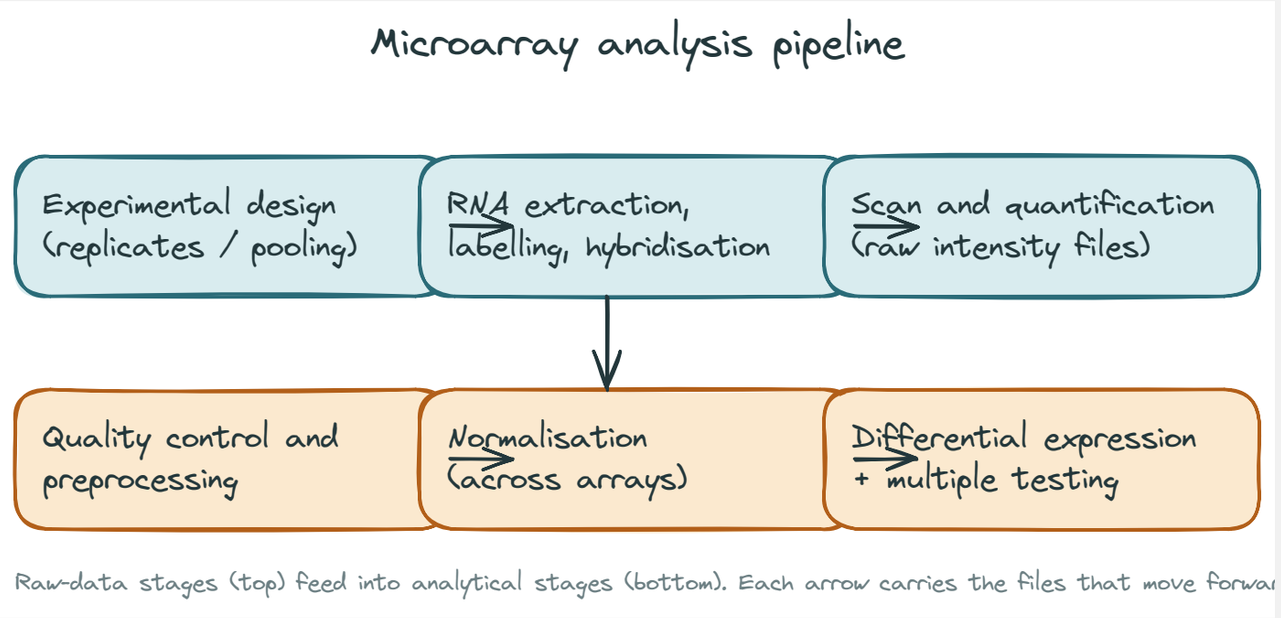
Preprocessing and differential expression

Bioinformatics · Universitat Politècnica de Catalunya

Alexandre Perera-Lluna

Workflow

1. Importing Data, Formatting Metadata.
2. Exploratory Analysis, Data Visualization.
3. Preprocessing.
4. Quality Assessment.
5. Differential Expression.
6. Functional Analysis.



The six stages shown are the steps this lecture covers in sequence.

Preprocessing

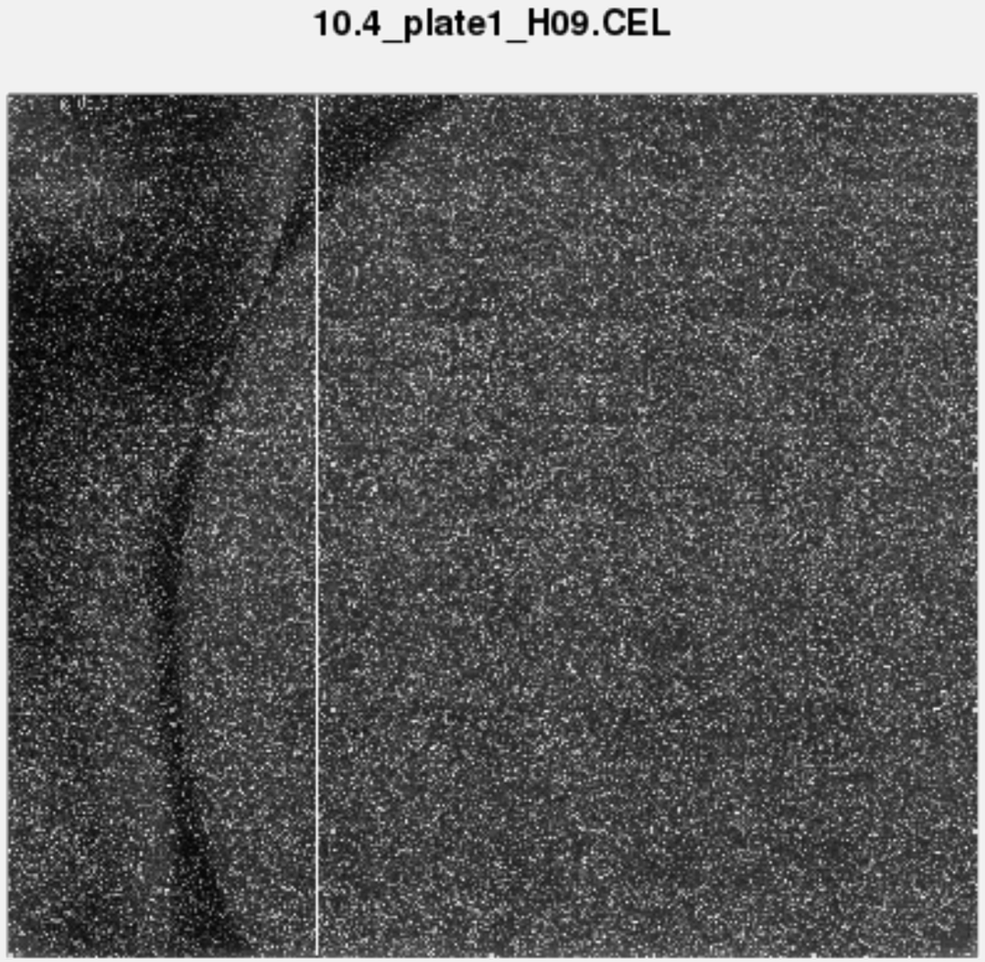
Adjusting systematic effects due to technology variation, not biology variation.

