Vall d'Hebron
Institut de Recerca
VHIR

Vall d'Hebron Institut de Recerca

Ricardo Gonzalo
ricardo.gonzalo@vhir.org

Alex Sánchez
alex.sanchez@vhir.org

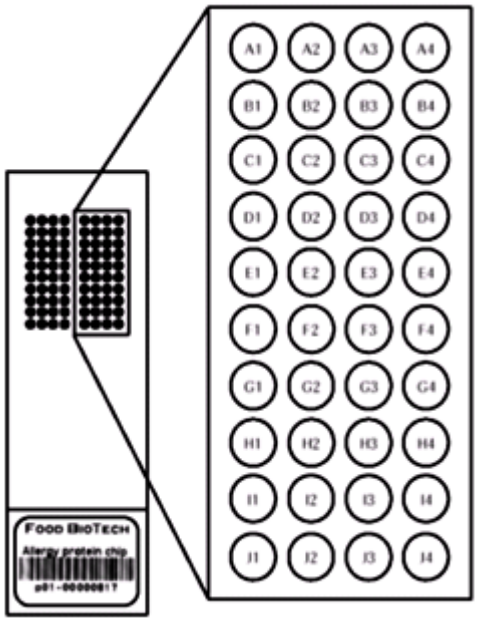
High Throughput Data Analysis. Microarrays Biological Significance

Table of contents

Table of contents

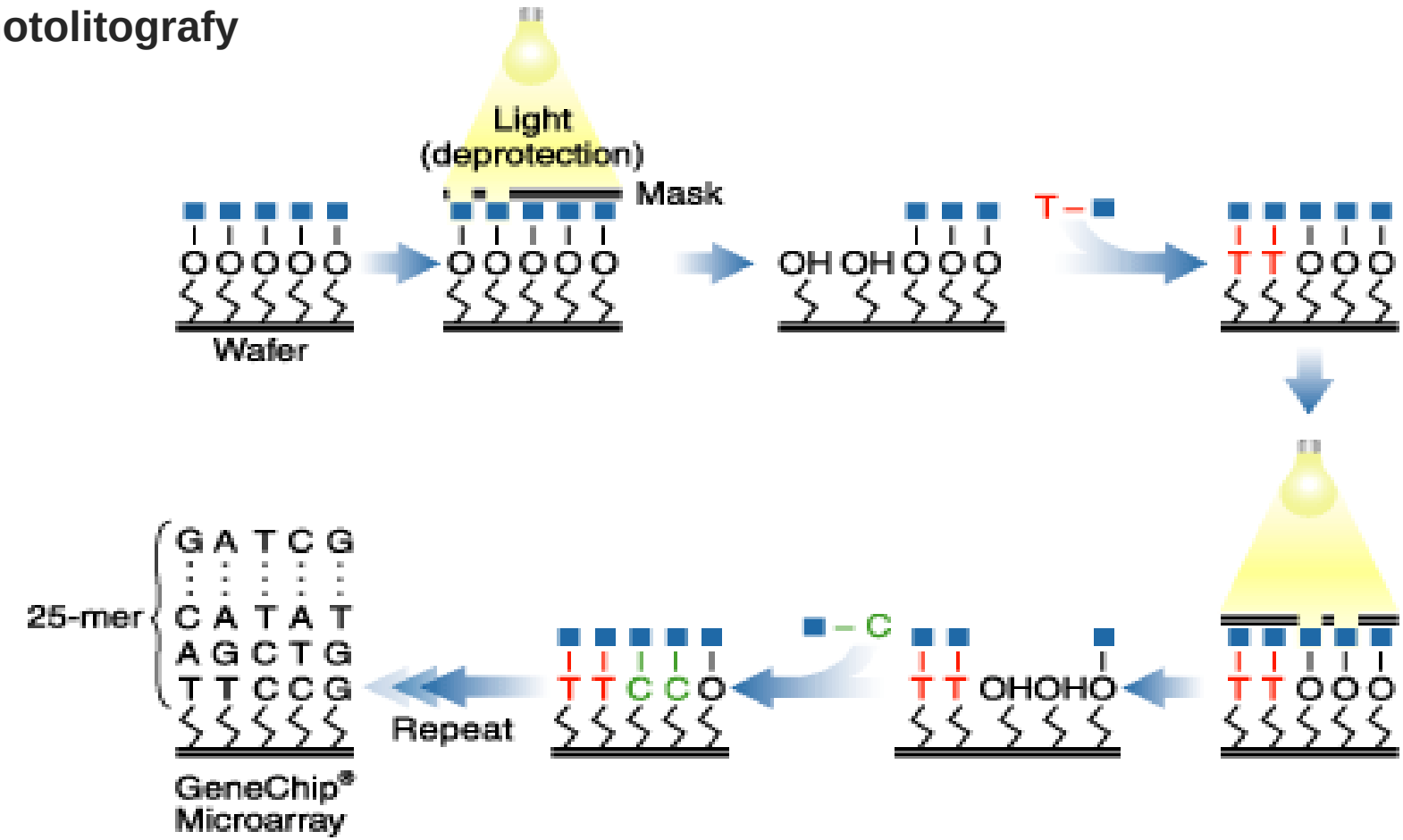
A microarray (of nucleic acids) in few words?

- DNA fixed to a solid surface (nylon, silica, glass,...)
  - is called ***probe***
- RNA "problem" is labelled, and have to bind to DNA fixed in the solid surface in an specific way.
  - is called ***target***

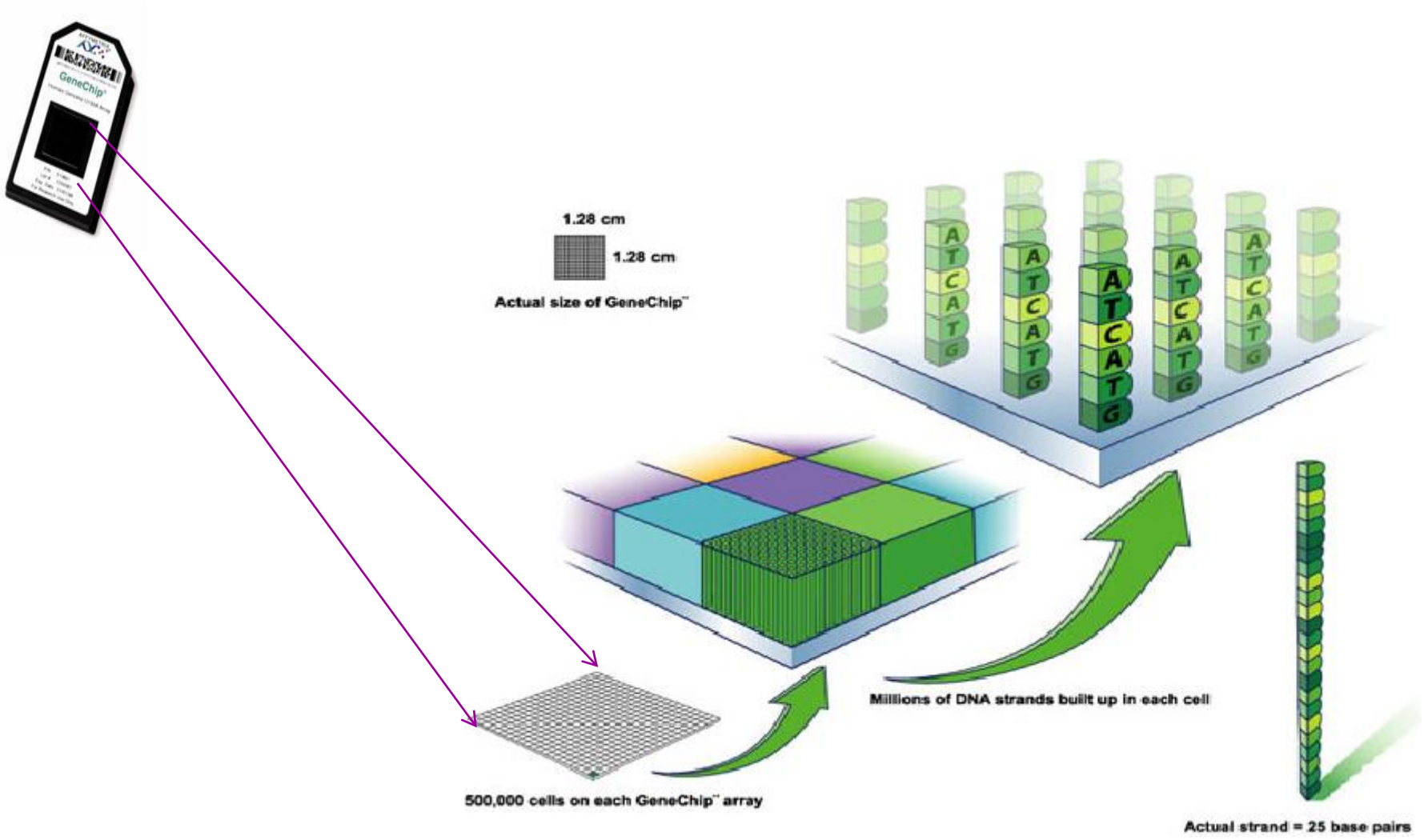


1. Introduction to microarrays technology

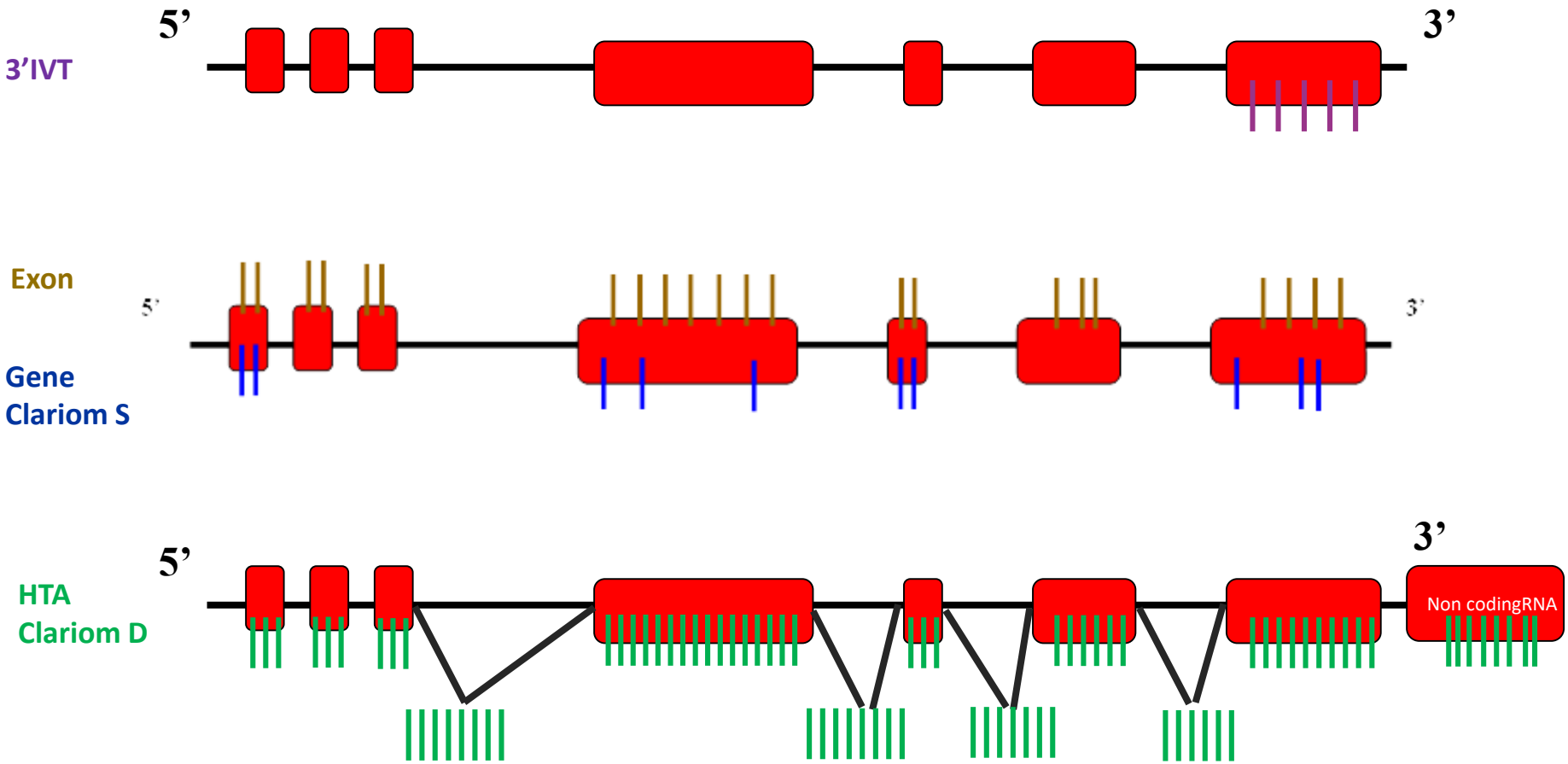
Photolitografy



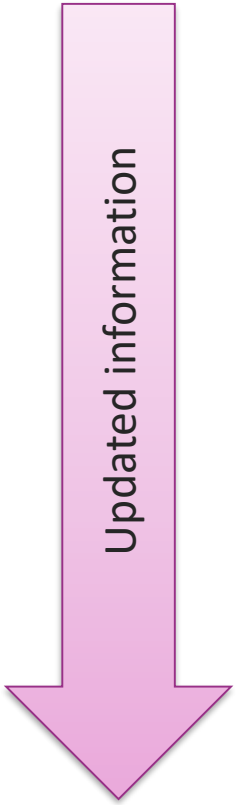
1. Introduction to microarrays technology



1. Introduction to microarrays technology



1. Introduction to microarrays technology



3' IVT ARRAYS

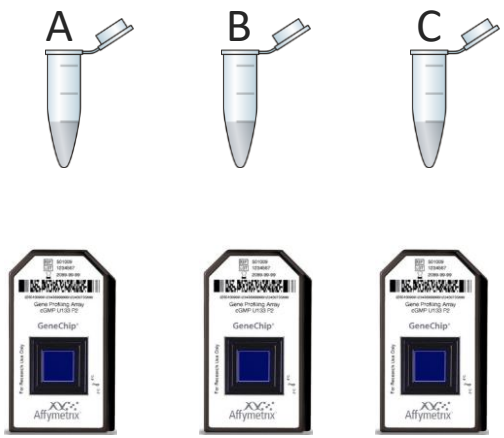
GENE/EXON ARRAYS

HUMAN TRANSCRIPTOME ARRAYS

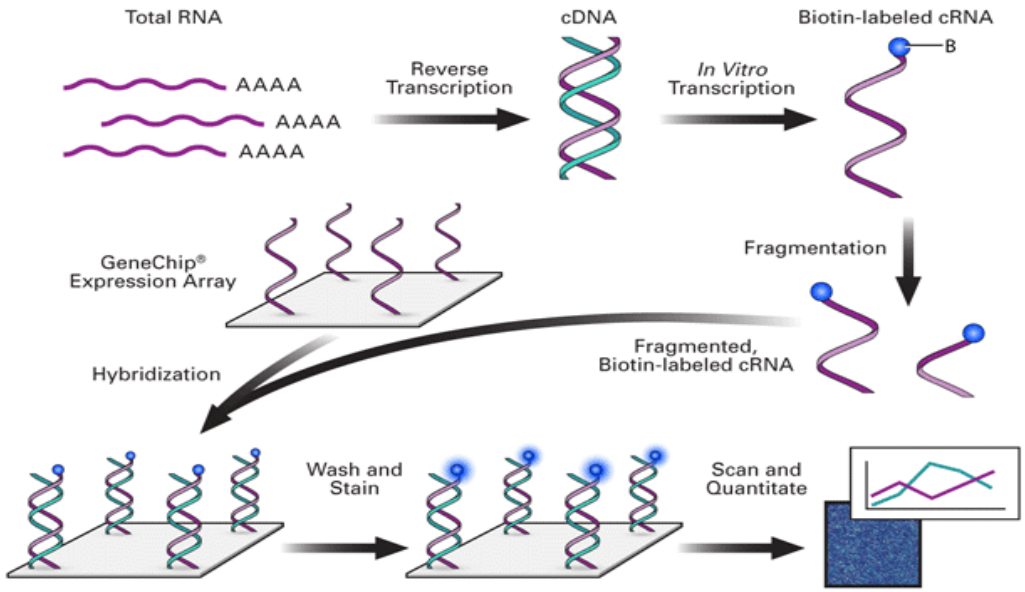
CLARIOM ARRAYS

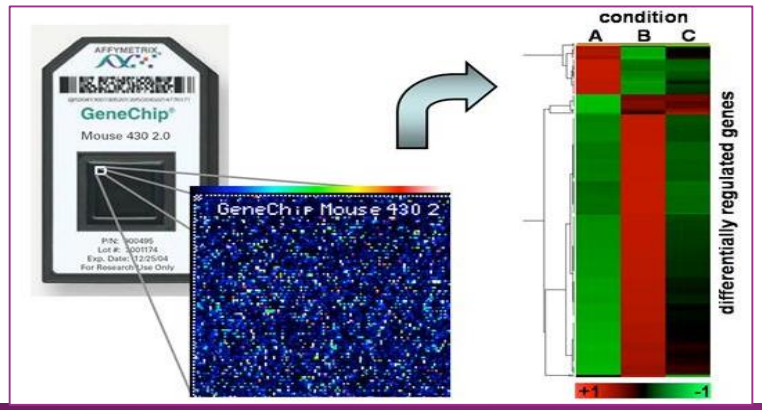
Cost increase with the amount of information analyzed in the array

