.
Contributor(s)	Arredouani MS, Lu B, Sanda M
Citation(s)	Arredouani MS, Lu B, Bhasin M, Eljanne M et al. Identification of the transcription factor single-minded homologue 2 as a potential biomarker and immunotherapy target in prostate cancer. *Clin Cancer Res* 2009 Sep 15;15(18):5794-802. PMID: 19737960
Submission date	Mar 17, 2014
Last update date	Mar 25, 2019
Contact name	M. Simo Arredouani
E-mail(s)	simarred@gmail.com
Organization name	Beth Israel Deaconess Medical Center
Department	Surgery
Street address	330 Brookline Ave.
City	Boston
State/province	MA
ZIP/Postal code	02215
Country	USA
Platforms (1)	GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array
Samples (21)	GSM1348933 Prostate Cancer_MS 36D6 GSM1348934 Prostate Cancer_MS 36C1

More...

Samples (21)
Less...

GSM1348933	Prostate Cancer_MS 36D6
GSM1348934	Prostate Cancer_MS 36C1
GSM1348935	Prostate Cancer_MS 36D7
GSM1348936	Prostate Cancer_MS 36D8
GSM1348937	Prostate Cancer_MS 36A6
GSM1348938	Prostate Cancer_MS 36C4
GSM1348939	Prostate Cancer_NUGEN TEST07
GSM1348940	Prostate Cancer_MS 36C8
GSM1348941	Prostate Cancer_MS 36A1
GSM1348942	Prostate Cancer_MS 36C2
GSM1348943	Prostate Cancer_MS 36C3
GSM1348944	Prostate Cancer_MS 36C9
GSM1348945	Prostate Cancer_NUGEN TEST05
GSM1348946	Normal_MS 36C6
GSM1348947	Normal_MS 36C7
GSM1348948	Normal_MS 36D2
GSM1348949	Normal_MS 36D3
GSM1348950	Normal_MS 36D4
GSM1348951	Normal_MS 36D5
GSM1348952	Normal_MS 36A4
GSM1348953	Normal_MS 36A5

Task 2

- **Click the button "Analyze with GEO2R" and start to define two groups of samples**

Overall design	Radical prostatectomy tissue samples were obtained from the Hershey Foundation Prostate Cancer Serum and Tumor Bank at our institution. Morphologic diagnosis was done by a pathologist. OCT blocks containing >30% of PCa tissue (with Gleason score of 6 or 7) were selected for RNA purification. A biopsy punch was used to select the PCa tissues from the OCT sample blocks. Benign or PCa tissues were homogenized using a TissueLyser (Qiagen) at 28 Hz for 5 min. Total RNA was isolated using Trizol reagent. RNA was quantified by NanoDrop ND-1000 spectrophotometer, and quality was evaluated with Agilent RNA 6000 NanoChip and the 2100 Bioanalyzer, with 28S/18S ratios and RIN determined by 2100 Expert software.
Contributor(s)	Arredouani MS, Lu B, Sanda M
Citation(s)	Arredouani MS, Lu B, Bhasin M, Eljanne M et al. Identification of the transcription factor single-minded homologue 2 as a potential biomarker and immunotherapy target in prostate cancer. *Clin Cancer Res* 2009 Sep 15;15(18):5794-802. PMID: 19737960
Submission date	Mar 17, 2014
Last update date	Mar 25, 2019
Contact name	M. Simo Arredouani
E-mail(s)	simarred@gmail.com
Organization name	Beth Israel Deaconess Medical Center
Department	Surgery
Street address	330 Brookline Ave.
City	Boston
State/province	MA
ZIP/Postal code	02215
Country	USA
Platforms (1)	GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array
Samples (21)	GSM1348933 Prostate Cancer_MS 36D6
More...	GSM1348934 Prostate Cancer_MS 36C1
	GSM1348935 Prostate Cancer_MS 36D7

Relations

BioProject PRJNA241405

Analyze with GEO2R

Task 2

- **Click the button "Analyze with GEO2R" and start to define two groups of samples**