1. Background Correction.
2. Normalization.
3. Summarization (some platforms).

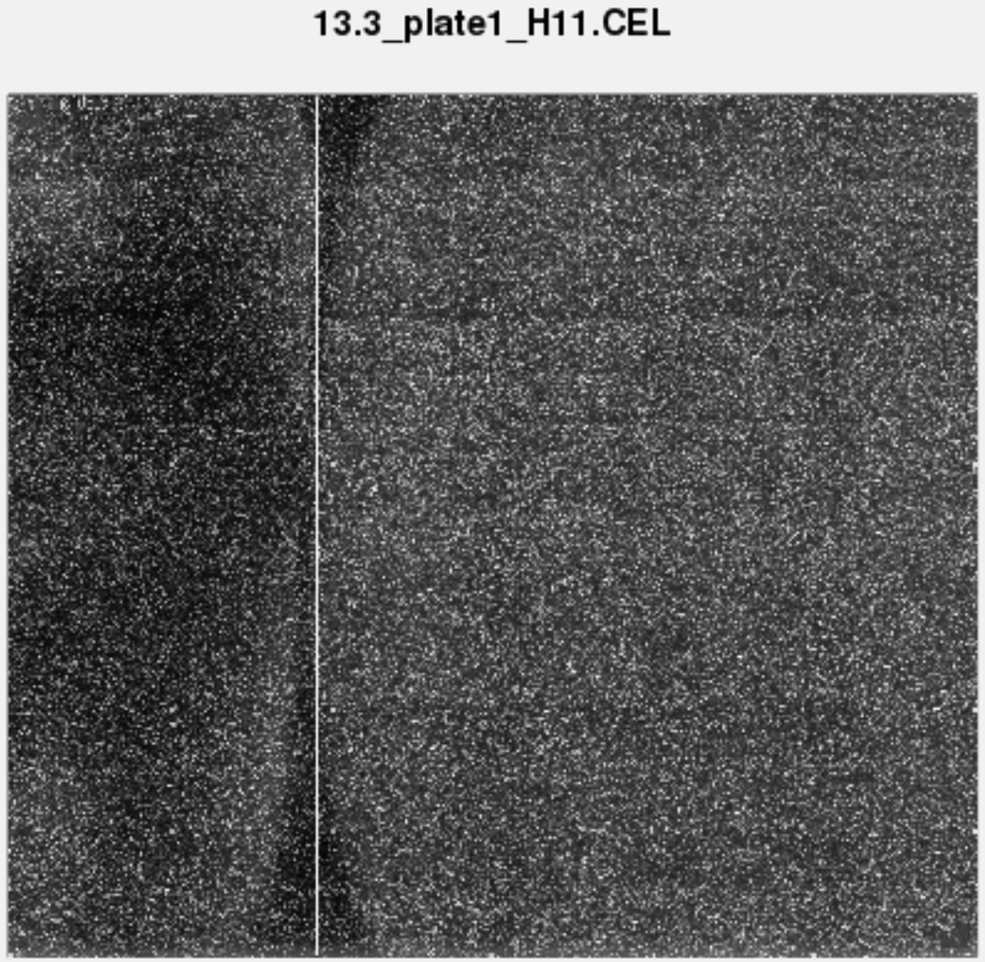
Background correction

Subtraction of the background intensities from the foreground at each spot. Analysis per-sample.

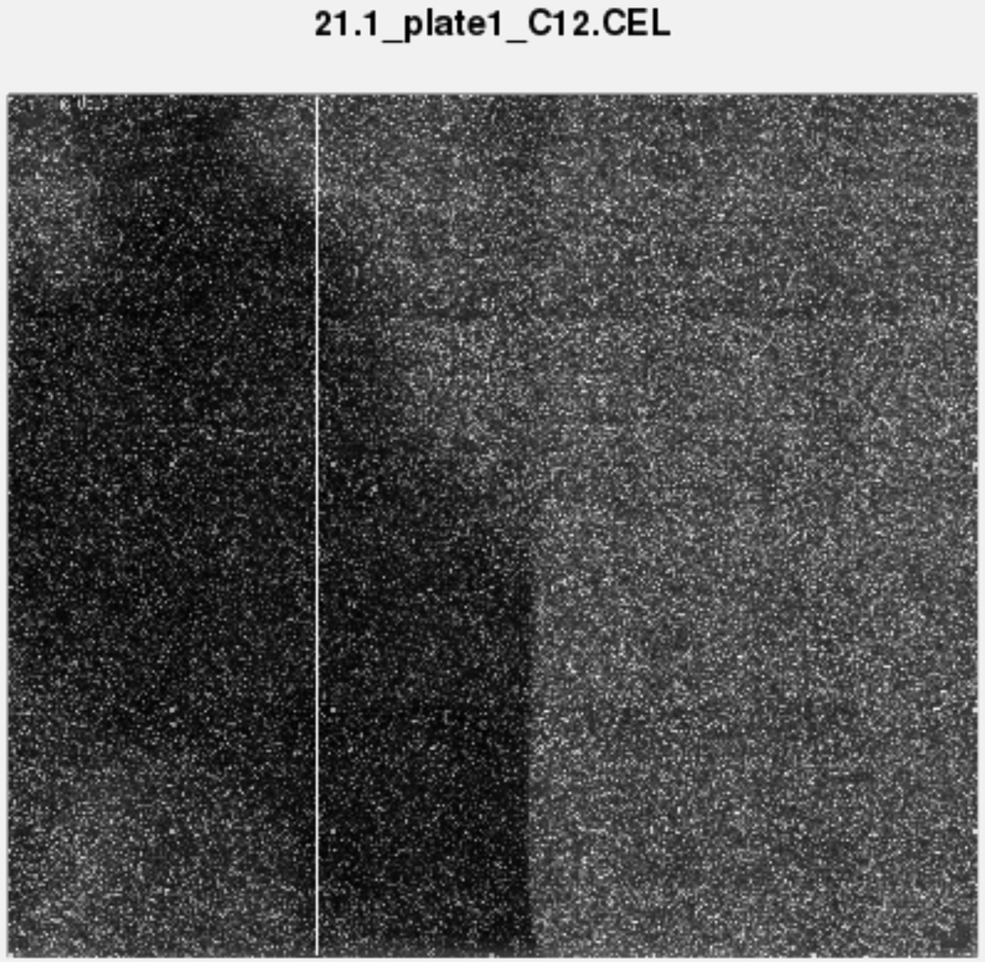
Exploratory Analysis (1/5)



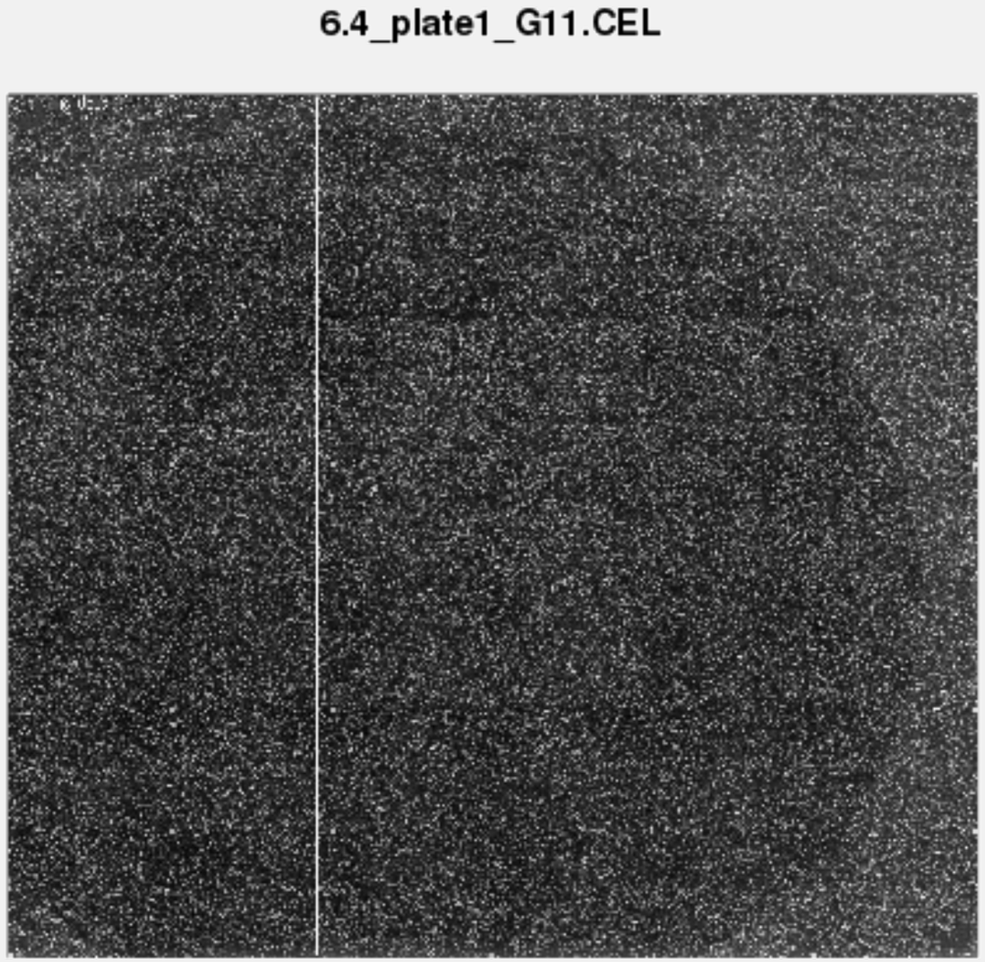
Exploratory Analysis (2/5)