1. Introduction to microarrays technology



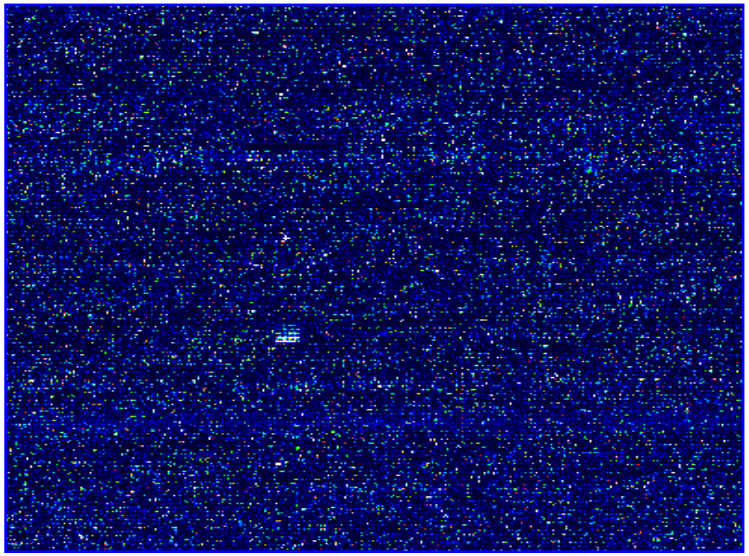
One sample hybridized per array





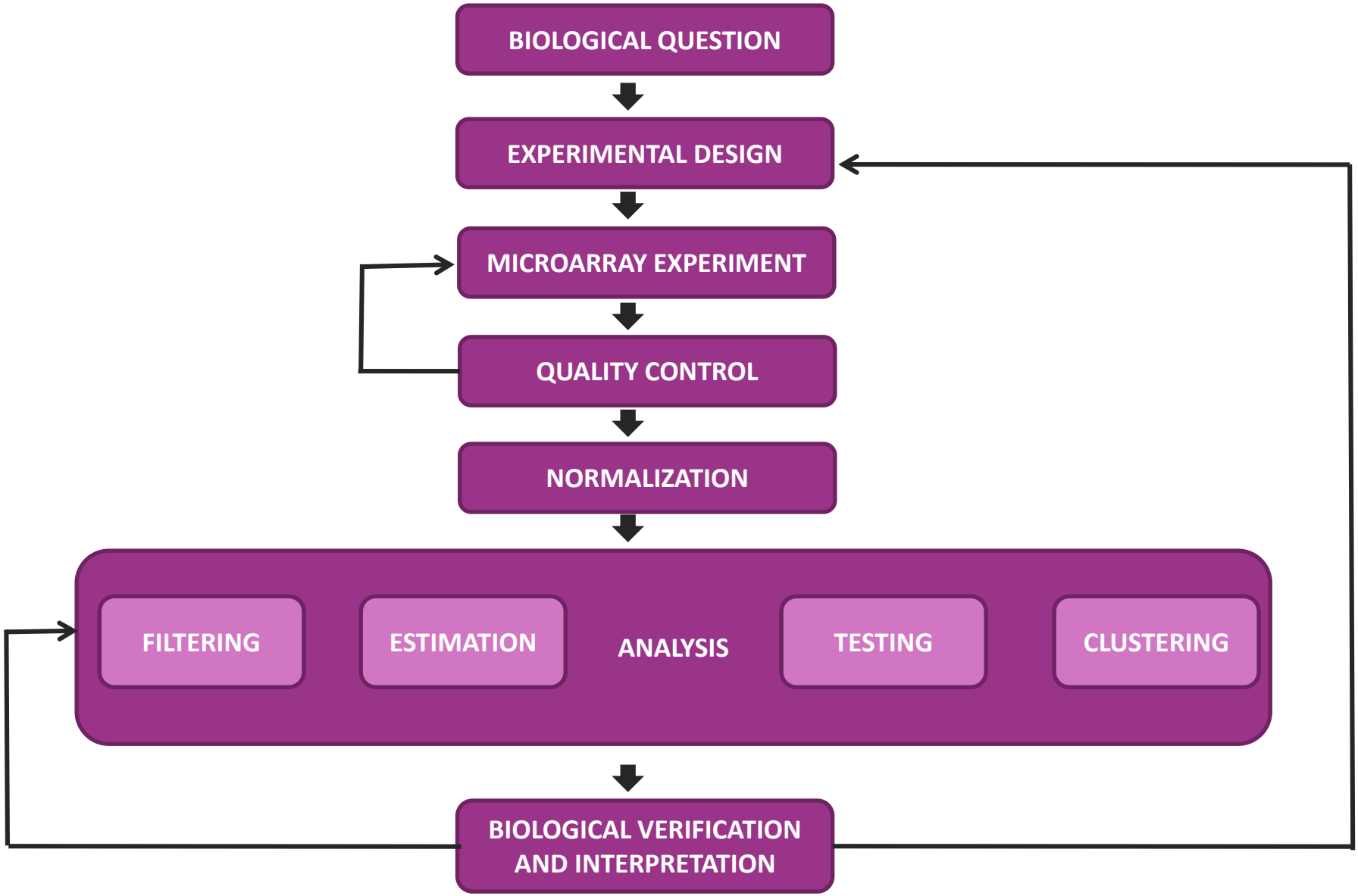
The sample is stained with **one dye (absolute fluorescence measure)**

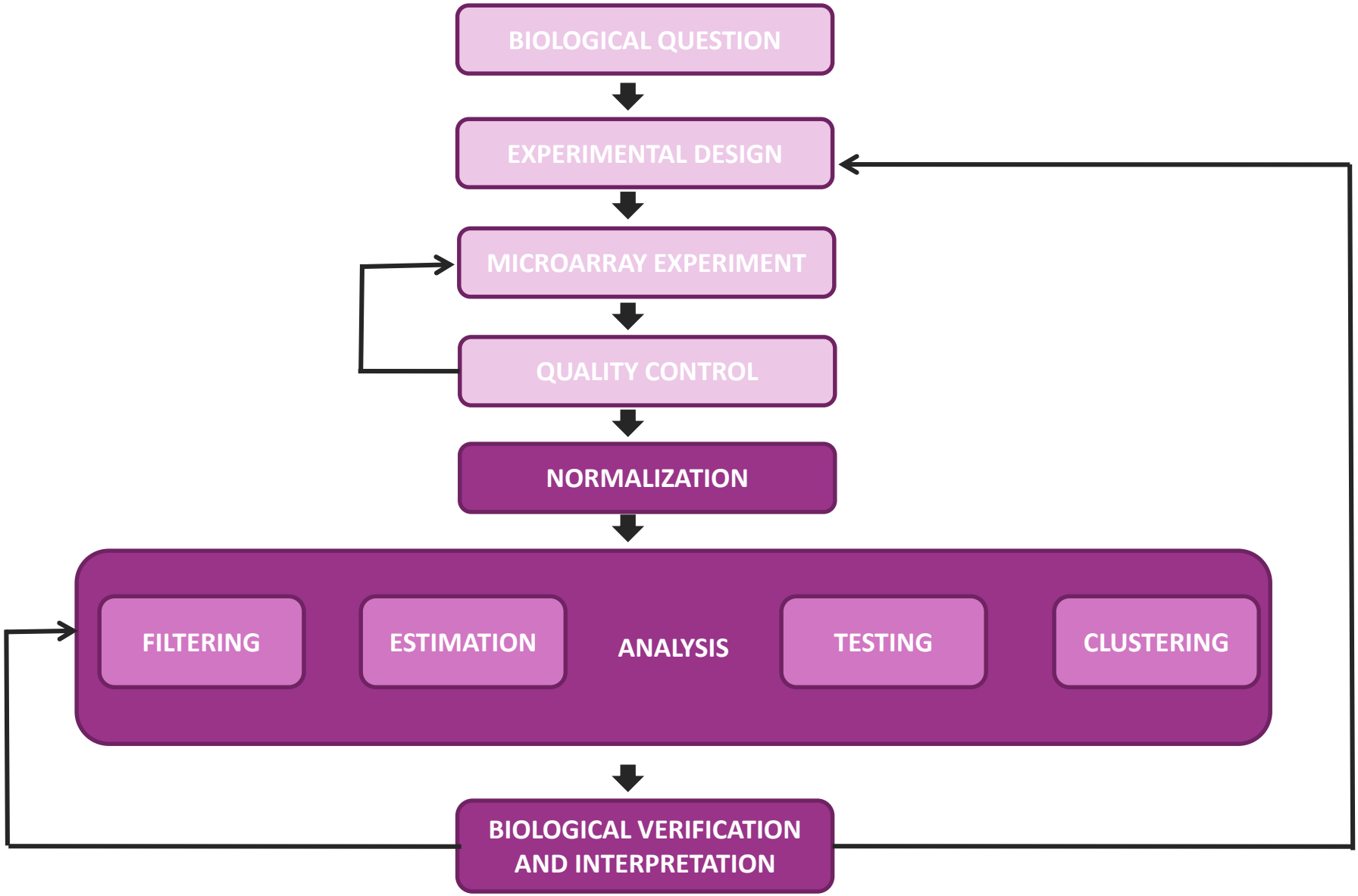
1. Introduction to microarrays technology



"CEL" FILES

Table of contents


BIOLOGICAL QUESTION
EXPERIMENTAL DESIGN
MICROARRAY EXPERIMENT
QUALITY CONTROL
NORMALIZATION
FILTERING
ESTIMATION
ANALYSIS
TESTING
CLUSTERING
BIOLOGICAL VERIFICATION AND INTERPRETATION


BIOLOGICAL QUESTION
EXPERIMENTAL DESIGN
MICROARRAY EXPERIMENT
QUALITY CONTROL
NORMALIZATION
FILTERING
ESTIMATION
ANALYSIS
TESTING
CLUSTERING
BIOLOGICAL VERIFICATION AND INTERPRETATION

2. Microarray data analysis with R. QUALITY CONTROL OF THE DATA

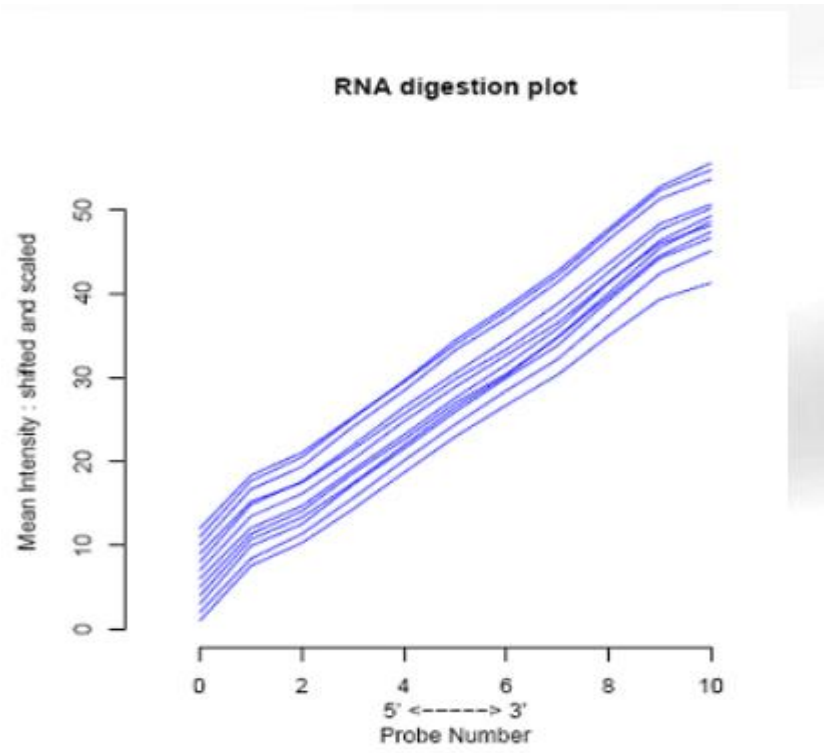
- First of all we have to decide if the data are good to work with.
- Microarray experiments generate huge quantities of data

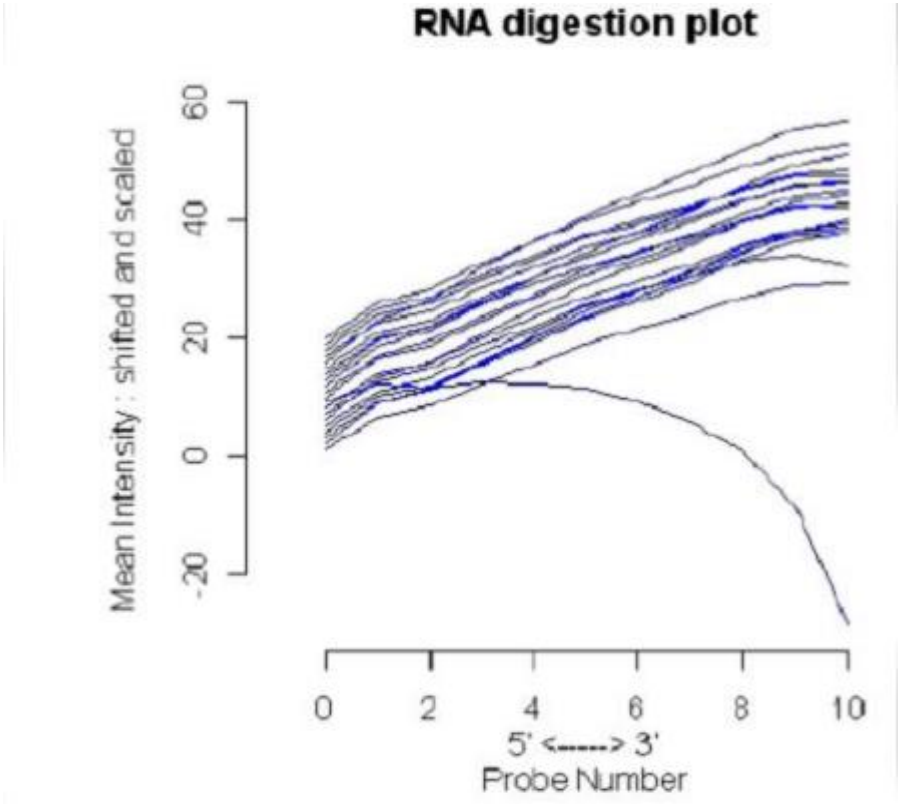
It is hard to decide if things "seem to be all right" just by looking at the numbers.

- Standard statistical approach **use plots** to check the quality
  - ✓ show all data together
  - ✓ highlight structures
  - ✓ may help to detect problems ("unusual patterns")

2. Microarray data analysis with R. QUALITY CONTROL OF THE DATA

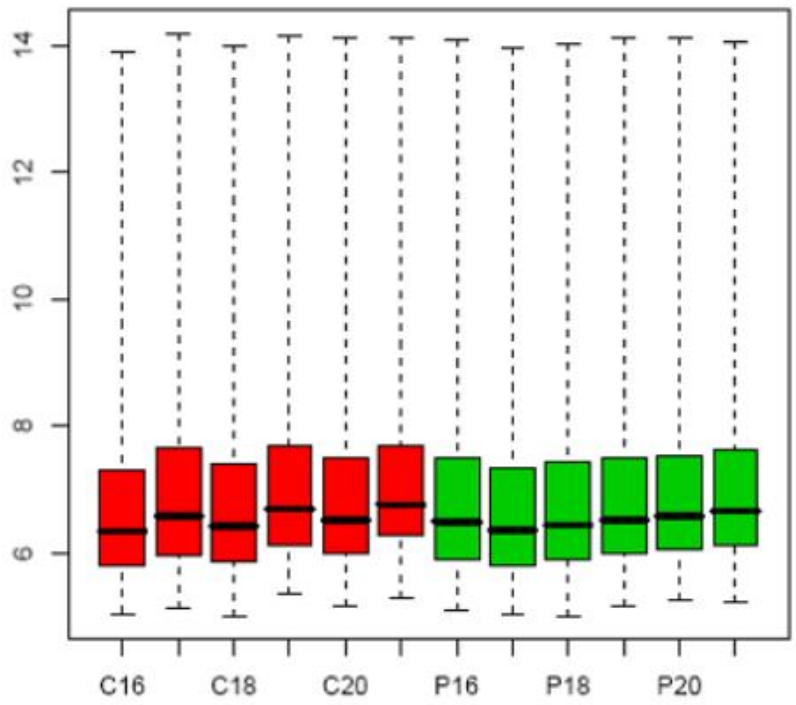
RNA digestion plot. Only for 3'Arrays.