Use GEO2R to compare two or more groups of Samples in order to identify genes that are differentially expressed across experimental conditions. Results are presented as a table of genes ordered by significance.
Full instructions

GEO accession GSE55945 Set Gene Expression Profiling of Prostate Benign and Malignant Tissue

Samples Define groups Selected **0** out of **21** samples

Enter a group name: List

Cancel selection
Normal
Cancer

Columns Set

Group	Accession		Source name	Tissue type	Tmprss2
-	GSM1348933		malignant tissue_TMPRSS2:ERG fusion negative	Malignant prostate tissue	ERG fusion status: negative
-	GSM1348934		malignant tissue_TMPRSS2:ERG fusion negative	Malignant prostate tissue	ERG fusion status: negative
-	GSM1348935		malignant tissue_TMPRSS2:ERG fusion negative	Malignant prostate tissue	ERG fusion status: negative
-	GSM1348936	Prostate Cancer_MS 36D8	malignant tissue_TMPRSS2:ERG fusion negative	Malignant prostate tissue	ERG fusion status: negative
-	GSM1348937	Prostate Cancer_MS 36A6	malignant tissue_TMPRSS2:ERG fusion negative	Malignant prostate tissue	ERG fusion status: negative
-	GSM1348938	Prostate Cancer_MS 36C4	malignant tissue_TMPRSS2:ERG fusion negative	Malignant prostate tissue	ERG fusion status: negative
-	GSM1348939	Prostate Cancer_NUGEN TEST07	malignant tissue_TMPRSS2:ERG fusion negative	Malignant prostate tissue	ERG fusion status: negative
-	GSM1348940	Prostate Cancer_MS 36C8	malignant tissue_TMPRSS2:ERG fusion positive	Malignant prostate tissue	ERG fusion status: positive
-	GSM1348941	Prostate Cancer_MS 36A1	malignant tissue_TMPRSS2:ERG fusion positive	Malignant prostate tissue	ERG fusion status: positive
-	GSM1348942	Prostate Cancer_MS 36C2	malignant tissue_TMPRSS2:ERG fusion positive	Malignant prostate tissue	ERG fusion status: positive
-	GSM1348943	Prostate Cancer_MS 36C3	malignant tissue_TMPRSS2:ERG fusion positive	Malignant prostate tissue	ERG fusion status: positive
-	GSM1348944	Prostate Cancer_MS 36C9	malignant tissue_TMPRSS2:ERG fusion positive	Malignant prostate tissue	ERG fusion status: positive
-	GSM1348945	Prostate Cancer_NUGEN TEST05	malignant tissue_TMPRSS2:ERG fusion positive	Malignant prostate tissue	ERG fusion status: positive
-	GSM1348946	Normal_MS 36C6	benign tissue	Benign prostate tissue	
-	GSM1348947	Normal_MS 36C7	benign tissue	Benign prostate tissue	
-	GSM1348948	Normal_MS 36D2	benign tissue	Benign prostate tissue	
-	GSM1348949	Normal_MS 36D3	benign tissue	Benign prostate tissue	
-	GSM1348950	Normal_MS 36D4	benign tissue	Benign prostate tissue	

Task 3

- **Run the analysis and examine the top 250 DEGs (potential markers)**
- What is LFC for the genes DXL1 and DXL2 ?

GEO accession GSE55945 Set Gene Expression Profiling of Prostate Benign and Malignant Tissue

Samples Define groups Selected **21** out of **21** samples

Enter a group name: List

Cancel selection

Normal (8 samples)

Cancer (13 samples)