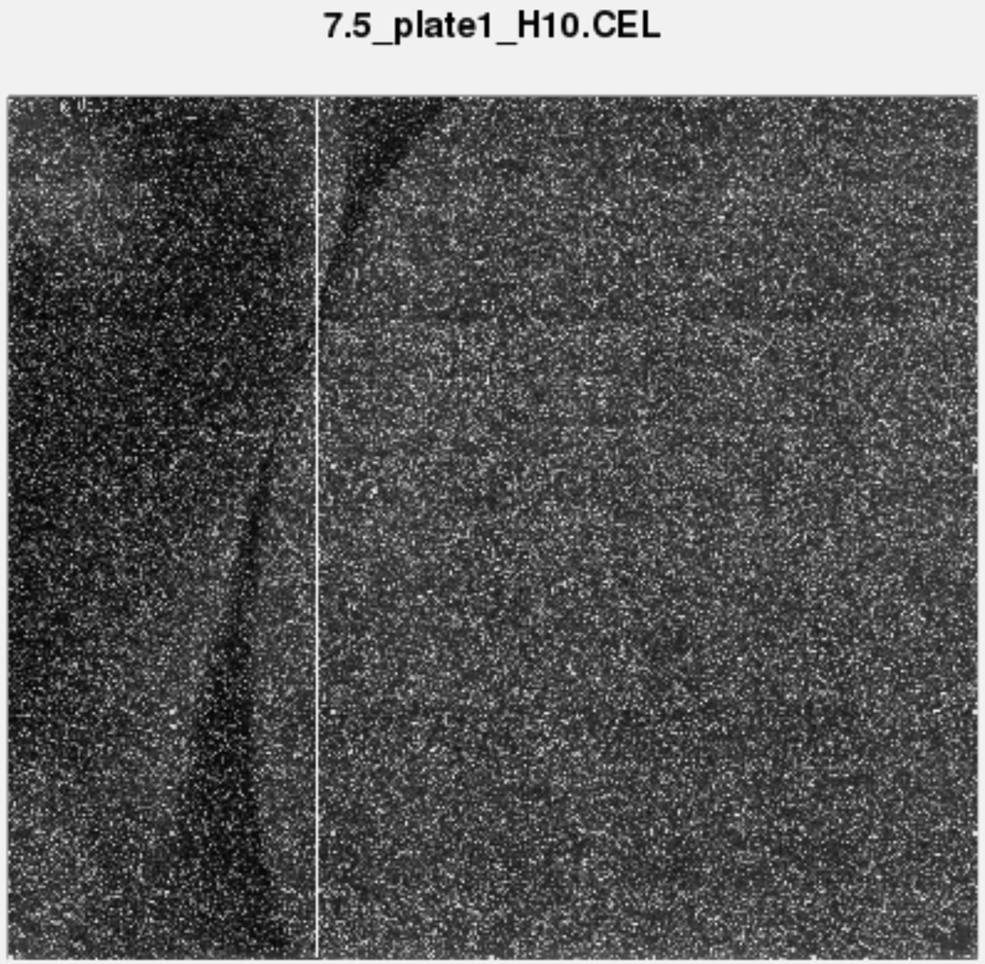
Exploratory Analysis (3/5)



Exploratory Analysis (4/5)



Exploratory Analysis (5/5)