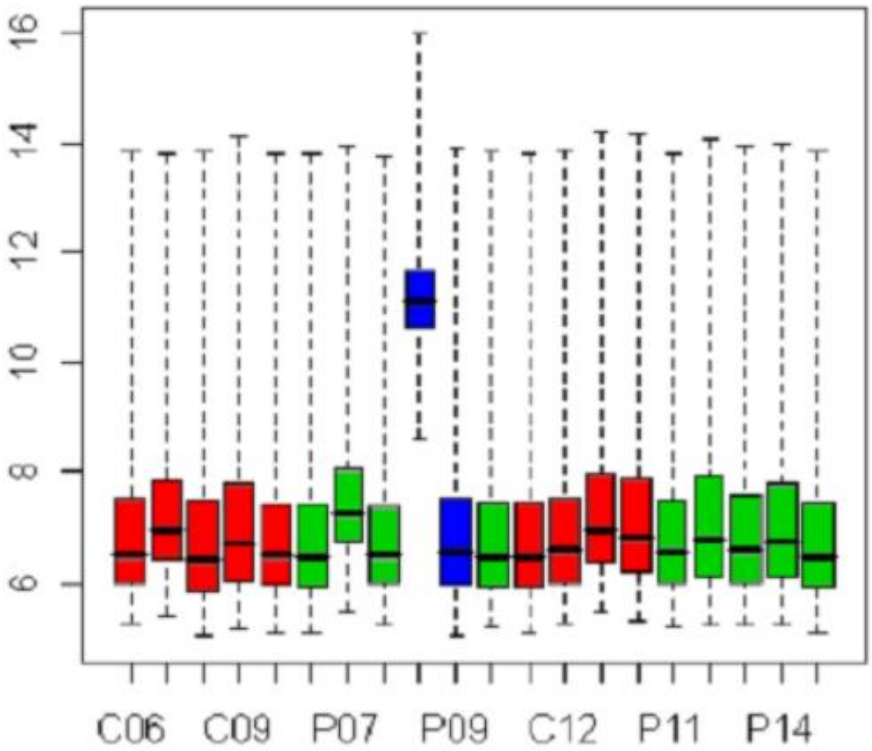




2. Microarray data analysis with R. QUALITY CONTROL OF THE DATA

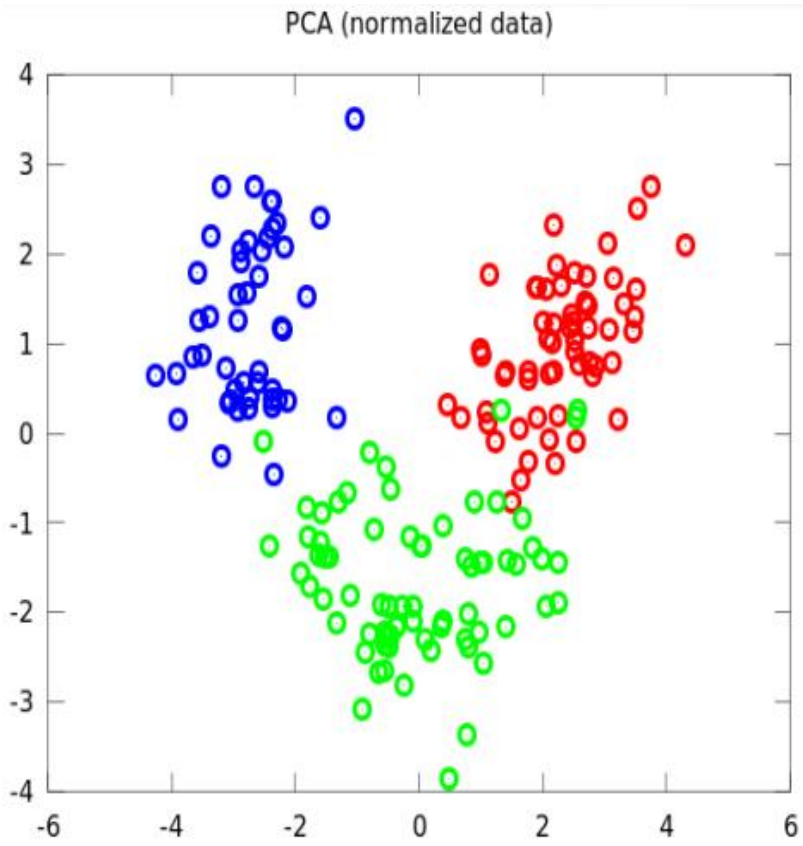
Boxplot intensities. Raw data/Normalized data

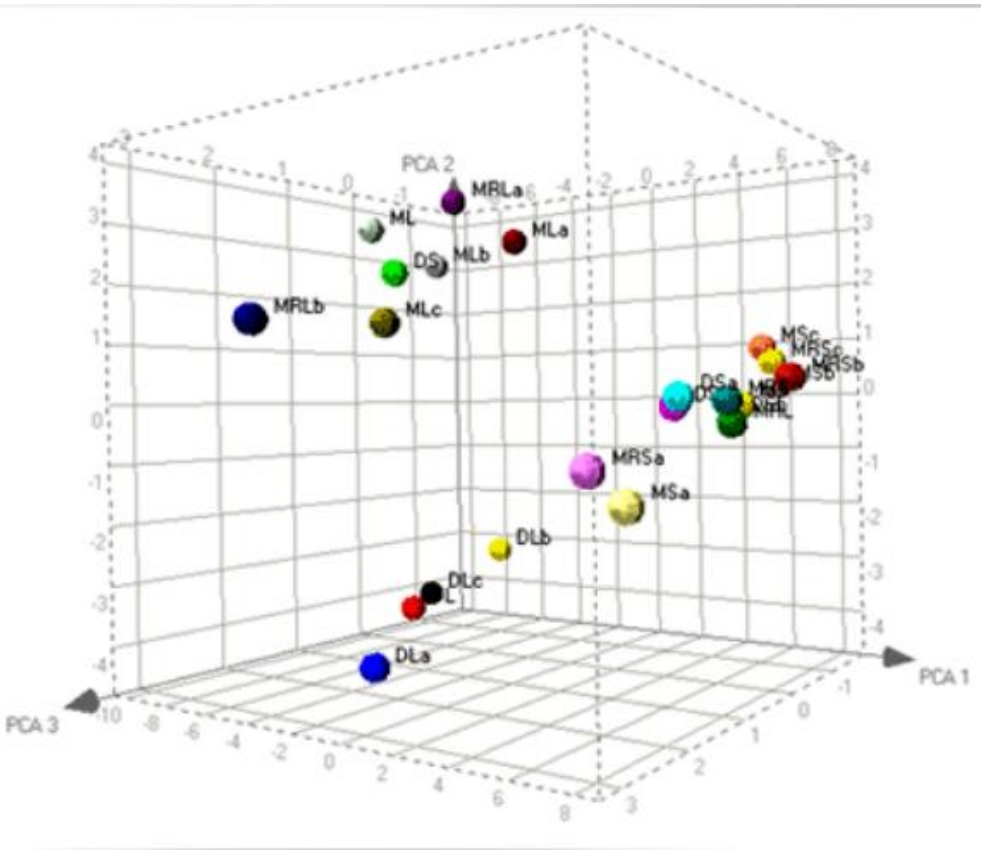


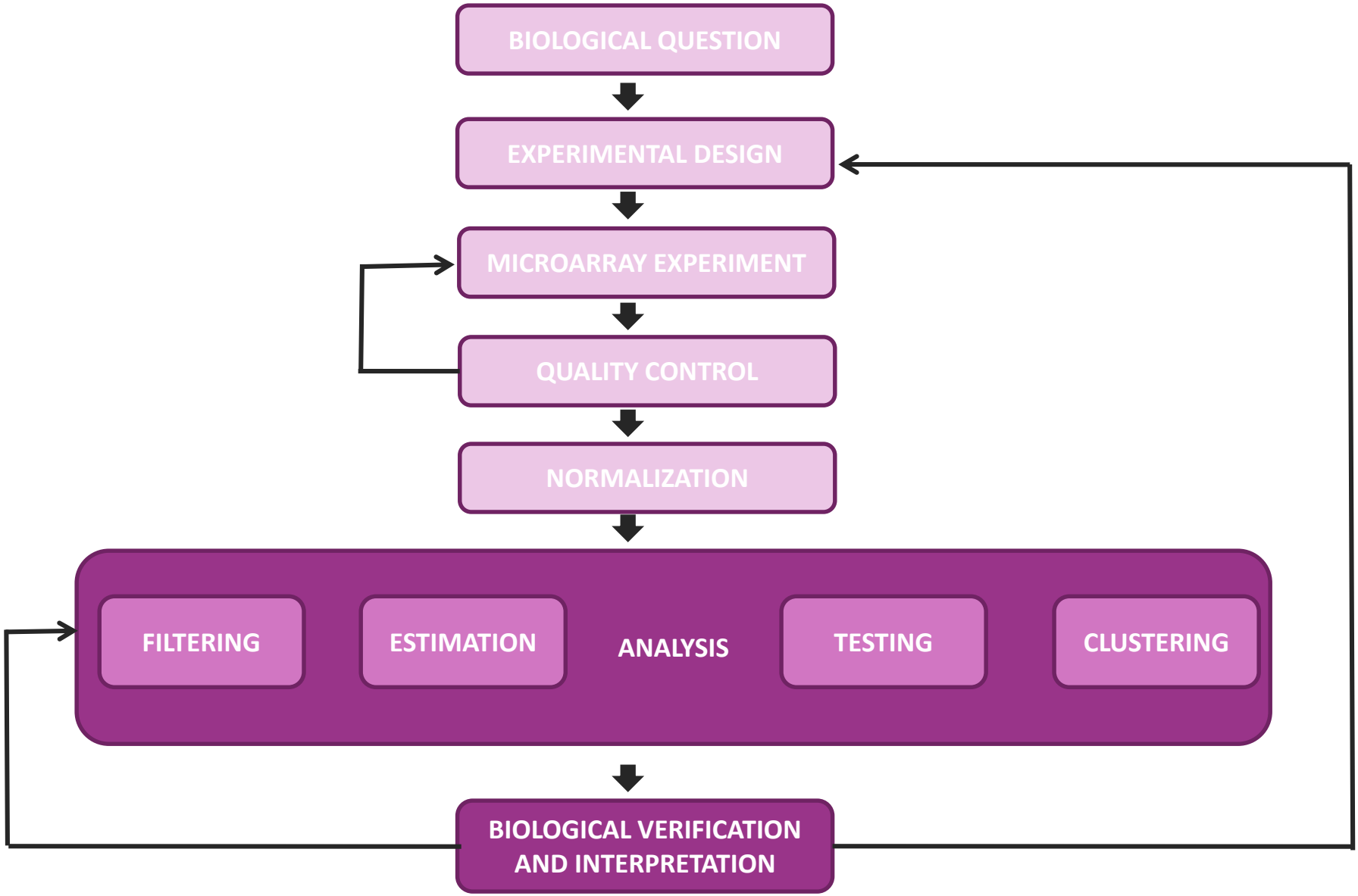


2. Example of a microarray analysis with R. QUALITY CONTROL OF THE DATA

Principal Component Analysis. Raw data/Normalized data






BIOLOGICAL QUESTION
EXPERIMENTAL DESIGN
MICROARRAY EXPERIMENT
QUALITY CONTROL
NORMALIZATION
FILTERING
ESTIMATION
ANALYSIS
TESTING
CLUSTERING
BIOLOGICAL VERIFICATION AND INTERPRETATION

2. Microarray data analysis with R. DATA NORMALIZATION

It is very important (essential) to normalize your data.

Why normalization?

1. To remove systematic biases:
   - Sample preparation
   - Variability in hybridization
   - Scanner settings
   - Experimenter bias
2. To achieve a measured scale such that:
   - Has the same origin for all spots
   - Linear relationship with mRNA quantity

2. Microarray data analysis with R. DATA NORMALIZATION

There exist different methods:

- **RMA (Robust Multiarray Average):** Performs background correction, normalization, and summarization in a modular way. RMA does not take in account unspecific probe hybridization in probe set background calculation (Irizarry et al., 2003)

- **GCRMA:** is a version of RMA with a background correction component that makes use of a probe sequence information (Wu et al., 2004)

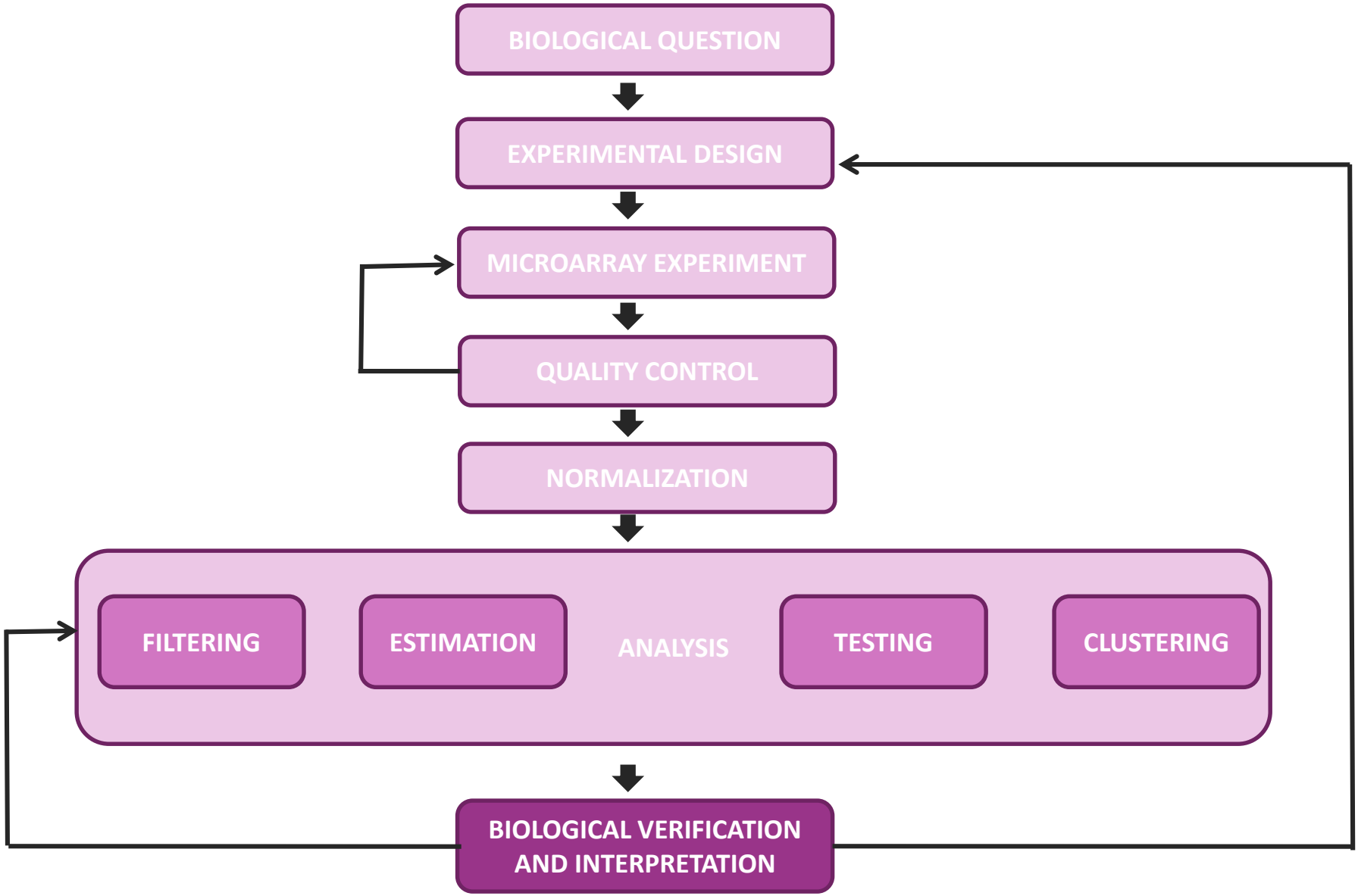
- **PLIER (Probe logarithmic error intensity estimate)**: this method produces an improved signal by accounting for experimentally observed patterns in probe behavior and handling error at the appropriately low and high signal values

2. Example of a microarray analysis with R. DATA NORMALIZATION

Nevertheless the steps they perform are common.