Columns Set

Cancer	GSM1348934	Cancer_MS 36C1	malignant tissue_TMPRSS2:ERG fusion negative	Malignant prostate tissue	ERG fusion status: negative
Cancer	GSM1348935	Cancer_MS 36D7	malignant tissue_TMPRSS2:ERG fusion negative	Malignant prostate tissue	ERG fusion status: negative
Cancer	GSM1348936	Cancer_MS 36D8	malignant tissue_TMPRSS2:ERG fusion negative	Malignant prostate tissue	ERG fusion status: negative
Cancer	GSM1348937	Cancer_MS 36A6	malignant tissue_TMPRSS2:ERG fusion negative	Malignant prostate tissue	ERG fusion status: negative
Cancer	GSM1348938	Prostate Cancer_MS 36C4	malignant tissue_TMPRSS2:ERG fusion negative	Malignant prostate tissue	ERG fusion status: negative
Cancer	GSM1348939	Prostate Cancer_NUGEN TEST07	malignant tissue_TMPRSS2:ERG fusion negative	Malignant prostate tissue	ERG fusion status: negative
Cancer	GSM1348940	Prostate Cancer_MS 36C8	malignant tissue_TMPRSS2:ERG fusion positive	Malignant prostate tissue	ERG fusion status: positive
Cancer	GSM1348941	Prostate Cancer_MS 36A1	malignant tissue_TMPRSS2:ERG fusion positive	Malignant prostate tissue	ERG fusion status: positive
Cancer	GSM1348942	Prostate Cancer_MS 36C2	malignant tissue_TMPRSS2:ERG fusion positive	Malignant prostate tissue	ERG fusion status: positive
Cancer	GSM1348943	Prostate Cancer_MS 36C3	malignant tissue_TMPRSS2:ERG fusion positive	Malignant prostate tissue	ERG fusion status: positive
Cancer	GSM1348944	Prostate Cancer_MS 36C9	malignant tissue_TMPRSS2:ERG fusion positive	Malignant prostate tissue	ERG fusion status: positive
Cancer	GSM1348945	Prostate Cancer_NUGEN TEST05	malignant tissue_TMPRSS2:ERG fusion positive	Malignant prostate tissue	ERG fusion status: positive
Normal	GSM1348946	Normal_MS 36C6	benign tissue	Benign prostate tissue	
Normal	GSM1348947	Normal_MS 36C7	benign tissue	Benign prostate tissue	
Normal	GSM1348948	Normal_MS 36D2	benign tissue	Benign prostate tissue	
Normal	GSM1348949	Normal_MS 36D3	benign tissue	Benign prostate tissue	
Normal	GSM1348950	Normal_MS 36D4	benign tissue	Benign prostate tissue	
Normal	GSM1348951	Normal_MS 36D5	benign tissue	Benign prostate tissue	
Normal	GSM1348952	Normal_MS 36A4	benign tissue	Benign prostate tissue	
Normal	GSM1348953	Normal_MS 36A5	benign tissue	Benign prostate tissue	

GEO2R | Value distribution | Options | Profile graph | R script

Quick start

- Specify a GEO Series accession and a Platform if prompted.
- Click 'Define groups' and enter names for the groups of Samples you plan to compare, e.g., test and control.
- Assign Samples to each group. Highlight Sample rows then click the group name to assign those Samples to the group. Use the Sample metadata (title, source and characteristics) columns to help determine which Samples belong to which group.
- Click 'Top 250' to perform the calculation with default settings.
- Results are presented as a table of genes ordered by significance. The top 250 genes are presented and may be viewed as profile graphs. Alternatively, the complete results table may be saved.
- You may change settings in Options tab.