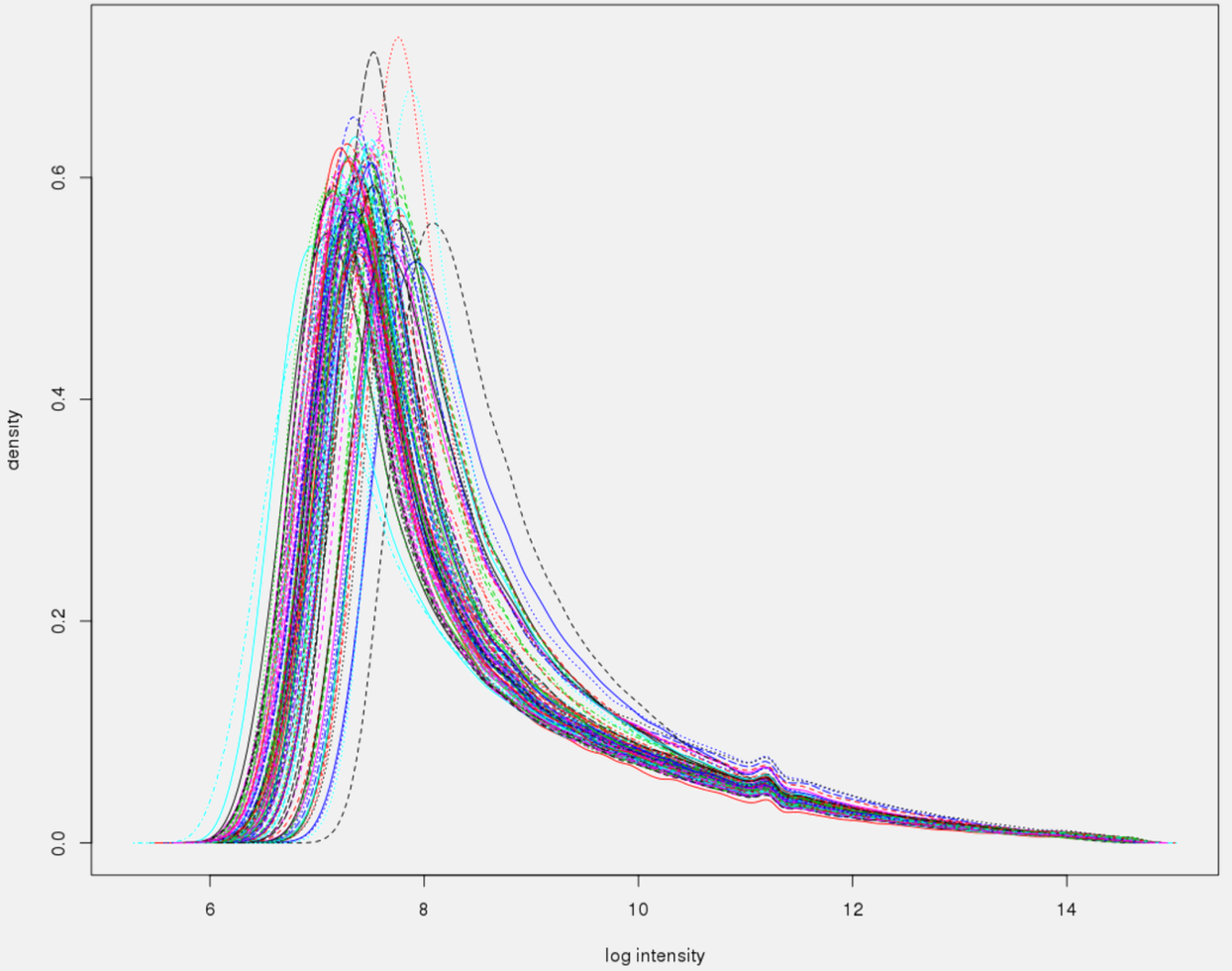


Normalization

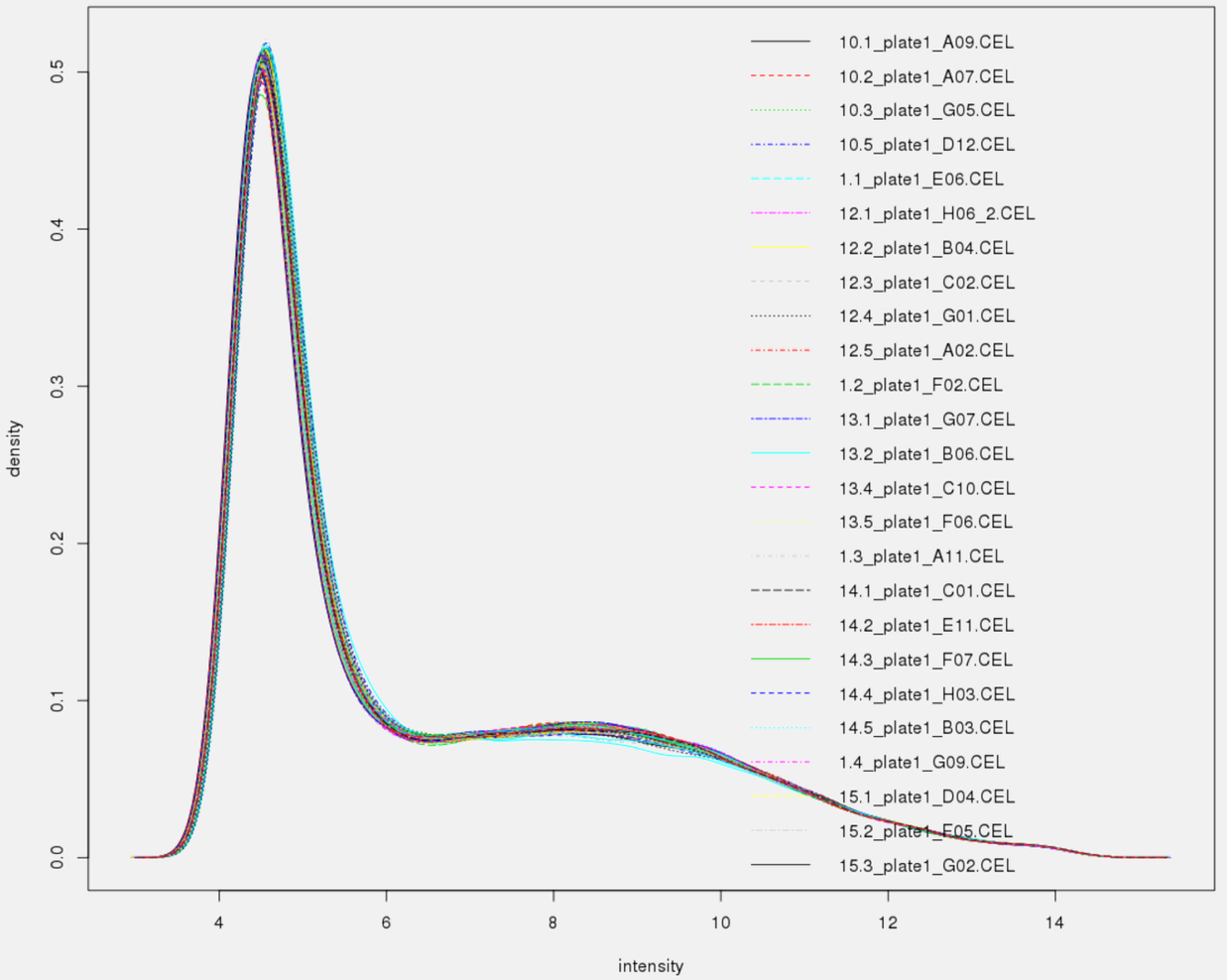
Remove systematic variation from the data, scaling data so that comparison can be made across samples.

- It is a critical step.
- Normalization Between Arrays, and Within Arrays.

Normalization within arrays



Post Normalization within arrays



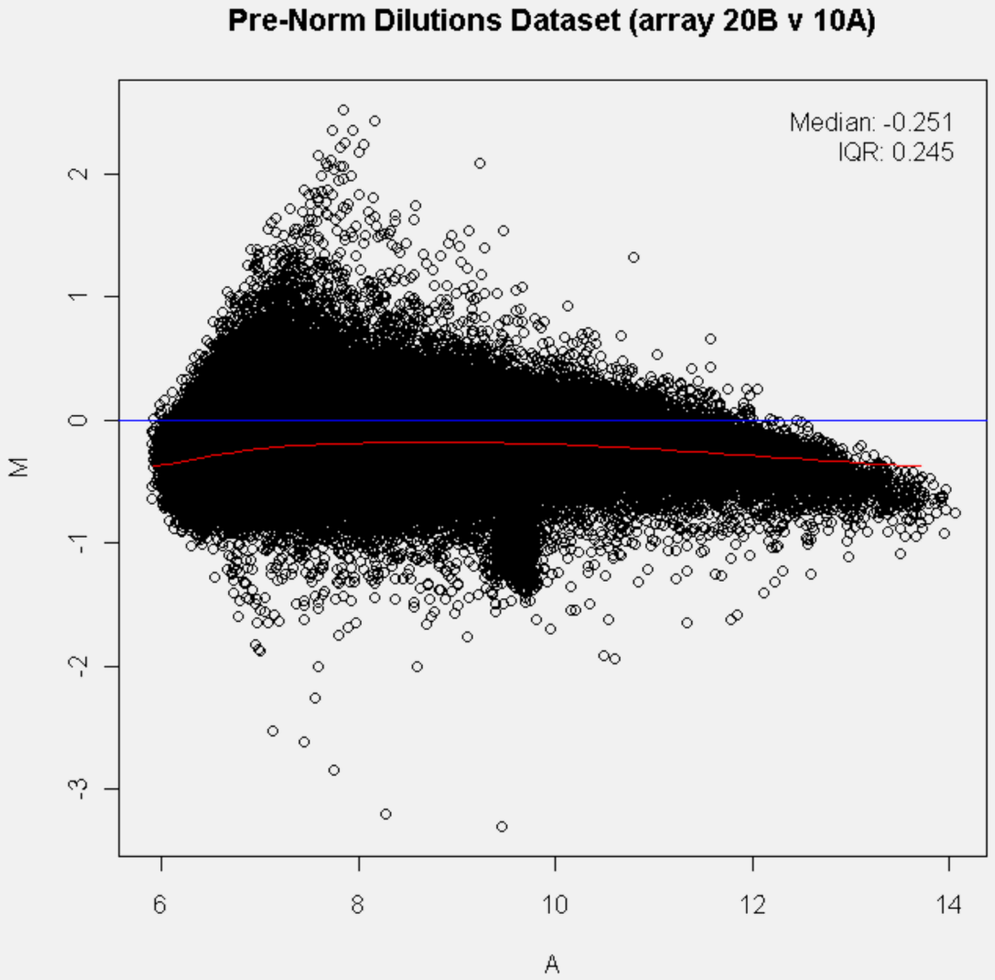
Red/green intensity ratio (M)

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

Average intensity (A)

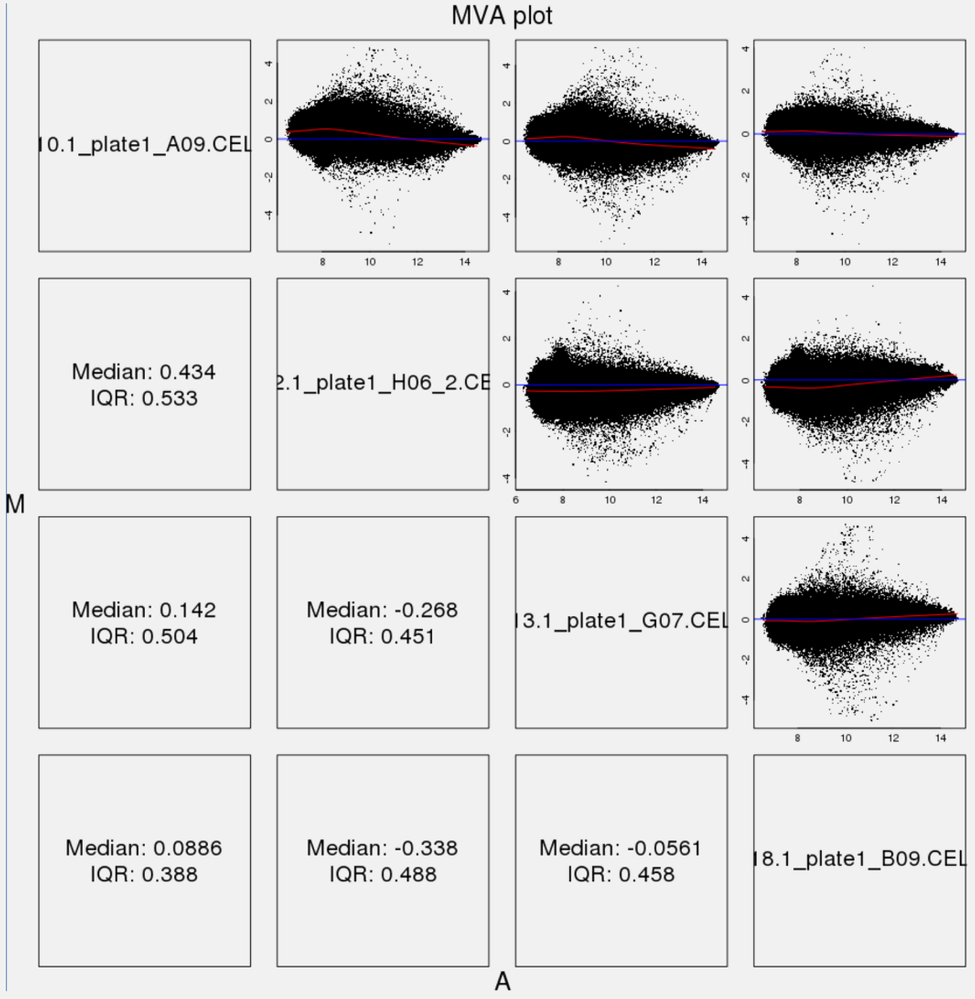
$$A = \frac{1}{2} log_2(RG) = \frac{log_2 R + log_2 G}{2}$$