General steps:

1. **Background** correction: correction of the scale origin for spots
2. **Normalization**: standardizing the scale unit. I**ntensity calculation**
3. **Summary** of information of several spots into a single measure for each gene

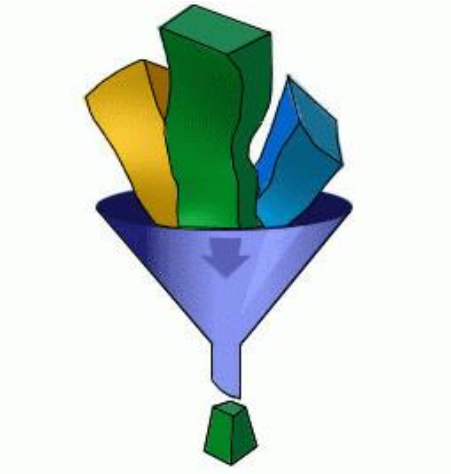

BIOLOGICAL QUESTION
EXPERIMENTAL DESIGN
MICROARRAY EXPERIMENT
QUALITY CONTROL
NORMALIZATION
FILTERING
ESTIMATION
ANALYSIS
TESTING
CLUSTERING
BIOLOGICAL VERIFICATION AND INTERPRETATION

2. Microarray data analysis with R. DATA FILTERING

- In a microarray experiment only a few hundreds/thousand of genes change their expression due to the different conditions

- Researcher is interested in keeping the number of tests/genes as low as possible while keeping the interesting genes in the selected subset.

Genes that do not change introduce noise, therefore is better not to be present when the statistical analysis is done



2. Microarray data analysis with R. DATA FILTERING

There exist different types of filtering:

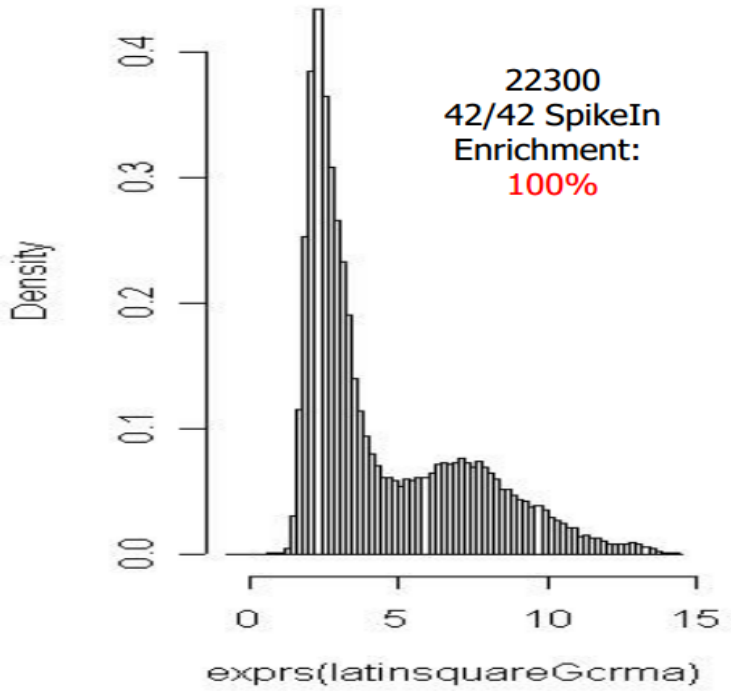
- Annotation features (specific):
  - Specific gene features (i.e. GO term, presence of transcriptional regulative elements in promoters, etc.)

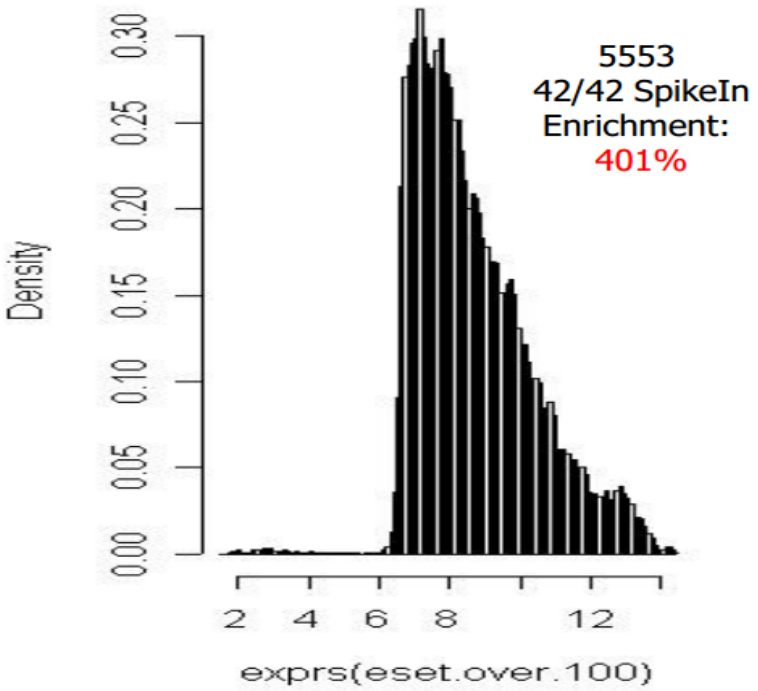
- Signal features (non specific)
  - % intensities greater of a user defined value
  - Interquartile range (IQR) greater of a defined value

- Variance between samples (non specific)
  - Genes that do not change among conditions, could be excluded

2. Microarray data analysis with R. DATA FILTERING

Signal filtering: This technique has as its premise the removal of genes that are deemed to be not expressed or unchanged according to some specific criterion that is under the control of the user.





2. Microarray data analysis with R. COMPARISONS

Statistical inference of differential expression

Class comparison problem:

- Identify genes whose expression is significantly associated with different conditions:
  - ✓Treatment, cell type …
  - ✓Dose, time,….
- Estimate effects/differences between groups.

2. Microarray data analysis with R. COMPARISONS

Which situations does one usually see (here the easiest)?

- Indirect comparisons: 2 groups, unpaired
  - E.g. 10 individuals: 5 suffer diabetes, 5 healthy
  - One sample from each individual
  - Test: Two sample t-test

- Direct comparisons: 2 groups, paired
  - E.g. 10 individuals with brain stroke
  - Two samples from each patient: one from healthy region1 and one from affected region
  - Test: Paired t-test

2. Microarray data analysis with R. COMPARISONS

Some issues in gene selection

- Some related with small sample sizes
  - Variance instability (very low variances produces a high t statistic value)
  - Non-normality of the data
- Related to the big number of variables (test to perform)
  - Multiple testing problem

Standard t test is not strictly correct to be used here, better to use a "modified version": *moderated t test*

2. Microarray data analysis with R. COMPARISONS

- Multiple testing problem: It is needed to control for the type I error (false positives). FALSE DISCOVERY RATE

- Finally we will be assigning a p-value for each test/gene. If the p-value is lower than an established threshold….

Vall d'Hebron Institut de Recerca

2. Microarray data analysis with R. Results presentation. TOP TABLES

Top 50 Candidate Genes for Differential Expression for (Bob+GFP+)-(Bob-GFP+)

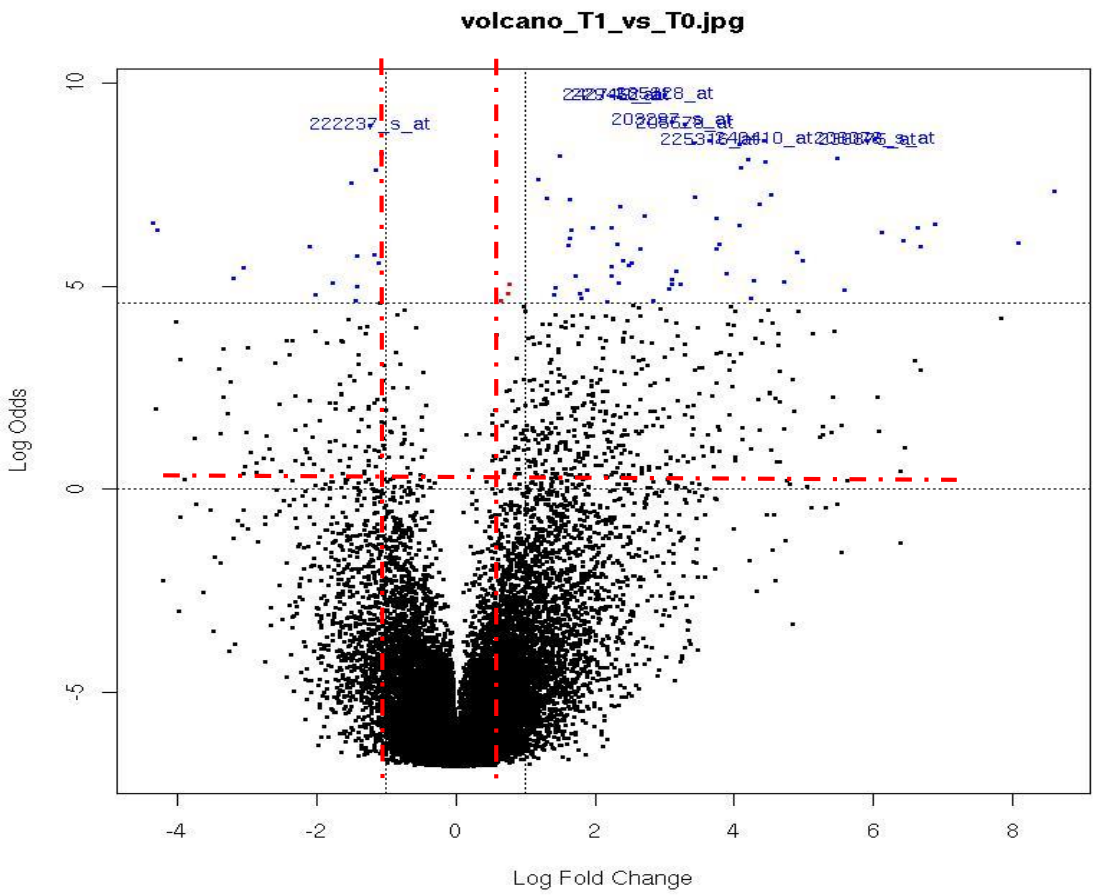
File Edit

ID (1)	Name (1)	M (2)	A (3)	t (4)	P.Value (5)	B (6)
NIA15k	H34599	0.4036	8.677	13.05	0.001097	7.996
NIA15k	H31324	-0.5197	7.795	-12.30	0.001920	7.5
NIA15k	H33309	0.4203	8.864	12.09	0.002262	7.353
NIA15k	H3440	0.5678	9.357	11.66	0.003164	7.049
NIA15k	H36795	0.46	11.46	11.61	0.003308	7.008
NIA15k	H3121	0.4409	8.071	11.36	0.004038	6.826
NIA15k	H36999	0.3807	5.925	11.28	0.004335	6.761
NIA15k	H3132	0.37	9.227	11.27	0.004357	6.756
NIA15k	H32838	1.640	7.065	11.21	0.004566	6.713
NIA15k	H36207	-0.3931	10.03	-11.14	0.004855	6.656
NIA15k	H37168	0.3909	8.516	10.84	0.006252	6.422
NIA15k	H31831	-0.3738	9.411	-10.71	0.007008	6.316
NIA15k	H32014	0.3630	6.999	10.57	0.007858	6.209
NIA15k	H34471	-0.3533	6.634	-10.50	0.008414	6.145
NIA15k	H37558	0.5319	7.714	10.49	0.008438	6.142
NIA15k	H3126	0.385	7.52	10.47	0.008632	6.12
NIA15k	H34360	-0.3409	7.847	-10.31	0.00993	5.989
NIA15k	H36794	0.4717	8.02	10.15	0.01149	5.851
NIA15k	H3329	0.4125	8.995	10.01	0.01301	5.733
NIA15k	H35017	0.4338	7.456	9.936	0.01392	5.67
NIA15k	H32367	0.4093	7.725	9.765	0.01629	5.52
NIA15k	H32678	0.4608	8.317	9.764	0.01631	5.518
NIA15k	H31232	-0.3717	7.509	-9.759	0.01639	5.514
NIA15k	H3111	0.3694	10.25	9.746	0.01658	5.502
NIA15k	H34258	0.2992	7.264	9.723	0.01695	5.482
NIA15k	H32159	0.4184	8.463	9.703	0.01726	5.464
NIA15k	H33192	-0.4095	6.425	-9.59	0.01919	5.363
NIA15k	H35961	-0.3624	7.346	-9.509	0.02073	5.289
NIA15k	H36025	0.4266	6.674	9.504	0.02082	5.284

1. Gene identifiers
2. Log2 Fold Change
3. Average intensity
4. t statistics
5. p-values
6. Log-odd statistics

2. Microarray data analysis with R. Results presentation. VOLCANO PLOTS

Statistics and biological significance representation



4. Example of a microarray analysis with R. Results presentation. CLUSTERING

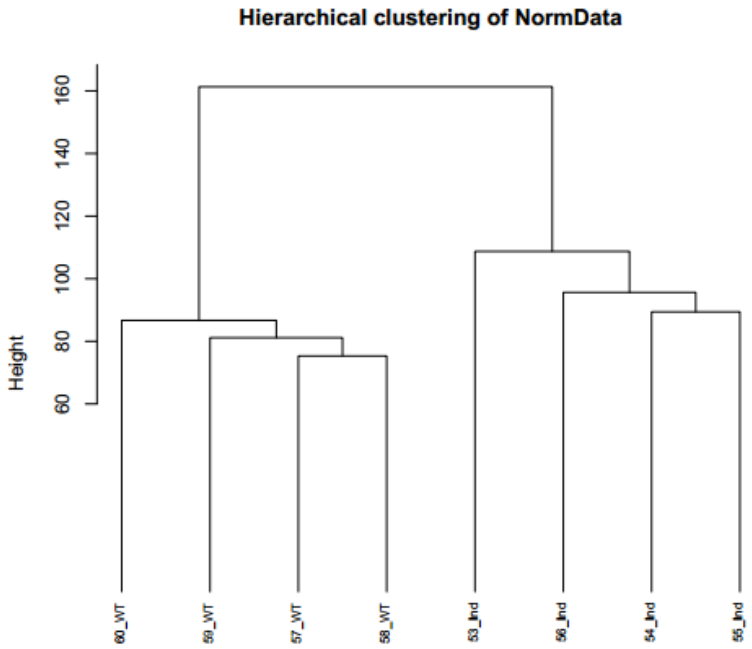
Types:

- **Supervised clustering** try to find the best partition for data that belong to a know set o classes
- **Unsupervised clustering** try to define the number and the size of the classes in which the transcription profiles can be fitted in.
- **Distances** between genes/samples are used to classify them (Euclidian distance, Manhattan distance, Mahalanovis distance….)

4. Example of a microarray analysis with R. Results presentation. CLUSTERING

Hierarchical Clustering (HCL)

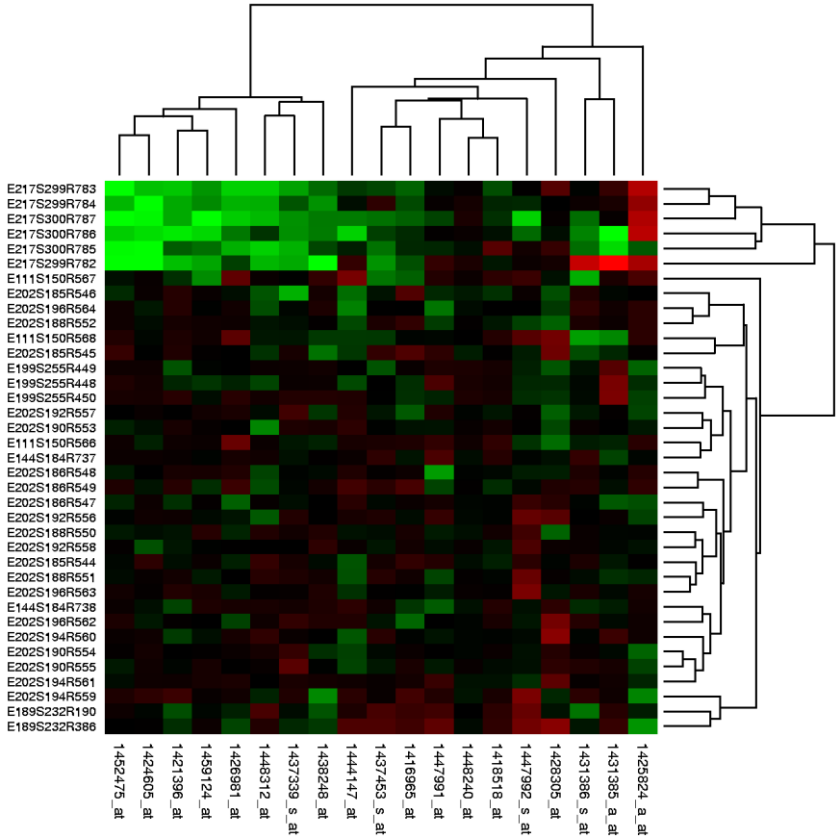
- HCL is an agglomerative /dividing clustering method.
- The iterative process continues until all groups are connected in a hierarchical tree.
- Samples more similar between them are closed.