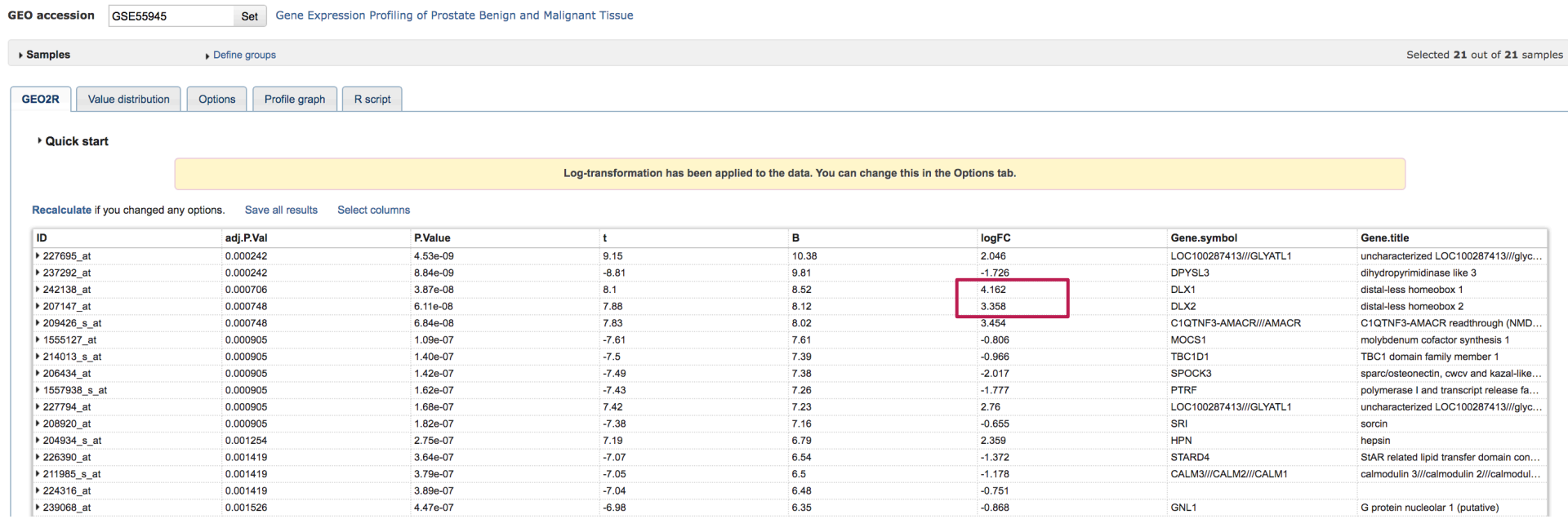
Top 250 Save all results

Task 3

- Run the analysis and examine the top 250 DEGs (potential markers)
- What are the LFC values for the genes DXL1 and DXL2 ?
- Show the gene expression profiles for both genes across all samples?

Task 3

- Run the analysis and examine the top 250 DEGs (potential markers)
- **What are the LFC values for the genes DXL1 and DXL2 ?**



GEO accession GSE55945 Set Gene Expression Profiling of Prostate Benign and Malignant Tissue

Samples Define groups Selected 21 out of 21 samples

GEO2R | Value distribution | Options | Profile graph | R script

Quick start

Log-transformation has been applied to the data. You can change this in the Options tab.

Recalculate if you changed any options. Save all results Select columns

ID	adj.P.Val	P.Value	t	B	logFC	Gene.symbol	Gene.title
227695_at	0.000242	4.53e-09	9.15	10.38	2.046	LOC100287413///GLYATL1	uncharacterized LOC100287413///glyc…
237292_at	0.000242	8.84e-09	-8.81	9.81	-1.726	DPYSL3	dihydropyrimidinase like 3
242138_at	0.000706	3.87e-08	8.1	8.52	4.162	DLX1	distal-less homeobox 1
207147_at	0.000748	6.11e-08	7.88	8.12	3.358	DLX2	distal-less homeobox 2
209426_s_at	0.000748	6.84e-08	7.83	8.02	3.454	C1QTNF3-AMACR///AMACR	C1QTNF3-AMACR readthrough (NMD…
1555127_at	0.000905	1.09e-07	-7.61	7.61	-0.806	MOCS1	molybdenum cofactor synthesis 1
214013_s_at	0.000905	1.40e-07	-7.5	7.39	-0.966	TBC1D1	TBC1 domain family member 1
206434_at	0.000905	1.42e-07	-7.49	7.38	-2.017	SPOCK3	sparc/osteonectin, cwcv and kazal-like…
1557938_s_at	0.000905	1.62e-07	-7.43	7.26	-1.777	PTRF	polymerase I and transcript release fa…
227794_at	0.000905	1.68e-07	7.42	7.23	2.76	LOC100287413///GLYATL1	uncharacterized LOC100287413///glyc…
208920_at	0.000905	1.82e-07	-7.38	7.16	-0.655	SRI	sorcin
204934_s_at	0.001254	2.75e-07	7.19	6.79	2.359	HPN	hepsin
226390_at	0.001419	3.64e-07	-7.07	6.54	-1.372	STARD4	StAR related lipid transfer domain con…
211985_s_at	0.001419	3.79e-07	-7.05	6.5	-1.178	CALM3///CALM2///CALM1	calmodulin 3///calmodulin 2///calmodul…
224316_at	0.001419	3.89e-07	-7.04	6.48	-0.751		
239068_at	0.001526	4.47e-07	-6.98	6.35	-0.868	GNL1	G protein nucleolar 1 (putative)