MA Plots


Pre-Norm Dilutions Dataset (array 20B v 10A)
Median: -0.251
IQR: 0.245
M
A
2
1
0
-1
-2
-3
6
8
10
12
14

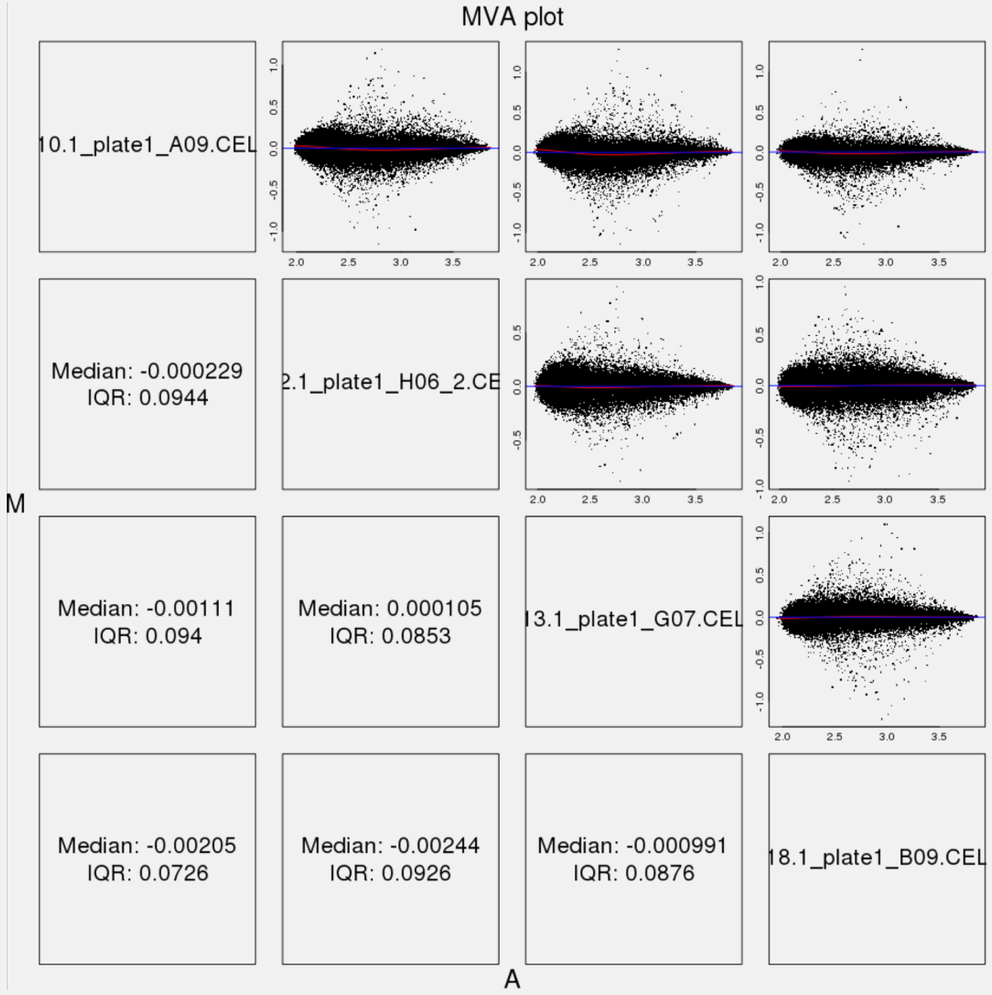
Prior to Normalization Between Arrays (1/2)

Pairwise scatter plots before between-array normalisation.



After Normalization Between Arrays

Pairwise scatter plots after between-array normalisation.



RMA

Background adjustment, quantile normalisation, and robust probe-set summarisation on the log scale.

GC-RMA

Extends background modelling to account for probe sequence composition and non-specific binding.

Li–Wong model

Model-based expression estimation accounts for probe-specific response. It is distinct from loess smoothing.

The appropriate method depends on array platform and experimental design.

Two approaches

1. Classical statistical tests.
2. Linear models.

workflow

- Generate Hypothesis.
- Choose appropriate statistical test.
- Compute p-value.
- Correct for multiple testing.
- Decide on confidence p-value threshold.
- List selected.

Choosing a t-test

Comparison	Test
One group mean against a specified value	One-sample t-test
Matched or repeated measurements	Paired t-test on within-pair differences
Two independent group means	Welch two-sample t-test

The pooled-variance test additionally assumes equal population variances. Inspect the study design, sample size, and distribution of errors or differences.

One-sample t-test (1/2)

- t-statistic:

$$t = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

```r
x <- rnorm(100, mean=2,sd=0.5)
t.test(x)
```

```
    One Sample t-test

data:  x
t = 38.277, df = 99, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 1.936873 2.148657
sample estimates:
mean of x
 2.042765
```

One-sample t-test (2/2)


```r
x <- rnorm(100, mean=0,sd=0.5)
t.test(x)
```

```
    One Sample t-test

data:  x
t = 1.7802, df = 99, p-value = 0.07811
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.01084199  0.20005835
sample estimates:
 mean of x
0.09460818
```

Two-sample t-tests (1/3)

For independent samples, the Welch statistic is

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - \Delta_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}.$$

Δ_0 is the difference specified by the null hypothesis. Welch's test estimates degrees of freedom from the sample variances.

The examples compare a pooled-variance test with Welch's test.

Two-sample t-tests (2/3)


```r
x = c(15, 10, 13, 7, 9, 8, 21, 9, 14, 8)
 y = c(15, 14, 12, 8, 14, 7, 16, 10, 15, 12)
t.test(x,y,alt="less",var.equal=TRUE)
```

```
    Two Sample t-test

data:  x and y
t = -0.53311, df = 18, p-value = 0.3002
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 2.027436
sample estimates:
mean of x mean of y
     11.4      12.3
```

Two-sample t-tests (3/3)