2. Microarray data analysis with R. Results presentation. CLUSTERING

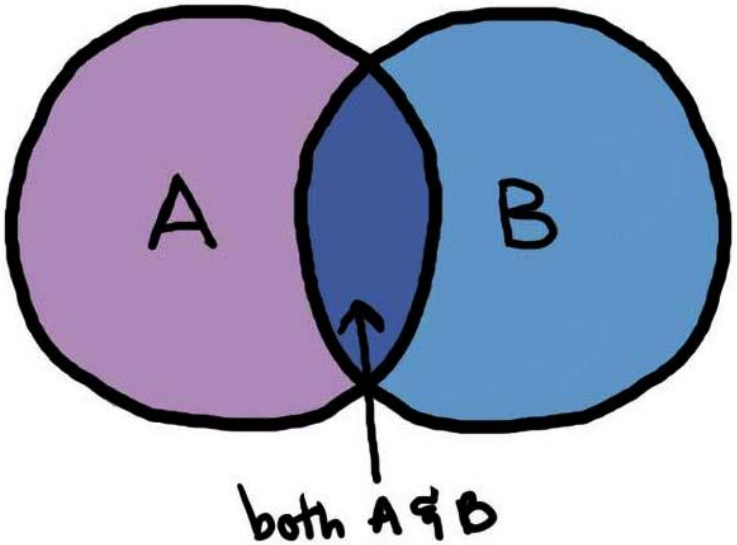
Heatmaps

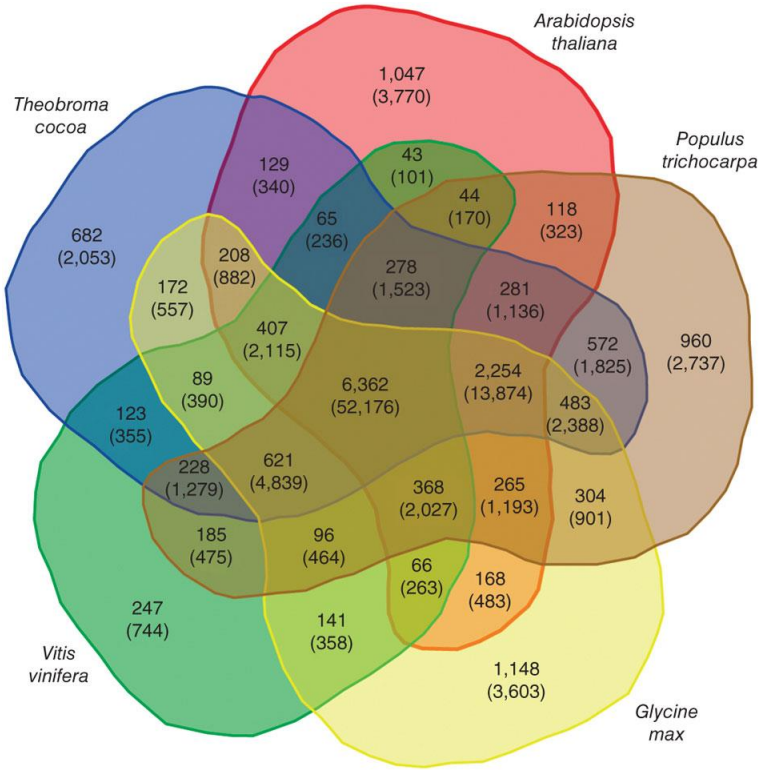
- Allow a quick visualization of the possible expression patterns that could exists among samples.



2. Microarray data analysis with R. Results presentation. VENN DIAGRAMS

- If the study have more than one comparisons it could be interesting to look for common genes in the gene lists (multiple comparisons)





2. Microarray data analysis with R. Results presentation. ANNOTATION

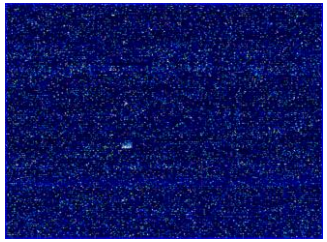
- Relation between probes sets and genes.
- An important issue in microarray data analysis is the specific association of probe identifiers with genome annotated transcripts.
- Not of the probes have a “genome annotated transcript”.
- Different database used (Entrez, Gene Symbol, Ensembl,…) generates different results.

Table of contents

3. Introduction to Biological Significance

You obtained a list of features!!!! … what is next?

- Select some genes for validation?
- Follow up experiments on some genes?
- Publish a huge table with all results?
- Try to learn on all genes in the list?



GNAQ
GNAS
DGKZ
GUCY1A3
PDE4B
PDE4D
ATP2A2
ATP2A3
NOS1
CNN1
GSTO1
NOS3
CNN2
MYLK2
CALD1
ACTA1
MYL2

my favourite gene

3. Introduction to Biological Significance

From gene lists to Pathway Analysis

- Gene lists contain useful information
  - This can be extracted from databases
  - Generically described as **Gene Annotation**

- Besides, we may obtain information from the analysis of *gene sets*
  - Genes don't act individually, rather in group (more realistic approach)
  - There are less gene sets tan individual genes (relatively simpler to manage)
  - Generically described as **Pathway analysis**

3. Introduction to Biological Significance

What do we need?

1. A way to identify genes relevant to the condition under study

2. A shared functional vocabulary

3. Systematic link between genes and functions

4. Statistical analysis

3. Introduction to Biological Significance

What do we need?

1. A way to identify genes relevant to the condition under study → Fold change, ranking, ANOVA
2. A shared functional vocabulary → Gene Ontology Annotation
3. Systematic link between genes and functions → Gene Ontology Annotation
4. Statistical analysis → Enrichment analysis, GSEA

3. Introduction to Biological Significance

Gene list and Annotations

- Identifiers (Ids) are ideally unique, stable names or numbers that help track database records (e.g. social insurance number, Entrez gene ID)

- But, information of features is stored in many databases
  - Genes have many Ids

- Records for: Gene, DNA, RNA, Protein

3. Introduction to Biological Significance

Common identifiers

Gene
Ensembl ENSG00000139618
Entrez Gene 675
Unigene Hs.34012

RNA transcript
GenBank BC026160.1
RefSeq NM_000059
Ensembl ENST00000380152

Protein
Ensembl ENSP00000369497
RefSeq NP_000050.2
UniProt BRCA2_HUMAN or A1YBP1_HUMAN
IPI IPI00412408.1
EMBL AF309413
PDB 1MIU

Species-specific
HUGO HGNC BRCA2
MGI MGI:109337
RGD 2219
ZFIN ZDB-GENE-060510-3
FlyBase CG9097
WormBase WBGene00002299 or ZK1067.1
SGD S000002187 or YDL029W

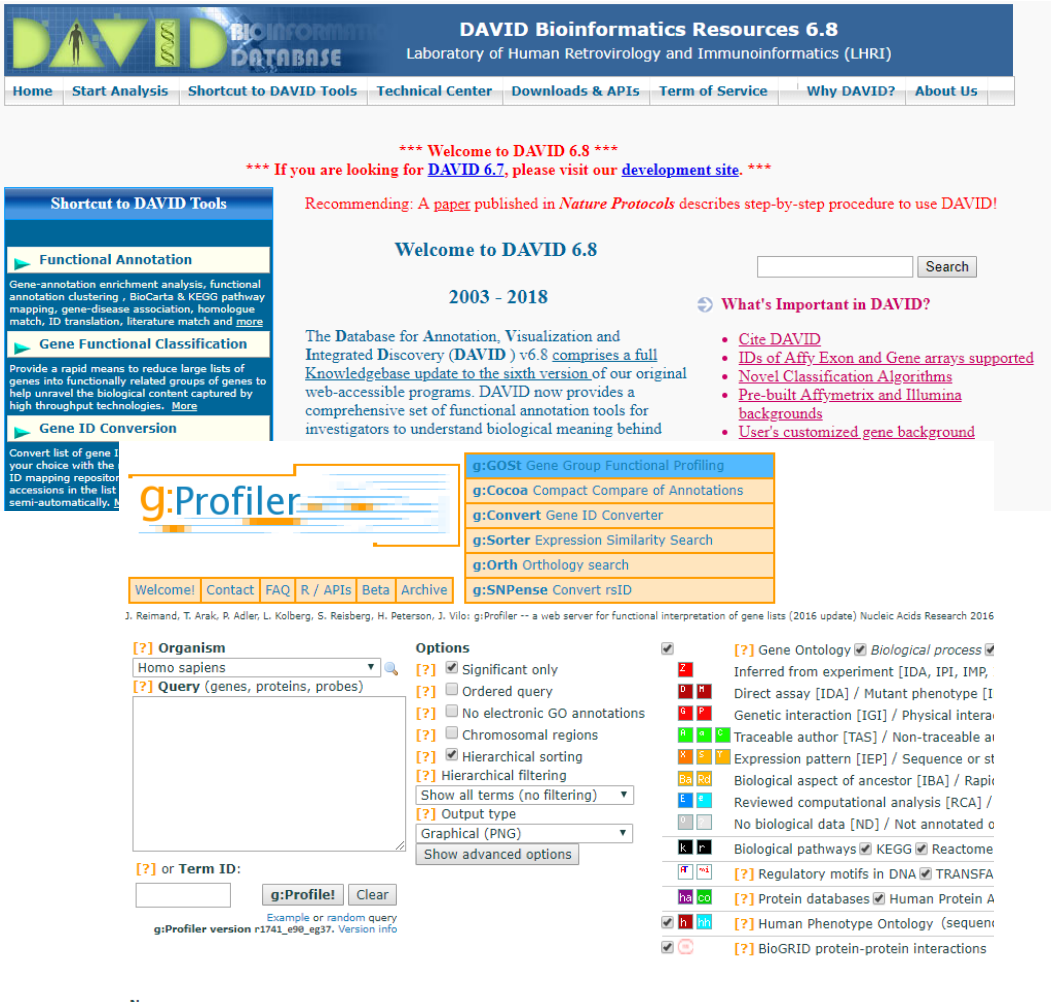
Annotations
InterPro IPR015252
OMIM 600185
Pfam PF09104
Gene Ontology GO:0000724
SNPs rs28897757

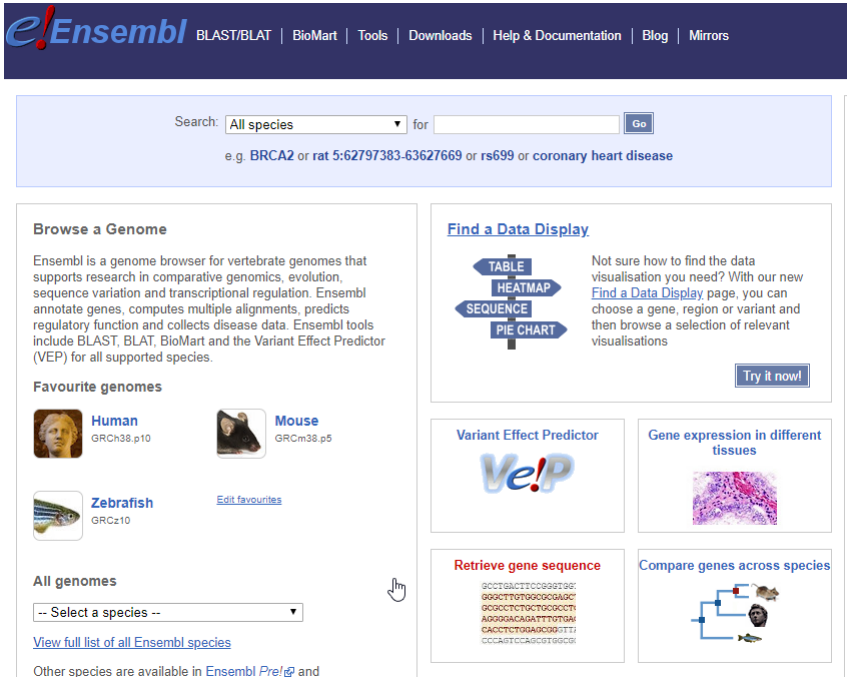
Experimental Platform
Affymetrix 208368_3p_s_at
Agilent A_23_P99452
CodeLink GE60169
Illumina GI_4502450-S

In Red = recommended

3. Introduction to Biological Significance

Use ID converters to prepare lists:





3. Introduction to Biological Significance

Gene Ontology Consortium

GENE ONTOLOGY (http://www.geneontology.org/)

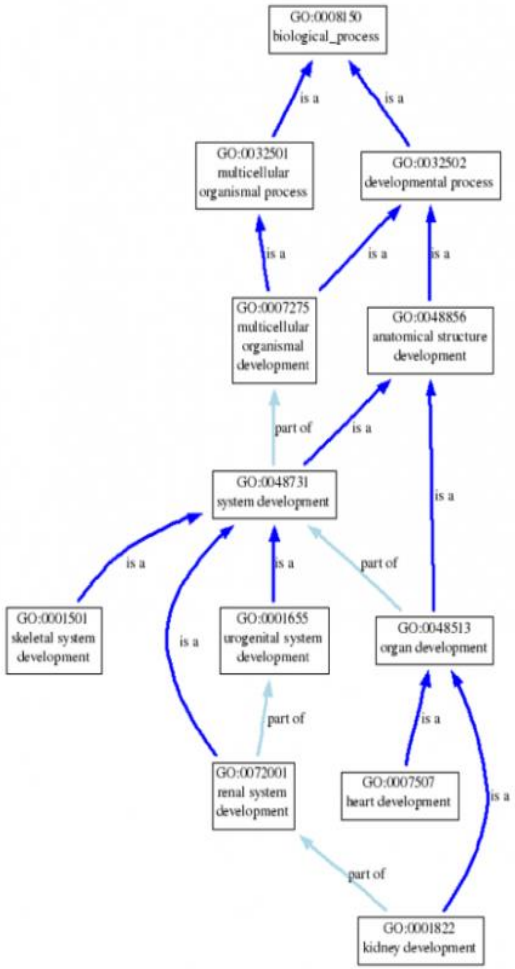
- A major bioinformatics initiative with the aim of standardizing the representation of gene and gene product attributes across species and databases.

- The Gene Ontology (GO) is a controlled vocabulary, a set of standard terms (words and phrases) used for indexing and retrieving information.

3. Introduction to Biological Significance

GENE ONTOLOGY STRUCTURE

- GO defines the relationships between the terms, making it a structured vocabulary.
- Terms are related within a hierarchy (“is a”, “is part of”)
- Terms can have more than one parent or child



3. Introduction to Biological Significance

GENE ONTOLOGY DOMAINS

1. MOLECULAR FUNCTION (MF): basic activity or task

   e.g. catalytic activity, calciom ion binding

2. BIOLOGICAL PROCESS (BP): broad objective or goal

   e.g. signal transduction, inmuno response

3. CELLULAR COMPONENT (CC): Location or complex

   e.g. nucleus, mitochondrion

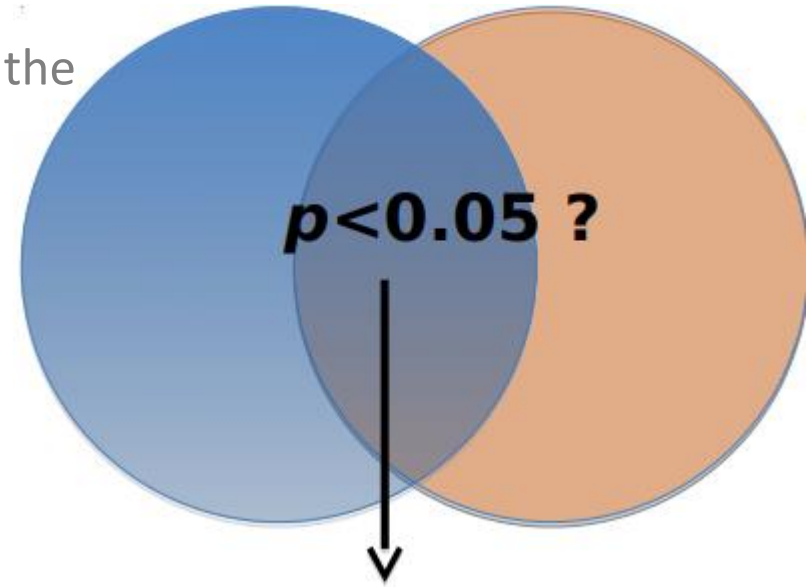
Cytochrome c: MF = oxidoreductase activity
BP = oxidative phosphorilation and induction of cell death
CC = mitochondrial matrix and mitochondrial inner membrane

3. Introduction to Biological Significance

Enrichment Analysis

Hypothesis: Drug sensitivity in brain cancer is related to reduced neurotransmitter signaling

Gene list from the experiment



All the genes known to be involved in Neurotransmitter signaling

Statistical test : are there more annotations in gene list than expected?