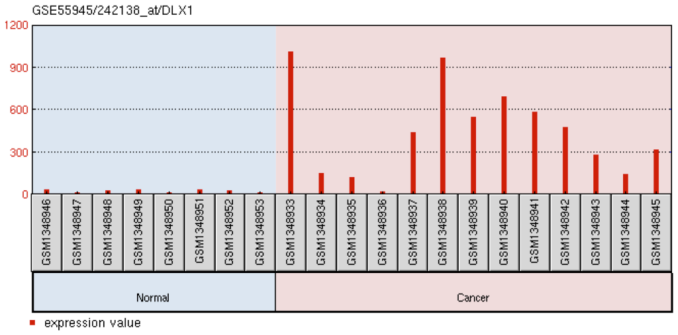
Task 3

- Run the analysis and examine the top 250 DEGs (potential markers)
- What are the LFC values for the genes DXL1 and DXL2 ?

Recalculate if you changed any options. Save all results Select columns

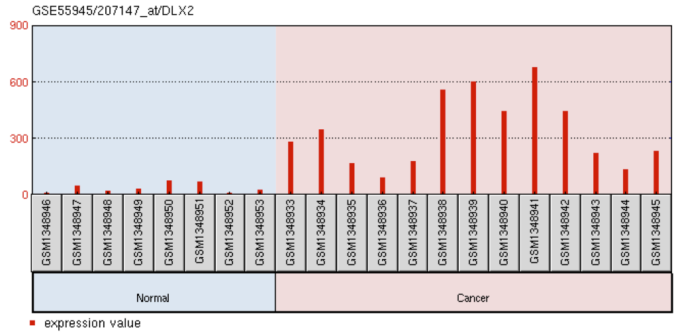
ID	adj.P.Val	P.Value	t	B	logFC	Gene.symbol	Gene.title
227695_at	0.000242	4.53e-09	9.15	10.38	2.046	LOC100287413///GLYATL1	uncharacterized LOC100287413///glyc…
237292_at	0.000242	8.84e-09	-8.81	9.81	-1.726	DPYSL3	dihydropyrimidinase like 3
242138_at	0.000706	3.87e-08	8.1	8.52	4.162	DLX1	distal-less homeobox 1

Sample values



207147_at	0.000748	6.11e-08	7.88	8.12	3.358	DLX2	distal-less homeobox 2

Sample values



209426_s_at	0.000748	6.84e-08	7.83	8.02	3.454	C1QTNF3-AMACR///AMACR	C1QTNF3-AMACR readthrough (NMD…
1555127_at	0.000905	1.09e-07	-7.61	7.61	-0.806	MOCS1	molybdenum cofactor synthesis 1