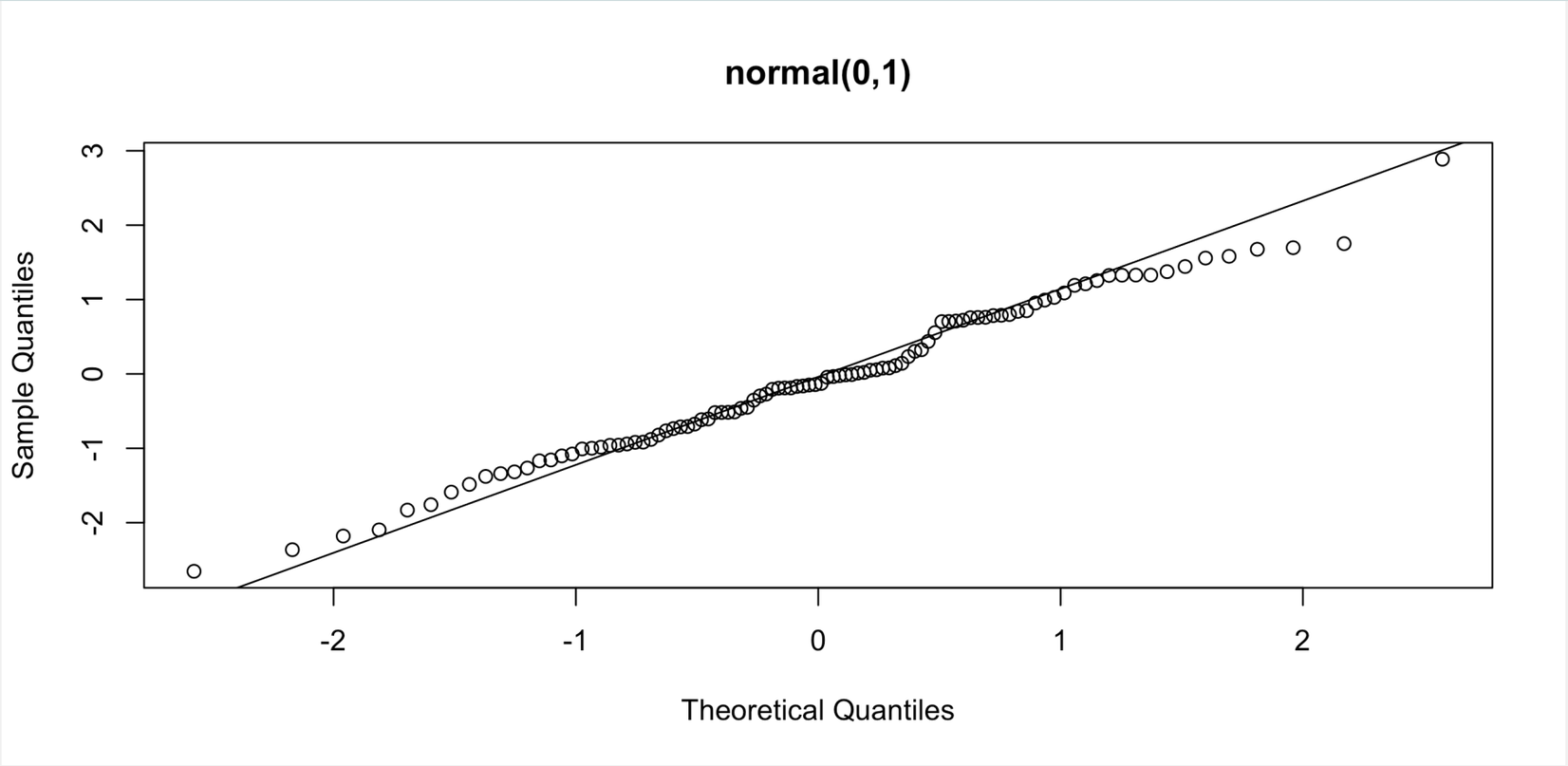

```r
t.test(x,y,alt="less")
```

```
    Welch Two Sample t-test

data:  x and y
t = -0.53311, df = 16.245, p-value = 0.3006
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 2.044664
sample estimates:
mean of x mean of y
     11.4      12.3
```

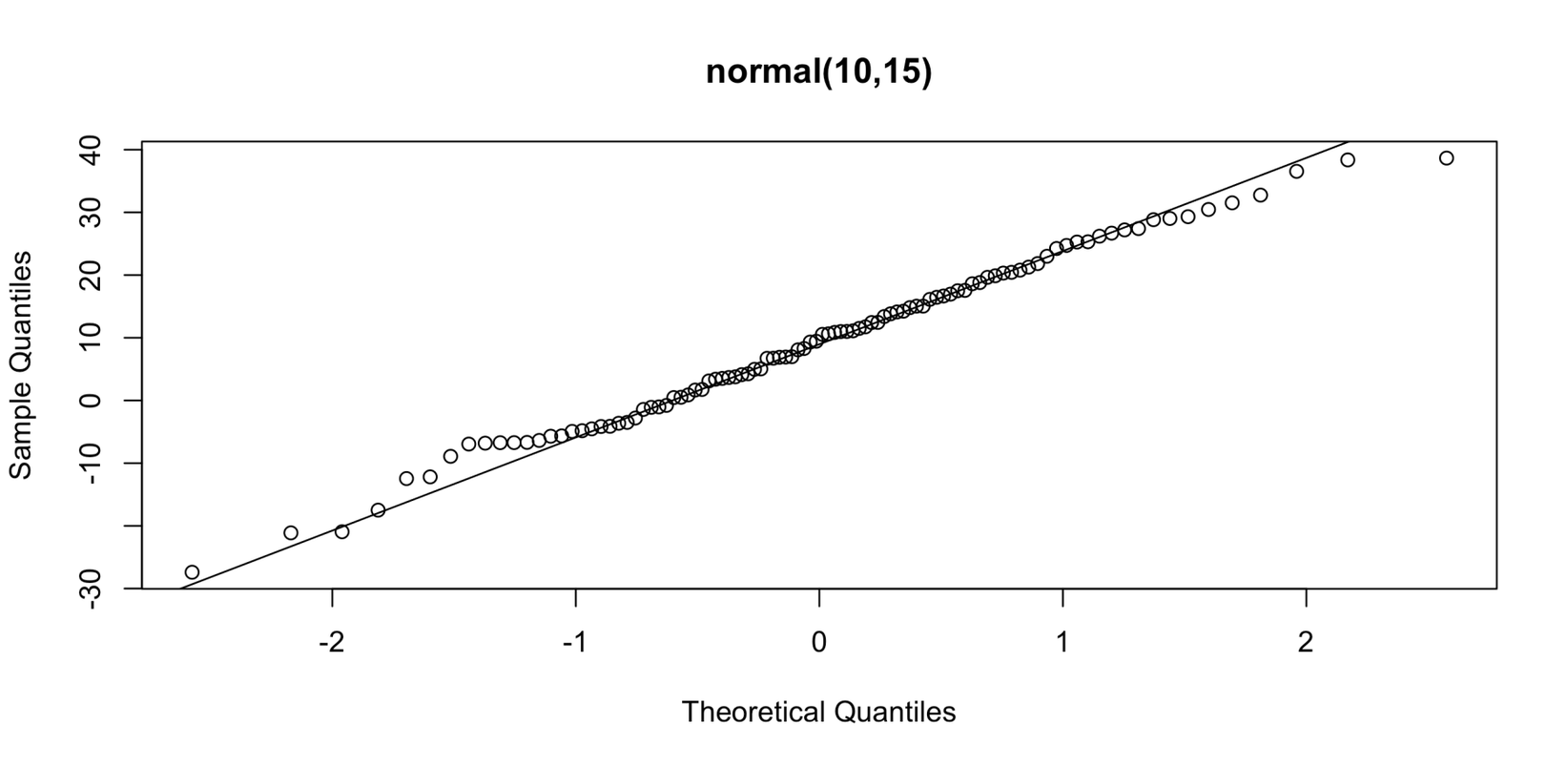
Non-Gaussian Data: Q–Q plots (1/2)

```
1  x = rnorm(100,0,1);qqnorm(x,main='normal(0,1)');qqline(x)
```