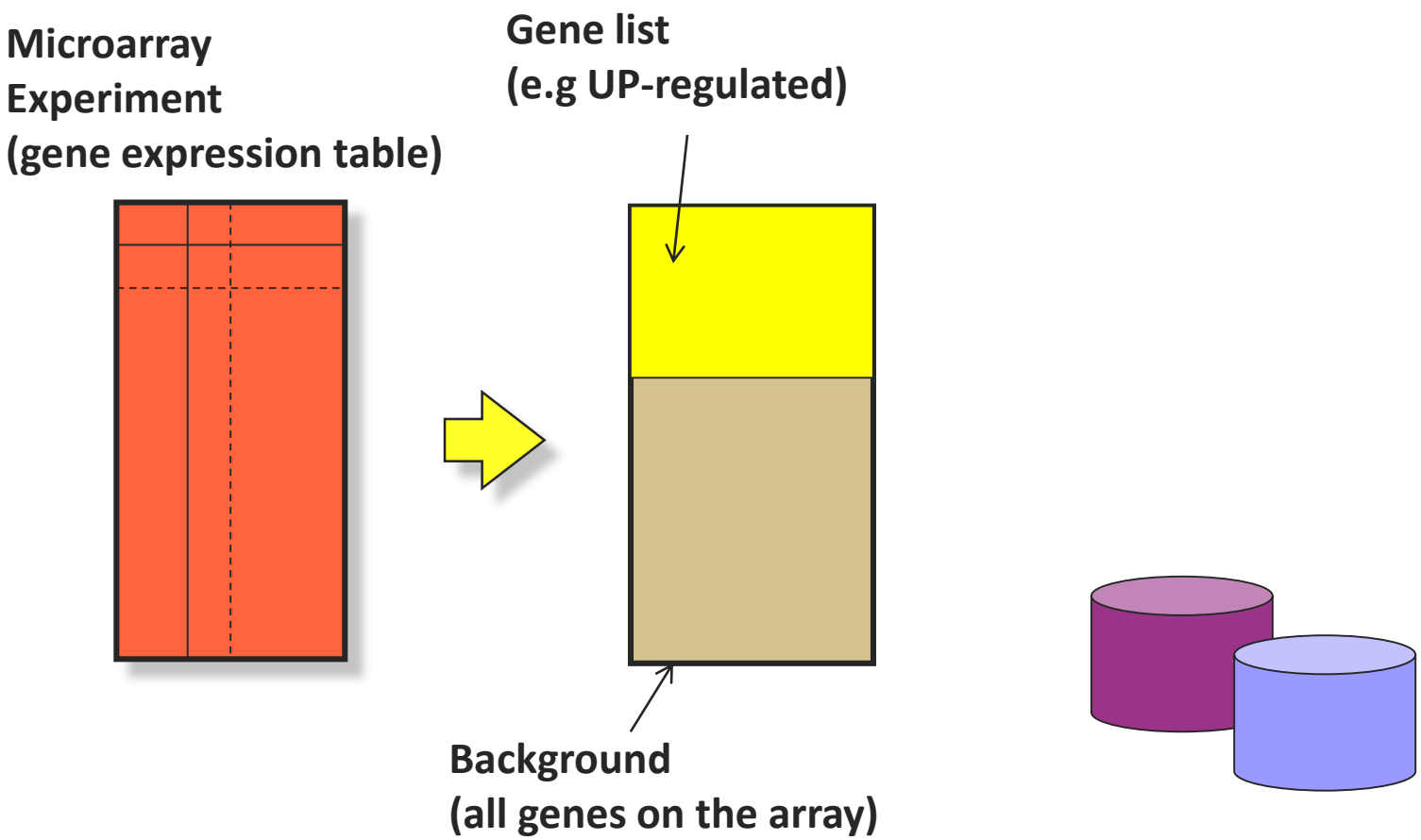
3. Introduction to Biological Significance

1. Gene list (e.g. expression change > 2-fold)
   - Answer the question: **Are any gene set surprisingly enriched or (depleted) in my gene list?**
   - Statistical test: **Fisher's Exact test (hypergeometric test)**

2. Ranked list (e.g. by differential expression)
   - Answer the question: **Are any gene set ranked surprisingly high or low in my ranked list of genes?**
   - Statistical test: mininum hypergeometric test

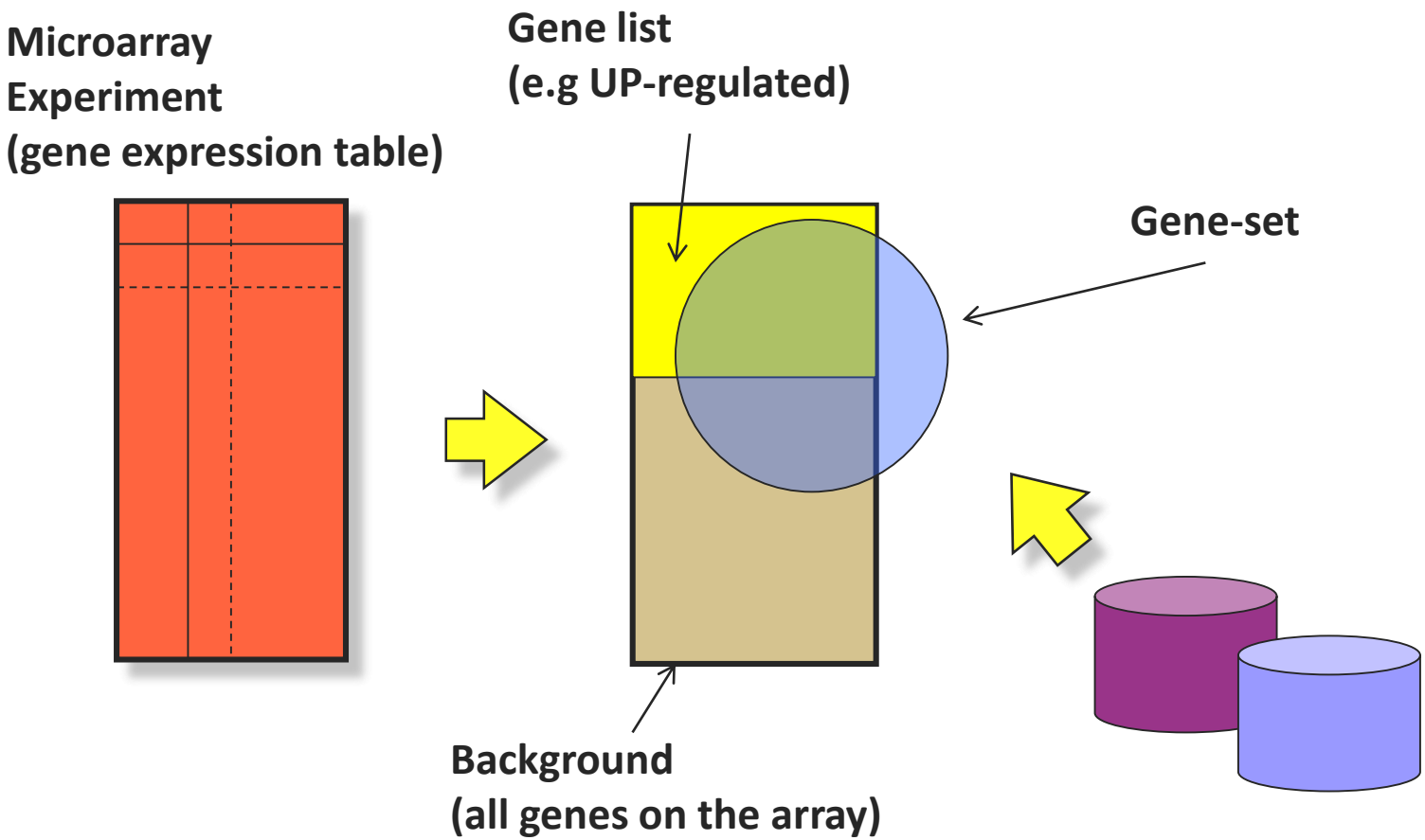
3. Introduction to Biological Significance

Hypergeometric test



3. Introduction to Biological Significance

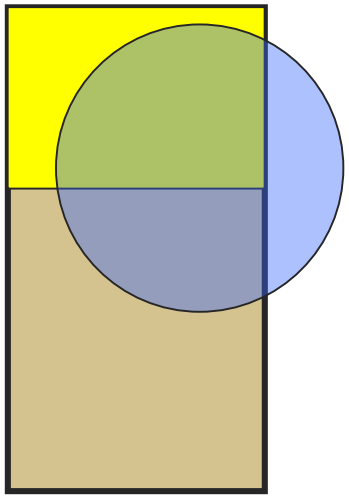
Hypergeometric test



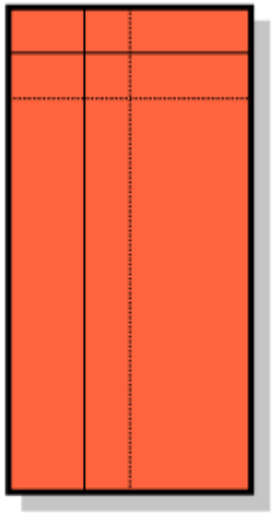
3. Introduction to Biological Significance

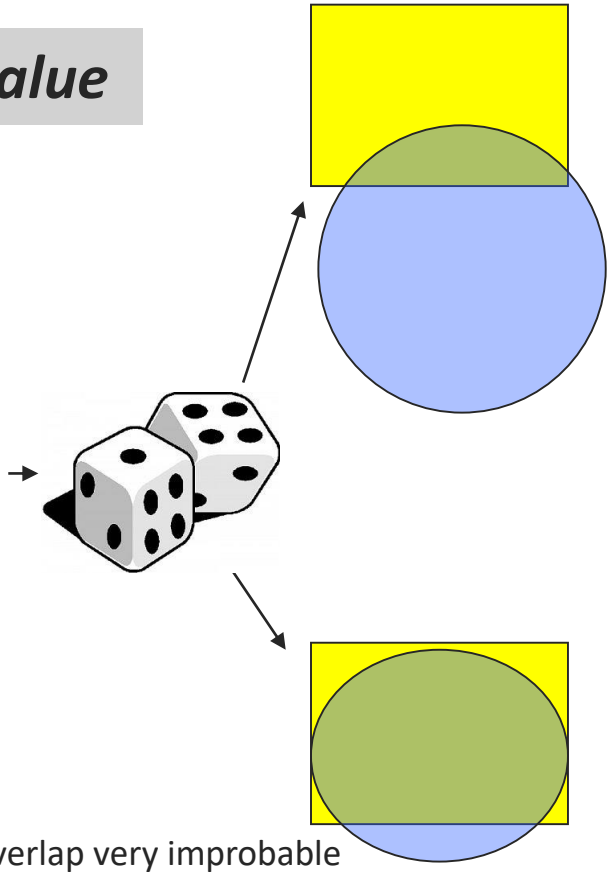
Hypergeometric test

The output of an enrichment test is a *P-value*



The P-value assesses the probability that the overlap is at least as large as observed by **random sampling** the array genes.





3. Introduction to Biological Significance

Recipe for gene list enrichment test:

1. Define gene list (e.g. FC >3) and background list (e.g. all genes in the array)
2. Select gene sets to test for enrichment
3. Run enrichment test and adjust for multiple testing if necessary
4. Interpret your enrichments

Table of contents

2. Example of a microarray analysis with R. LOAD THE DATA

GENE EXPRESSION ONMIBUS DATABASE

- Public functional genomic repository from NCBI
- Array and sequence-based data are accepted
- It is mandatory to upload your microarrays CEL files before publishing any article about them

http://www.ncbi.nlm.nih.gov/geo/

4. Example of a microarray analysis with R. LOAD THE DATA

GENE EXPRESSION ONMIBUS DATABASE

The data for the example: GDS4155

Keyword or GEO Accession | Search

Search for GDS4155[ACCN] | Search | Clear | Show All | Advanced Search

DataSet Record GDS4155: Expression Profiles | Data Analysis Tools | Sample Subsets

Title:	Dopaminergic transcription factors Ascl1, Lmx1a, Nurr1 combined effect on embryonic fibroblasts		
Summary:	Analysis of induced dopaminergic (iDA) neurons generated from E14.5 mouse embryonic fibroblasts (MEFs) reprogrammed by infection with lentiviruses expressing dopaminergic transcription factors Ascl1, Lmx1a and Nurr1. Results provide insight into the molecular basis of MEF to iDA reprogramming.		
Organism:	*Mus musculus*		
Platform:	GPL6246: [MoGene-1_0-st] Affymetrix Mouse Gene 1.0 ST Array [transcript (gene) version]		
Citation:	Caiazzo M, Dell'Anno MT, Dvoretskova E, Lazarevic D et al. Direct generation of functional dopaminergic neurons from mouse and human fibroblasts. *Nature* 2011 Jul 3;476(7359):224-7. PMID: 21725324		
Reference Series:	GSE27174	**Sample count:**	8
Value type:	transformed count	**Series published:**	2011/07/04

4. Example of a microarray analysis with R. LOAD THE DATA

Sample	Title
GSM671653	Fibroblasts dopaminergic induced rep1
GSM671654	Fibroblasts dopaminergic induced rep2
GSM671655	Fibroblasts dopaminergic induced rep3
GSM671656	Fibroblasts dopaminergic induced rep4
GSM671657	Fibroblasts not induced rep1
GSM671658	Fibroblasts not induced rep2
GSM671659	Fibroblasts not induced rep3
GSM671660	Fibroblasts not induced rep4

INDUCED

WT

4. Example of a microarray analysis with R. LOAD THE DATA

Two types of files are necessary to begin the data analysis:

1. CEL files
2. Targets file

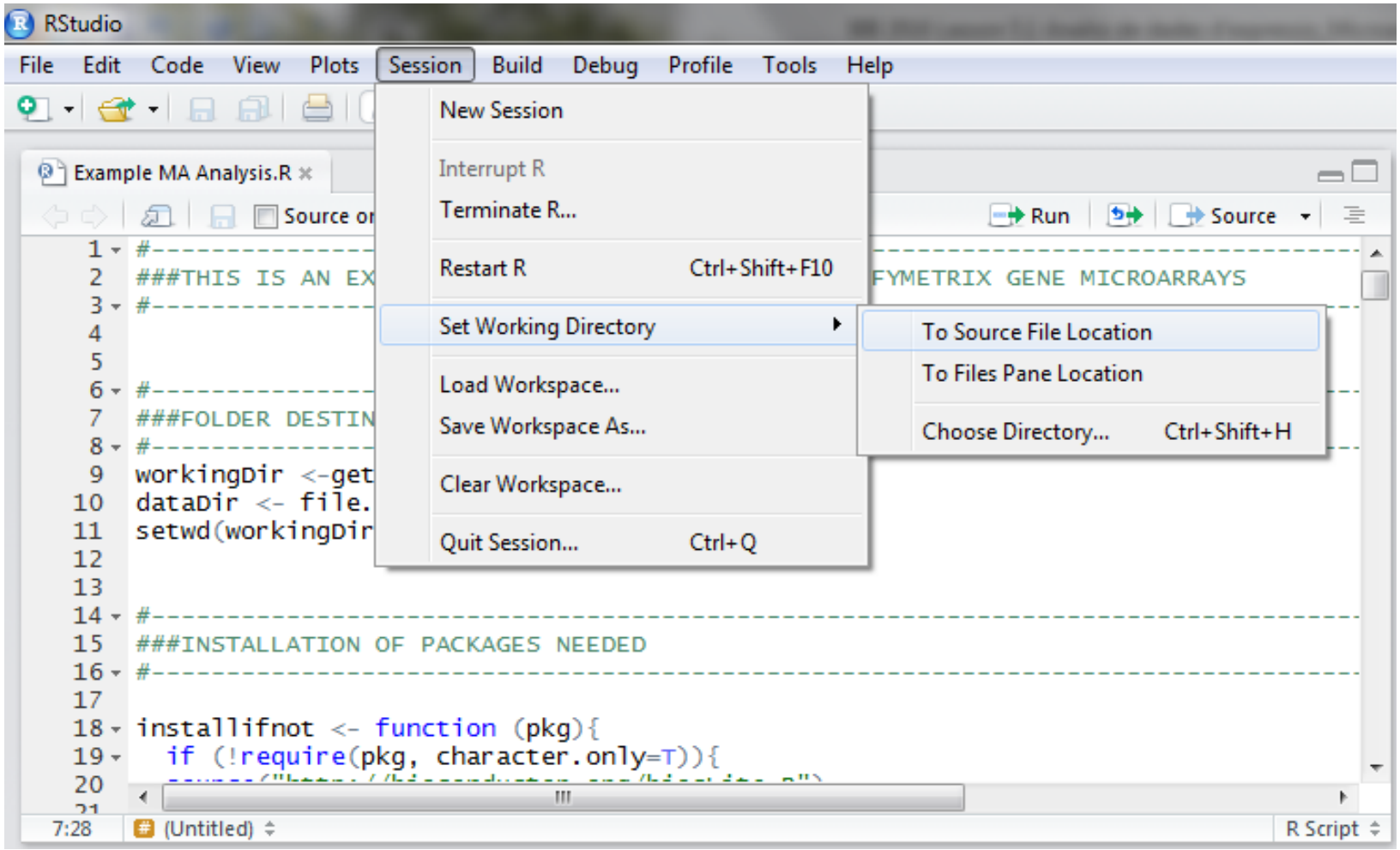
fileName	grupos	ShortName	Colors
GSM671653.CEL	Induced	53_Ind	red
GSM671654.CEL	Induced	54_Ind	red
GSM671655.CEL	Induced	55_Ind	red
GSM671656.CEL	Induced	56_Ind	red
GSM671657.CEL	WT	57_WT	blue
GSM671658.CEL	WT	58_WT	blue
GSM671659.CEL	WT	59_WT	blue
GSM671660.CEL	WT	60_WT	blue