Task 4

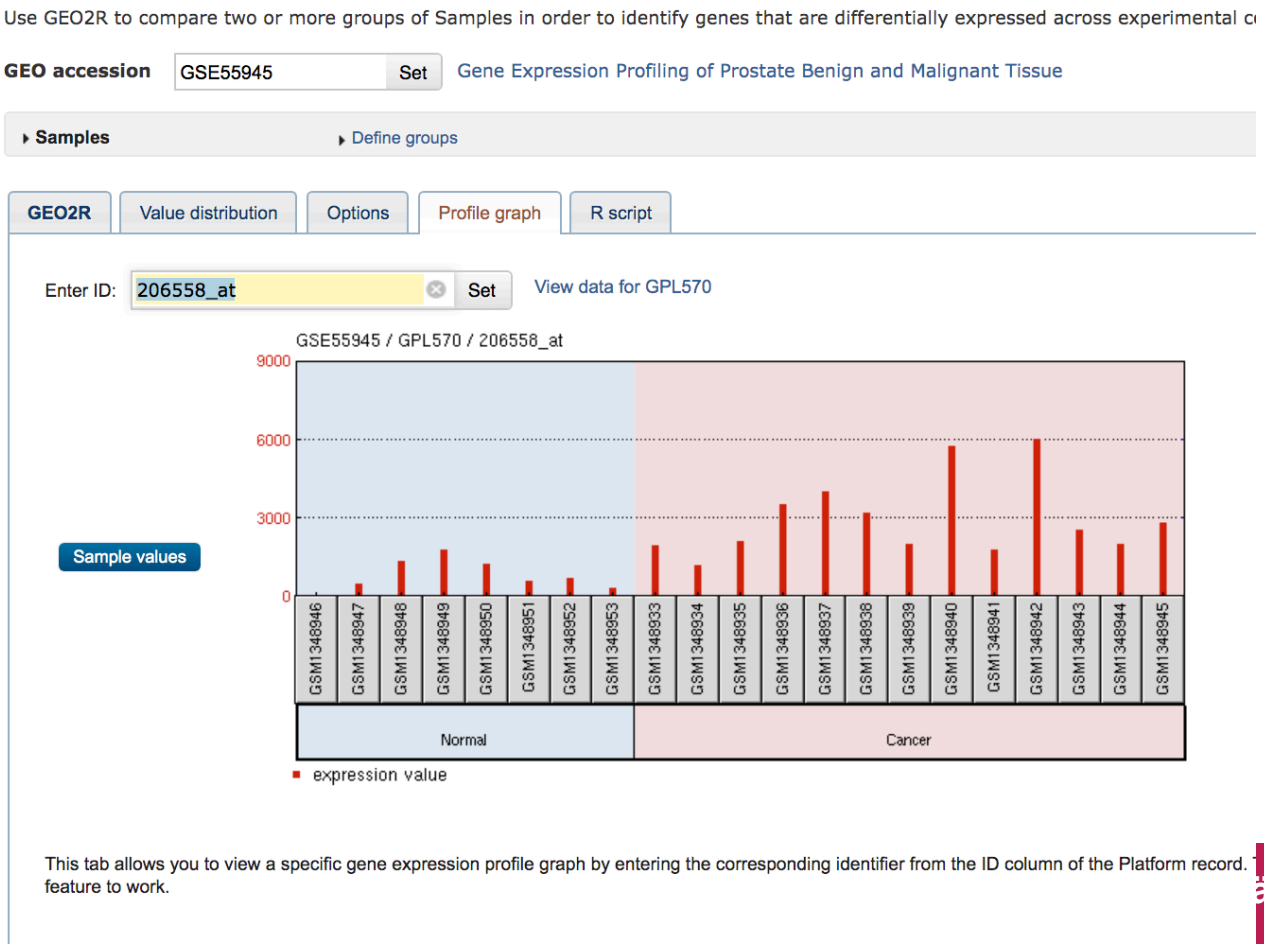
- Show the expression profile for the gene “SIM2”

  Tip: The corresponding **Probe ID** for SIM2 is : **206558_at**.

Task 4

- Show the expression profile for the gene “SIM2”

  Tip: The corresponding **Probe ID** for SIM2 is : **206558_at**.



Task 5 (Bonus)

- Redefine the two sample groups based on the prostate cancer subtype (TMPRSS2:ERG fusion gene)
- What is the LFC and adj-pval of the ERG gene?
- And look at the expression profile of ERG across all samples. Can we consider it as a Biomarker?