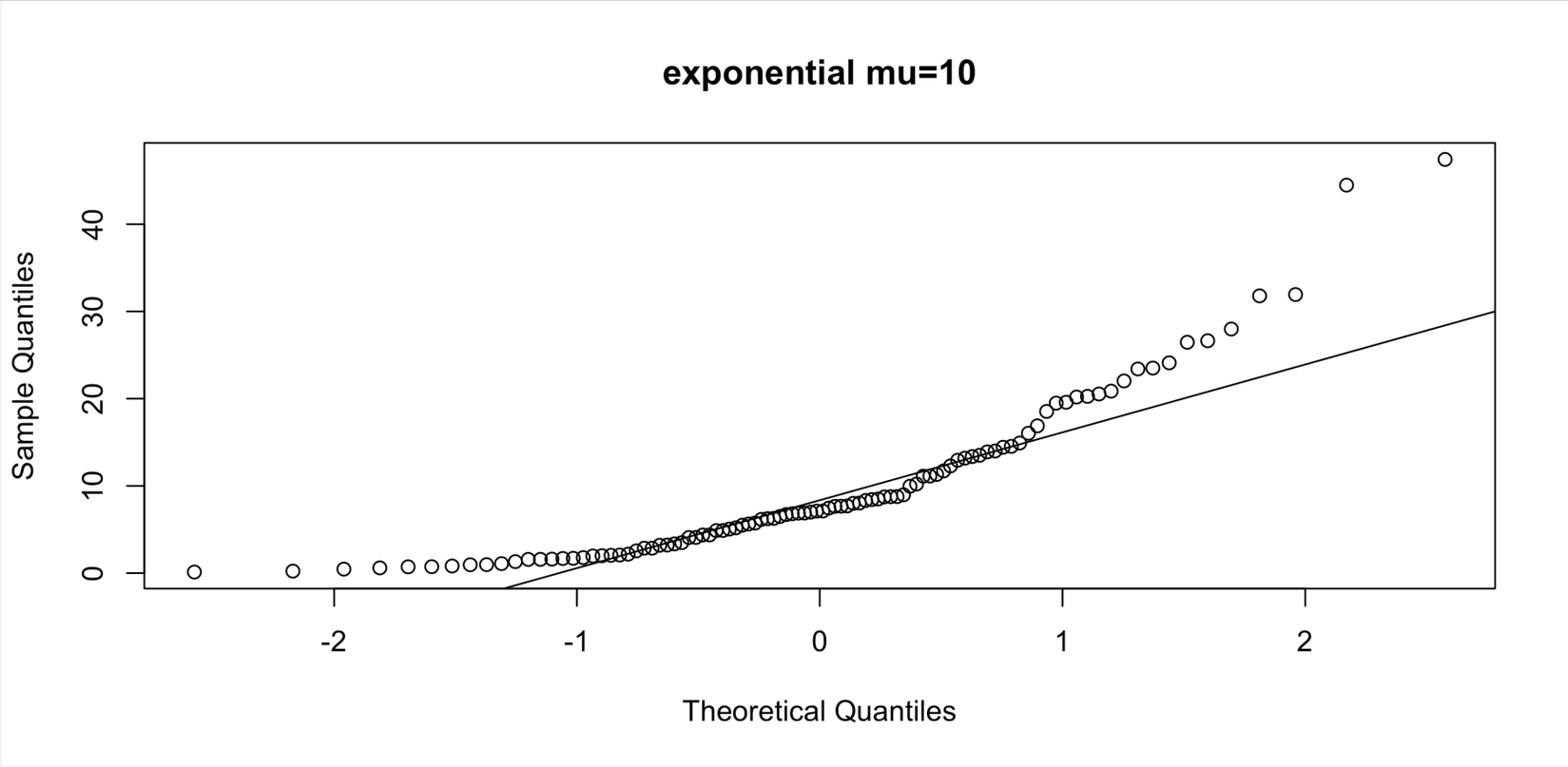
Non-Gaussian Data: Q–Q plots (2/2)

```
1  x = rnorm(100,10,15);qqnorm(x,main='normal(10,15)');qqline(x)
```



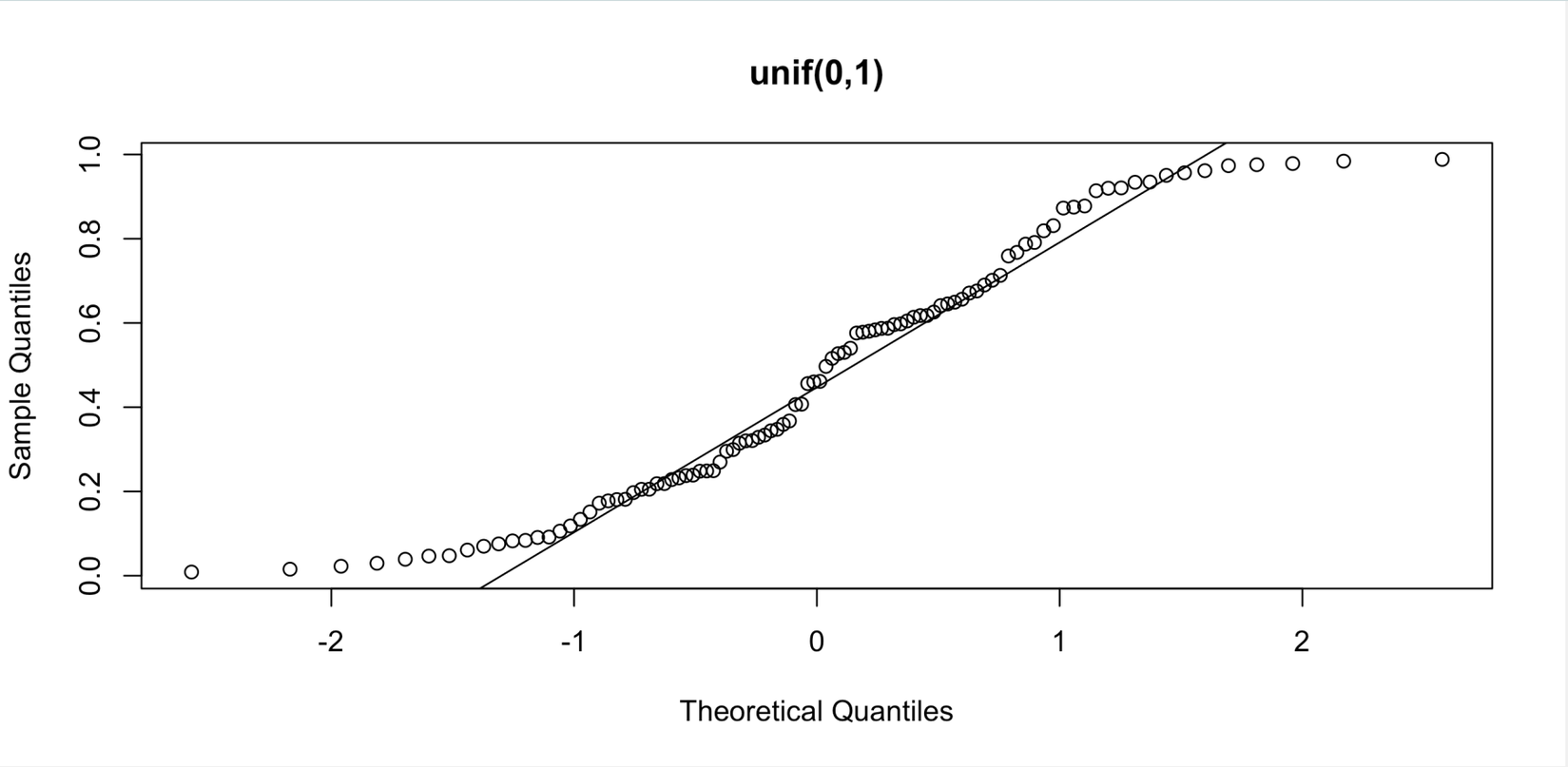
Non-Gaussian data: Q–Q plots (1/2)

```
1  x = rexp(100,1/10);qqnorm(x,main='exponential mu=10');qqline(x)
```



Non-Gaussian data: Q–Q plots (2/2)

```
1  x = runif(100,0,1);qqnorm(x,main='unif(0,1)');qqline(x)
```



Mann–Whitney / Wilcoxon rank-sum

Compares two independent samples. Its general null concerns distributions; a median-shift interpretation needs additional shape assumptions.

Wilcoxon signed-rank

Uses paired differences and assumes a symmetric distribution of those differences for its usual location interpretation.

These tests can be used with small samples, but discreteness, ties, and the available exact or approximate method affect inference.

Chi-squared goodness-of-fit test (1/2)

- Allows a statistical test for categorical data
- χ^2-test:

$$\chi^2 = \sum_{i=1}^{n} \frac{(f_i - e_i)^2}{e_i}$$

- Assumption (Cochran's rule): every expected count is at least 1, and no more than 20% of the cells have expected counts below 5.

```r
freqs <- c(22,21,22,27,22,36)
 probs <- rep(1/6,6)
 chisq.test(freqs,p=probs)
```

```
    Chi-squared test for given probabilities

data:  freqs
X-squared = 6.72, df = 5, p-value = 0.2423
```

Chi-squared goodness-of-fit test (2/2)


```r
freqs <- c(22,31,12,37,12,36)
 probs <- rep(1/6,6)
 chisq.test(freqs,p=probs)
```

```
    Chi-squared test for given probabilities

data:  freqs
X-squared = 25.92, df = 5, p-value = 9.248e-05
```

Matching a question to a statistical method (1/2)

Question	Common starting point
One mean versus a reference	One-sample t-test
Two independent means	Welch t-test
Two paired means	Paired t-test
More than two independent means	ANOVA or Welch ANOVA
Association between categorical variables	Chi-square or Fisher’s exact test

Matching a question to a statistical method (2/2)

Question	Common starting point
Linear association	Pearson correlation
Monotonic association	Spearman correlation
Continuous response with predictors	Linear regression
Binary response with predictors	Logistic regression
Time-to-event response	Survival analysis

Choose according to the design and assumptions. Repeated measures, censoring, confounding, and multiple testing require explicit treatment.