4. Example of a microarray analysis with R. LOAD THE DATA

We have **to define the folders** before begin to analyze

- make a folder name it for exemple “microarrays”
- inside this folder create two more:
  - name the second “dades”
  - name the second “results”
- save the CEL and target files in the “dades” folder
- open “MA-Analysis-Example.R” with RStudio

4. Example of a microarray analysis with R. LOAD THE DATA



4. Example of a microarray analysis with R. LOAD THE DATA

We have to **define the working folders:**

```
workingDir <- getwd()
dataDir <- file.path(workingDir, "data")
dataDir
resultsDir <- file.path(workingDir, "results")
resultsDir
```

Define the function to install the packages:

```
installifnot <- function (pkg){
  if (!require(pkg, character.only=T)){
  source("http://bioconductor.org/biocLite.R")
  biocLite(pkg)
}else{
  require(pkg, character.only=T)
  }
}
```

4. Example of a microarray analysis with R. LOAD THE DATA

Install the necessary packages:

```
installifnot("pd.mogene.1.0.st.v1")
installifnot("mogene10sttranscriptcluster.db")
installifnot("oligo")
installifnot("limma")
installifnot("Biobase")
installifnot("arrayQualityMetrics")
installifnot("genefilter")
installifnot("multtest")
installifnot("annotate")
installifnot("xtable")
installifnot("gplots")
installifnot("scatterplot3d")
```

To be sure, all the packages are installed and loaded, execute last code twice.

4. Example of a microarray analysis with R. LOAD THE DATA

We load the data:

```
#-----------------------------------------------------------------------------
###LOAD DATA: TARGETS AND CEL FILES.
#-----------------------------------------------------------------------------

#TARGETS
targets <-read.csv(file=file.path(dataDir,"targets.csv"), header = TRUE, sep=";")
targets

#CELFILES
CELfiles <- list.celfiles(file.path(dataDir))
CELfiles
rawData <- read.celfiles(file.path(dataDir,CELfiles))

#DEFINE SOME VARIABLES FOR PLOTS
sampleNames <- as.character(targets$ShortName)
sampleColor <- as.character(targets$Colors)
```

4. Example of a microarray analysis with R. QUALITY CONTROL OF THE DATA

Let's do with our data:

```
#-----------------------------------------------------------------------------
###QUALITY CONTROL OF ARRAYS: RAW DATA
#-----------------------------------------------------------------------------


#BOXPLOT
boxplot(rawData, which="all",las=2, main="Intensity distribution of RAW data",
        cex.axis=0.6, col=sampleColor, names=sampleNames)

#HIERARQUICAL CLUSTERING
clust.euclid.average <- hclust(dist(t(exprs(rawData))),method="average")
plot(clust.euclid.average, labels=sampleNames, main="Hierarchical clustering
of RawData",
     cex=0.7,  hang=-1)
```

4. Example of a microarray analysis with R. QUALITY CONTROL OF THE DATA

Let's do with our data:

```
#PRINCIPAL COMPONENT ANALYSIS

plotPCA <- function ( X, labels=NULL, colors=NULL, dataDesc="", scale=FALSE,
formapunts=NULL, myCex=0.8,...)
{
  pcX<-prcomp(t(X), scale=scale) # o prcomp(t(X))
  loads<- round(pcX$sdev^2/sum(pcX$sdev^2)*100,1)
  xlab<-c(paste("PC1",loads[1],"%"))
  ylab<-c(paste("PC2",loads[2],"%"))
  if (is.null(colors)) colors=1
  plot(pcX$x[,1:2],xlab=xlab,ylab=ylab, col=colors, pch=formapunts,
       xlim=c(min(pcX$x[,1])-100000,
max(pcX$x[,1])+100000),ylim=c(min(pcX$x[,2])-100000, max(pcX$x[,2])+100000))
  text(pcX$x[,1],pcX$x[,2], labels, pos=3, cex=myCex)
  title(paste("Plot of first 2 PCs for expressions in", dataDesc, sep=" "),
cex=0.8)
}
```

4. Example of a microarray analysis with R. QUALITY CONTROL OF THE DATA

Let's do with our data:

```
#PRINCIPAL COMPONENT ANALYSIS

plotPCA(exprs(rawData), labels=sampleNames, dataDesc="raw data",
colors=sampleColor,
        formapunts=c(rep(16,4),rep(17,4)), myCex=0.6)

#-----------------------------------------------------------------------
#SAVE TO A FILE
pdf(file.path(resultsDir, "QCPlots_Raw.pdf"))
boxplot(rawData, which="all",las=2, main="Intensity distribution of RAW data",
        cex.axis=0.6, col=sampleColor, names=sampleNames)
plot(clust.euclid.average, labels=sampleNames, main="Hierarchical clustering
of samples of RawData",
     cex=0.7,  hang=-1)
plotPCA(exprs(rawData), labels=sampleNames, dataDesc="raw data",
colors=sampleColor,
        formapunts=c(rep(16,4),rep(17,4)), myCex=0.6)
dev.off()
```

4. Example of a microarray analysis with R. DATA NORMALIZATION

Let's do with our data:

```
#-----------------------------------------------------------------------------
###DATA NORMALIZATION
#-----------------------------------------------------------------------------
eset<-rma(rawData)

write.exprs(eset, file.path(resultsDir, "NormData.txt"))


## We can see data previous normalization
head(exprs(rawData))
dim(exprs(rawData))

## We can see data after normalization
head(exprs(normData))
dim(exprs(normData))
```

4. Example of a microarray analysis with R. QUALITY CONTROL. NORMALIZED DATA

Let's do with our data:

```
#-----------------------------------------------------------------------
###QUALITY CONTROL OF ARRAYS: NORMALIZED DATA
#-----------------------------------------------------------------------

#BOXPLOT
boxplot(eset, las=2, main="Intensity distribution of Normalized data", cex.axis=0.6,
        col=sampleColor, names=sampleNames)

#HIERARQUICAL CLUSTERING
clust.euclid.average <- hclust(dist(t(exprs(eset))),method="average")
plot(clust.euclid.average, labels=sampleNames, main="Hierarchical clustering of NormData",
     cex=0.7,  hang=-1)
```

4. Example of a microarray analysis with R. QUALITY CONTROL. NORMALIZED DATA

Let's do with our data:

```
#PRINCIPAL COMPONENT ANALYSIS
plotPCA <- function ( X, labels=NULL, colors=NULL, dataDesc="", scale=FALSE, formapunts=NULL, myCex=0.8,...)
{
  pcX<-prcomp(t(X), scale=scale) # o prcomp(t(X))
  loads<- round(pcX$sdev^2/sum(pcX$sdev^2)*100,1)
  xlab<-c(paste("PC1",loads[1],"%"))
  ylab<-c(paste("PC2",loads[2],"%"))
  if (is.null(colors)) colors=1
  plot(pcX$x[,1:2],xlab=xlab,ylab=ylab, col=colors, pch=formapunts,
       xlim=c(min(pcX$x[,1])-10, max(pcX$x[,1])+10),ylim=c(min(pcX$x[,2])-10, max(pcX$x[,2])+10))
  text(pcX$x[,1],pcX$x[,2], labels, pos=3, cex=myCex)
  title(paste("Plot of first 2 PCs for expressions in", dataDesc, sep=" "), cex=0.8)
}
```

```
plotPCA(exprs(eset), labels=sampleNames, dataDesc="NormData", colors=sampleColor,
        formapunts=c(rep(16,4),rep(17,4)), myCex=0.6)
```

4. Example of a microarray analysis with R. QUALITY CONTROL. NORMALIZED DATA

Let's do with our data:

```
#SAVE TO A FILE
pdf(file.path(resultsDir, "QCPlots_Norm.pdf"))
boxplot(eset, las=2, main="Intensity distribution of Normalized data", cex.axis=0.6,
        col=sampleColor, names=sampleNames)
plot(clust.euclid.average, labels=sampleNames, main="Hierarchical clustering of NormData",
     cex=0.7,  hang=-1)
plotPCA(exprs(eset), labels=sampleNames, dataDesc="selected samples", colors=sampleColor,
        formapunts=c(rep(16,4),rep(17,4)), myCex=0.6)
dev.off()
```

```
#ARRAY QUALITY METRICS
arrayQualityMetrics(eset,  reporttitle="QualityControl", force=TRUE)
```

4. Example of a microarray analysis with R. DATA FILTERING

Let's do with our data:

```
#-----------------------------------------------------------------------------
###FILTER OUT THE DATA
#-----------------------------------------------------------------------------

annotation(eset) <- "org.Mm.eg.db“

eset_filtered <- nsFilter(eset, var.func=IQR, var.cutoff=0.75, var.filter=TRUE,
         filterByQuantile=TRUE)

#NUMBER OF GENES OUT
print(eset_filtered$filter.log$numLowVar)

#NUMBER OF GENES IN
print(eset_filtered$eset)
```

4. Example of a microarray analysis with R. COMPARISONS

Let's do with our data:

```
#-----------------------------------------------------------------------------------------------
###DIFERENTIAL EXPRESSED GENES SELECTION. LINEAR MODELS. COMPARITIONS
#-----------------------------------------------------------------------------------------------

#CONTRAST MATRIX.lINEAR MODEL
treat <- targets$grupos
lev <- factor(treat, levels = unique(treat))
design <-model.matrix(~0+lev)
colnames(design) <- levels(lev)
rownames(design) <- sampleNames
print(design)


#COMPARISON
cont.matrix1 <- makeContrasts(
        Induced.vs.WT = Induced-WT,
        levels = design)
comparison1 <- "Effect of Induction"


#MODEL FIT
fit1 <- lmFit(eset_filtered$eset, design)
fit.main1 <- contrasts.fit(fit1, cont.matrix1)
fit.main1 <- eBayes(fit.main1)
```

4. Example of a microarray analysis with R. Results presentation. TOP TABLES

Let's do with our data:

```
#-----------------------------------------------------------------------------------
###DIFERENTIAL EXPRESSED GENES LISTS.TOPTABLES
#-----------------------------------------------------------------------------------

#FILTER BY FALSE DISCOVERY RATE AND FOLD CHANGE
topTab <-  topTable (fit.main1, number=nrow(fit.main1), coef="Induced.vs.WT",
adjust="fdr",lfc=abs(3))

#EXPORTED TO CSV AND HTML FILE
write.csv2(topTab, file= file.path(resultsDir,paste("Selected.Genes.in.comparison.",
                                                    comparison1, ".csv", sep = "")))

print(xtable(topTab,align="lllllll"),type="html",html.table.attributes="",
      file=paste("Selected.Genes.in.comparison.",comparison1,".html", sep=""))
```

4. Example of a microarray analysis with R. Results presentation. VOLCANO PLOTS

Let's do with our data:

```
#-----------------------------------------------------------------------------------
###VOLCANO PLOTS
#-----------------------------------------------------------------------------------

volcanoplot(fit.main1, highlight=10, names=fit.main1$ID,
            main = paste("Differentially expressed genes", colnames(cont.matrix1),
sep="\n"))
abline(v = c(-3, 3))


pdf(file.path(resultsDir,"Volcanos.pdf"))
volcanoplot(fit.main1, highlight = 10, names = fit.main1$ID,
            main = paste("Differentially expressed genes", colnames(cont.matrix1),
sep = "\n"))
abline(v = c(-3, 3))
dev.off()
```

4. Example of a microarray analysis with R. Results presentation. CLUSTERING

Let's do with our data:

```
#-----------------------------------------------------------------------------------------
###HEATMAP PLOTS
#-----------------------------------------------------------------------------------------

#PREPARE THE DATA
my_frame <- data.frame(exprs(eset))
head(my_frame)
HMdata <- merge(my_frame, topTab, by.x = 0, by.y = 0)
rownames(HMdata) <- HMdata$Row.names
HMdata <- HMdata[, -c(1,10:15)]
head(HMdata)
HMdata2 <- data.matrix(HMdata, rownames.force=TRUE)
head(HMdata2)
write.csv2(HMdata2, file = file.path(resultsDir,"Data2HM.csv"))

#HEATMAP PLOT
my_palette <- colorRampPalette(c("blue", "red"))(n = 299)

heatmap.2(HMdata2,
          Rowv=TRUE,
          Colv=TRUE,
          main="HeatMap Induced.vs.WT FC>=3",
          scale="row",
          col=my_palette,
          sepcolor="white",
          sepwidth=c(0.05,0.05),
          cexRow=0.5,
          cexCol=0.9,
          key=TRUE,
          keysize=1.5,
          density.info="histogram",
          ColSideColors=c(rep("red",4),rep("blue",4)),
          tracecol=NULL,
          srtCol=30)
```

4. Example of a microarray analysis with R. Results presentation. CLUSTERING

Let's do with our data:

```
#EXPORT TO PDF FILE
pdf(file.path(resultsDir,"HeatMap InducedvsWT.pdf"))
heatmap.2(HMdata2,
          Rowv=TRUE,
          Colv=TRUE,
          main="HeatMap Induced.vs.WT FC>=3",
          scale="row",
          col=my_palette,
          sepcolor="white",
          sepwidth=c(0.05,0.05),
          cexRow=0.5,
          cexCol=0.9,
          key=TRUE,
          keysize=1.5,
          density.info="histogram",
          ColSideColors=c(rep("red",4),rep("blue",4)),
          tracecol=NULL,
          srtCol=30)
dev.off()
```

4. Example of a microarray analysis with R. Results presentation. ANNOTATION

Let's do with our data:

```
#-----------------------------------------------------------------------------------------
###DATA ANNOTATION
#-----------------------------------------------------------------------------------------

all_anota<-data.frame(exprs(eset))
Annot <- data.frame(SYMBOL=sapply(contents(mogene10sttranscriptclusterSYMBOL), paste, collapse=", "),
                    DESC=sapply(contents(mogene10sttranscriptclusterGENENAME), paste, collapse=", "))
Annot<-Annot[!Annot$SYMBOL=="NA",]
Annot<-Annot[!Annot$DESC=="NA",]
head(Annot)

anotaGenes <- merge(Annot,all_anota, by.x=0,by.y=0)
head(anotaGenes)
write.table(anotaGenes, file ="data.ann.txt",sep="\t")

rownames(anotaGenes) <- anotaGenes[,1]
anotaGenes <- anotaGenes[,-1]
anotaGenes.end <- merge(anotaGenes, topTab, by.x=0,by.y=0)
#reordenamos las columnas
topTab.end <- anotaGenes.end[,c(1:3,12:17,4:11)]
topTab.end <- topTab.end[order(-topTab.end$B),]

rownames(topTab.end) <- topTab.end[,1]
topTab.end <- topTab.end[, -1]
write.csv(topTab.end, file = file.path(resultsDir,"TopTable.end.csv"))
```

3. Introduction to Biological Significance

Our case study

Affy Ids that we may need to map to other (gene Ids, gene Symbols,…)

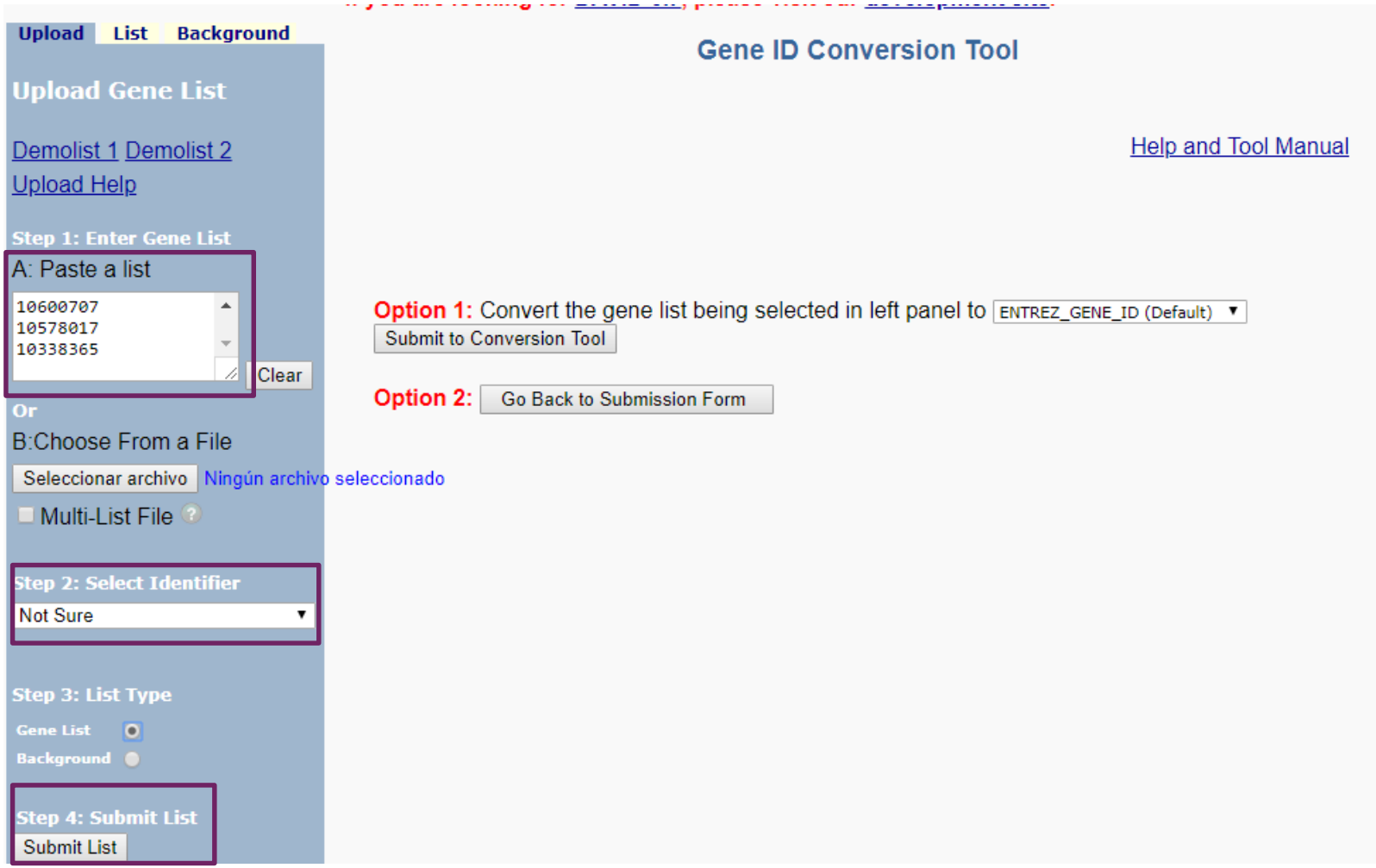
AffyID	logFC	AveExpr	t	P.Value	adj.P.Val	B
10470175	6,70741636	7,36410713	33,6741217	1,19E-10	6,64E-07	14,4533744
10351443	6,77982467	9,60024622	32,8302066	1,49E-10	6,64E-07	14,2928807
10403796	6,65419808	8,31661974	30,9631689	2,50E-10	6,78E-07	13,9121895
10522388	6,53131792	8,43170671	29,7723358	3,52E-10	6,78E-07	13,6493
10469358	-6,37121011	7,74774275	-29,5017644	3,82E-10	6,78E-07	13,5872221
10531869	6,75197558	8,78504933	28,4970669	5,17E-10	7,00E-07	13,3486709
10400926	6,28127721	9,12688557	28,2473608	5,58E-10	7,00E-07	13,2873456
10499189	-4,9732455	7,91668064	-27,8612518	6,30E-10	7,00E-07	13,1908635
10474524	-4,38129973	6,24078024	-27,4403222	7,20E-10	7,00E-07	13,0833286
10455942	6,00894252	7,93734767	26,6954373	9,16E-10	7,00E-07	12,8867945
10482772	4,71424943	10,3274567	26,6745827	9,22E-10	7,00E-07	12,8811741
10464370	6,24393064	8,40464488	26,6002636	9,45E-10	7,00E-07	12,8610914
10382341	4,76224664	8,9009275	25,2626123	1,48E-09	9,39E-07	12,4848696
10362372	4,51831162	7,64924398	25,1782749	1,53E-09	9,39E-07	12,4601768
10345791	-4,24641517	8,08063532	-24,948231	1,66E-09	9,39E-07	12,392212
10497713	4,66592009	8,44015956	24,8900286	1,69E-09	9,39E-07	12,3748735
10517513	-4,24650481	7,90145012	-24,4628279	1,97E-09	1,03E-06	12,2458086
10466200	-6,23115591	7,83898265	-23,9419475	2,37E-09	1,17E-06	12,0840392
10469816	-4,19348905	8,74058362	-23,6413874	2,65E-09	1,24E-06	11,9884213
10540401	6,92746622	8,76188071	23,2762611	3,03E-09	1,28E-06	11,869954
10477986	4,37136952	10,4194869	23,2759003	3,03E-09	1,28E-06	11,8698357
10386211	4,19200484	7,71904509	23,0536971	3,30E-09	1,33E-06	11,7964699
10560919	6,08134785	9,0248543	22,628519	3,88E-09	1,46E-06	11,6533367
10585484	4,15947657	7,35673793	22,5806864	3,95E-09	1,46E-06	11,6370038
10363082	-4,96086344	7,76193693	-21,6524512	5,69E-09	2,02E-06	11,3104631
10569370	6,92471584	8,96414504	21,3010725	6,56E-09	2,18E-06	11,1819164
10563597	8,45161149	11,2652164	21,1684004	6,93E-09	2,18E-06	11,1326469
10446965	-4,58754189	6,95461918	-21,1425264	7,00E-09	2,18E-06	11,1229907
10476633	4,91102252	8,8463601	21,0998861	7,13E-09	2,18E-06	11,1070432
10547657	-6,01048925	8,957912	-20,7390712	8,28E-09	2,45E-06	10,9703812
10513467	7,11895464	7,53720571	20,638981	8,64E-09	2,48E-06	10,9319188
10606835	6,32944055	8,94956227	20,2952985	9,99E-09	2,74E-06	10,7979801
10573979	6,59597743	9,91300765	20,1942547	1,04E-08	2,74E-06	10,7580421
10602896	-6,77562265	8,22448493	-20,1804935	1,05E-08	2,74E-06	10,752583
10606868	6,25711539	7,82975845	19,9770021	1,15E-08	2,91E-06	10,671294

3. Introduction to Biological Significance

Our case study: List with genes fold change > 3

https://david.ncifcrf.gov/home.jsp

3. Introduction to Biological Significance



3. Introduction to Biological Significance

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

Upload | List | Background

Gene List Manager

Select to limit annotations by one or more species Help

- Use All Species -
Mus musculus(308)

Select Species

List Manager Help

new_converted_list

Select List to:

Use | Rename
Remove | Combine
Show Gene List

Gene ID Conversion Tool

Help and Tool Manual

You are either not sure which identifier type your list contains, or less than 80% of your list has mapped to your chosen identifier type. Please use the Gene Conversion Tool to determine the identifier type.

Option 1: Convert the gene list to ENTREZ_GENE_ID (Default) Submit to Conversion Tool

Option 2: Go Back to Submission Form

3. Introduction to Biological Significance

Gene Accession Conversion Tool

Gene Accession Conversion Statistics

Conversion Summary		
ID Count	In DAVID DB	Conversion
0	Yes	Successful
0	Yes	None
17	No	None
309	Ambiguous	Pending
Total Unique User IDs: 326		

Summary of Ambiguous Gene IDs		
ID Count	Possible Source	Convert All
3	MRNA_GI_ACCESSION	
129	ENTREZ_GENE_ID	
2	PROTEIN_GI_ACCESSION	
308	AFFYMETRIX_EXON_ID	

All Possible Sources For Ambiguous IDs		
Ambiguous ID	Possiblity	Convert
10358557	AFFYMETRIX_EXON_ID	
10358557	ENTREZ_GENE_ID	
10517513	AFFYMETRIX_EXON_ID	
10517513	ENTREZ_GENE_ID	
10601903	AFFYMETRIX_EXON_ID	
10358555	AFFYMETRIX_EXON_ID	
10358555	ENTREZ_GENE_ID	
10578017	AFFYMETRIX_EXON_ID	
10456363	AFFYMETRIX_EXON_ID	
10490665	AFFYMETRIX_EXON_ID	
10363224	AFFYMETRIX_EXON_ID	
10564159	AFFYMETRIX_EXON_ID	
10564159	ENTREZ_GENE_ID	
10368495	AFFYMETRIX_EXON_ID	

3. Introduction to Biological Significance

Gene Accession Conversion Tool

Gene Accession Conversion Statistics

Conversion Summary		
ID Count	In DAVID DB	Conversion
0	Yes	Successful
0	Yes	None
17	No	None
309	Ambiguous	Pending
Total Unique User IDs: 326		

Summary of Ambiguous Gene IDs		
ID Count	Possible Source	Convert All
3	MRNA_GI_ACCESSION	
129	ENTREZ_GENE_ID	
2	PROTEIN_GI_ACCESSION	
308	AFFYMETRIX_EXON_ID	

All Possible Sources For Ambiguous IDs		
Ambiguous ID	Possiblity	Convert
10358557	AFFYMETRIX_EXON_ID	
10358557	ENTREZ_GENE_ID	
10517513	AFFYMETRIX_EXON_ID	
10517513	ENTREZ_GENE_ID	
10601903	AFFYMETRIX_EXON_ID	
10358555	AFFYMETRIX_EXON_ID	
10358555	ENTREZ_GENE_ID	
10578017	AFFYMETRIX_EXON_ID	
10456363	AFFYMETRIX_EXON_ID	
10490665	AFFYMETRIX_EXON_ID	
10363224	AFFYMETRIX_EXON_ID	
10564159	AFFYMETRIX_EXON_ID	
10564159	ENTREZ_GENE_ID	
10368495	AFFYMETRIX_EXON_ID	

Download File

Submit Converted List to DAVID as a Gene List | Submit Converted List to DAVID as a Background

From	To	Species	David Gene Name
10358565	545370	Mus musculus	hemicentin 1(Hmcn1)
10577641	69068	Mus musculus	RIKEN cDNA 1810011O10 gene(1810011O10Rik)
10358567	545370	Mus musculus	hemicentin 1(Hmcn1)
10358561	545370	Mus musculus	hemicentin 1(Hmcn1)
10355960	20254	Mus musculus	secretogranin II(Scg2)
10358563	545370	Mus musculus	hemicentin 1(Hmcn1)
10474700	21825	Mus musculus	thrombospondin 1(Thbs1)
10584024	330908	Mus musculus	opioid binding protein/cell adhesion molecule-like(Opcml)
10476512	20614	Mus musculus	synaptosomal-associated protein 25(Snap25)
10578017	100101427	Mus musculus	predicted gene 9911(Gm9911)
10522208	22223	Mus musculus	ubiquitin carboxy-terminal hydrolase L1(Uchl1)
10450325	14962	Mus musculus	complement factor B(Cfb)
10498337	229320	Mus musculus	clarin 1(Clrn1)
10382341	20606	Mus musculus	somatostatin receptor 2(Sstr2)
10601385	279572	Mus musculus	toll-like receptor 13(Tlr13)
10428388	239405	Mus musculus	R-spondin 2(Rspo2)
10498885	14800	Mus musculus	glutamate receptor, ionotropic, AMPA2 (alpha 2)(Gria2)
10358557	545370	Mus musculus	hemicentin 1(Hmcn1)

3. Introduction to Biological Significance

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

Upload | List | Background

Gene List Manager

Select to limit annotations by one or more species Help

- Use All Species -
Mus musculus(308)

Select Species

List Manager Help

new_converted_list
new_converted_list

Select List to:

Use | Rename
Remove | Combine
Show Gene List

Analysis Wizard

Tell us how you like the tool
Contact us for questions

Step 1. Successfully submitted gene list
Current Gene List: new_converted_list
Current Background: Mus musculus

Step 2. Analyze above gene list with one of DAVID tools

Which DAVID tools to use?

- Functional Annotation Tool
  - Functional Annotation Clustering
  - Functional Annotation Chart
  - Functional Annotation Table
- Gene Functional Classification Tool
- Gene ID Conversion Tool
- Gene Name Batch Viewer

3. Introduction to Biological Significance

*** Welcome to DAVID 6.8 ***

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

Upload | List | Background

Gene List Manager

Select to limit annotations by one or more species Help

- Use All Species -
Mus musculus(308)

Select Species

List Manager Help

new_converted_list
new_converted_list

Select List to:

Use | Rename | Remove | Combine | Show Gene List

Annotation Summary Results

Help and Tool Manual

Current Gene List: new_converted_list — 229 DAVID IDs

Current Background: Mus musculus — Check Defaults ☑ — Clear All

- Functional_Categories (3 selected)
- Gene_Ontology (3 selected)
- General_Annotations (0 selected)
- Literature (0 selected)
- Main_Accessions (0 selected)
- Pathways (2 selected)
- Protein_Domains (3 selected)
- Protein_Interactions (0 selected)
- Tissue_Expression (0 selected)

Red annotation categories denote DAVID defined defaults

Combined View for Selected Annotation

Functional Annotation Clustering

Functional Annotation Chart

Functional Annotation Table

3. Introduction to Biological Significance

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

Functional Annotation Clustering

Help and Manual

Current Gene List: new_converted_list
Current Background: Mus musculus
229 DAVID IDs

Options Classification Stringency Medium

Rerun using options Create Sublist

50 Cluster(s) Download File

Annotation Cluster 1	Enrichment Score: 13.68		Count	P_Value	Benjamini
UP_KEYWORDS	Glycoprotein	RT	95	1.6E-19	3.7E-17
UP_SEQ_FEATURE	signal peptide	RT	87	7.3E-18	5.1E-15
UP_KEYWORDS	Disulfide bond	RT	81	3.6E-17	4.1E-15
UP_KEYWORDS	Signal	RT	97	2.7E-15	2.0E-13
UP_KEYWORDS	Secreted	RT	53	6.7E-14	3.8E-12
UP_SEQ_FEATURE	disulfide bond	RT	69	3.4E-13	1.2E-10
UP_SEQ_FEATURE	glycosylation site:N-linked (GlcNAc...)	RT	82	8.3E-12	1.9E-9
GOTERM_CC_DIRECT	extracellular region	RT	52	7.3E-11	1.8E-8
GOTERM_CC_DIRECT	extracellular space	RT	46	5.2E-10	3.1E-8
Annotation Cluster 2	**Enrichment Score: 7.85**		**Count**	**P_Value**	**Benjamini**
UP_KEYWORDS	Synapse	RT	21	4.1E-10	1.9E-8
GOTERM_CC_DIRECT	synapse	RT	25	1.9E-9	9.0E-8
UP_KEYWORDS	Cell junction	RT	25	3.8E-8	1.2E-6
GOTERM_CC_DIRECT	cell junction	RT	25	1.3E-6	2.9E-5
Annotation Cluster 3	**Enrichment Score: 6.85**		**Count**	**P_Value**	**Benjamini**
UP_KEYWORDS	Cleavage on pair of basic residues	RT	18	4.2E-10	1.6E-8
GOTERM_BP_DIRECT	neuropeptide signaling pathway	RT	9	1.7E-6	2.0E-3
GOTERM_CC_DIRECT	secretory granule	RT	10	4.1E-6	8.2E-5
Annotation Cluster 4	**Enrichment Score: 6.12**		**Count**	**P_Value**	**Benjamini**
GOTERM_CC_DIRECT	terminal bouton	RT	13	4.2E-9	1.7E-7
GOTERM_CC_DIRECT	synaptic vesicle	RT	12	1.5E-7	4.4E-6
GOTERM_CC_DIRECT	synaptic vesicle membrane	RT	8	4.6E-6	8.6E-5
GOTERM_BP_DIRECT	chemical synaptic transmission	RT	10	1.1E-4	3.3E-2
Annotation Cluster 5	**Enrichment Score: 3.81**		**Count**	**P_Value**	**Benjamini**
GOTERM_BP_DIRECT	neuropeptide signaling pathway	RT	9	1.7E-6	2.0E-3