Why multiple testing matters

Suppose we run **100,000 valid tests**, all with true null hypotheses.

With uniform null p-values and a threshold of 0.05, the expected number of false positives is:

$$100{,}000 \times 0.05 = 5{,}000.$$

This is an expected count, not the exact realised count or the probability of at least one false positive.

Multiple-testing adjustments (1/2)

Let $p_{(1)} \leq \cdots \leq p_{(m)}$ be the sorted p-values for the **full family of tests**.

Bonferroni

$$p_i^{\text{adj}} = \min(1, mp_i)$$

Controls the family-wise error rate.

Holm

$$p_{(i)}^{\text{adj}} = \min\left(1, \max_{j \leq i}(m - j + 1)p_{(j)}\right)$$

Controls the family-wise error rate using a step-down procedure.

Benjamini–Hochberg

$$p_{(i)}^{\text{adj}} = \min\left(1, \min_{j \geq i}\frac{m}{j}p_{(j)}\right)$$

Multiple-testing adjustments (2/2)

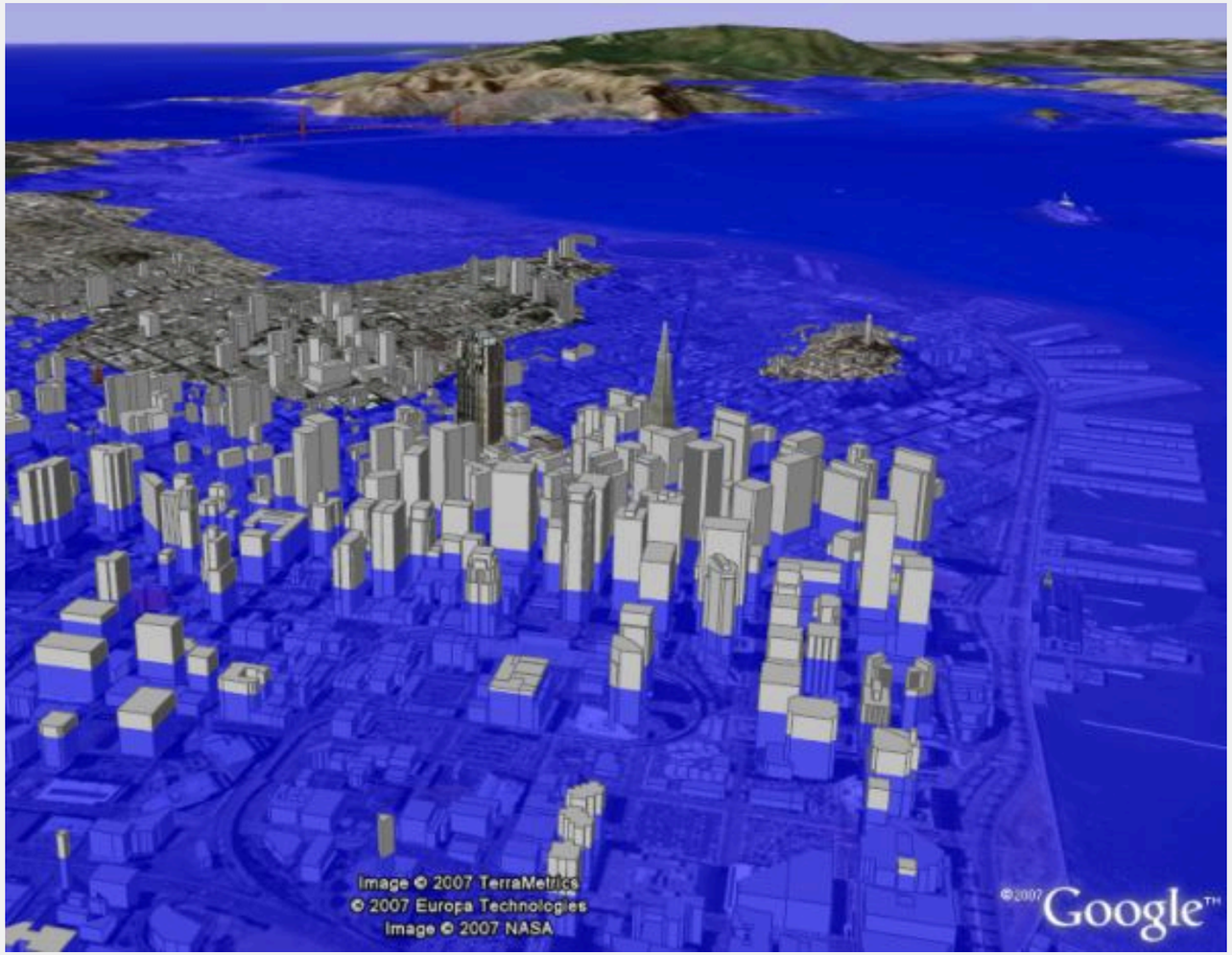
Controls the false discovery rate under independence or appropriate positive dependence.
Restore the original test order after adjustment.

Adjustments in R

```r
p <- c(0.01, 0.04, 0.03, 0.20)
cbind(p, Bonferroni=p.adjust(p, "bonferroni"),
      Holm=p.adjust(p, "holm"), BH=p.adjust(p, "BH"))
```

```
      p Bonferroni Holm         BH
[1,] 0.01       0.04 0.04 0.04000000
[2,] 0.04       0.16 0.09 0.05333333
[3,] 0.03       0.12 0.09 0.05333333
[4,] 0.20       0.80 0.20 0.20000000
```

Signal to noise?



San Francisco under a flooded-sea scenario, visualised in Google Earth. As a metaphor: the “signal” (the city skyline) remains identifiable, but the surrounding “noise” (the raised sea) drowns everything at the same scale. When many hypotheses are tested, the noise floor rises with the number of tests — and a “significant” gene is a building that seems to rise above water, with no guarantee that it truly does.

Enrichment techniques

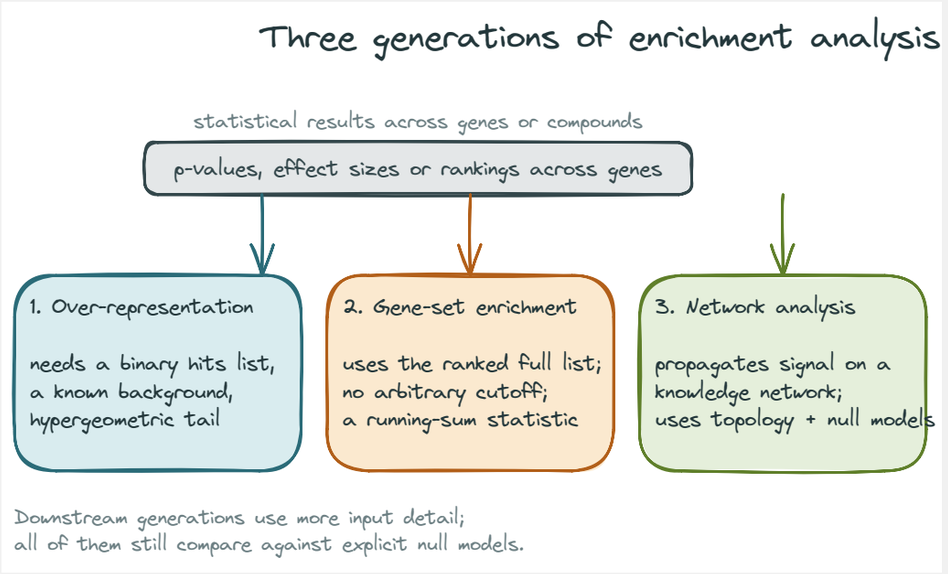
Experimental data can be understood in the framework of the **known biological context**. Enrichment techniques relate this data to **biological pathways**.

Enrichment in gene expression

> The challenge no longer lies in obtaining gene expression profiles, but rather in interpreting the results to gain insights into biological mechanisms.

(Subramanian et al PNAS 2005)

Three generations appear on later slides: **over-representation** (binary hits list, known background), **gene-set enrichment** (ranked full list, no cutoff), **network analysis** (signal propagated over a knowledge network).



Genes and gene products

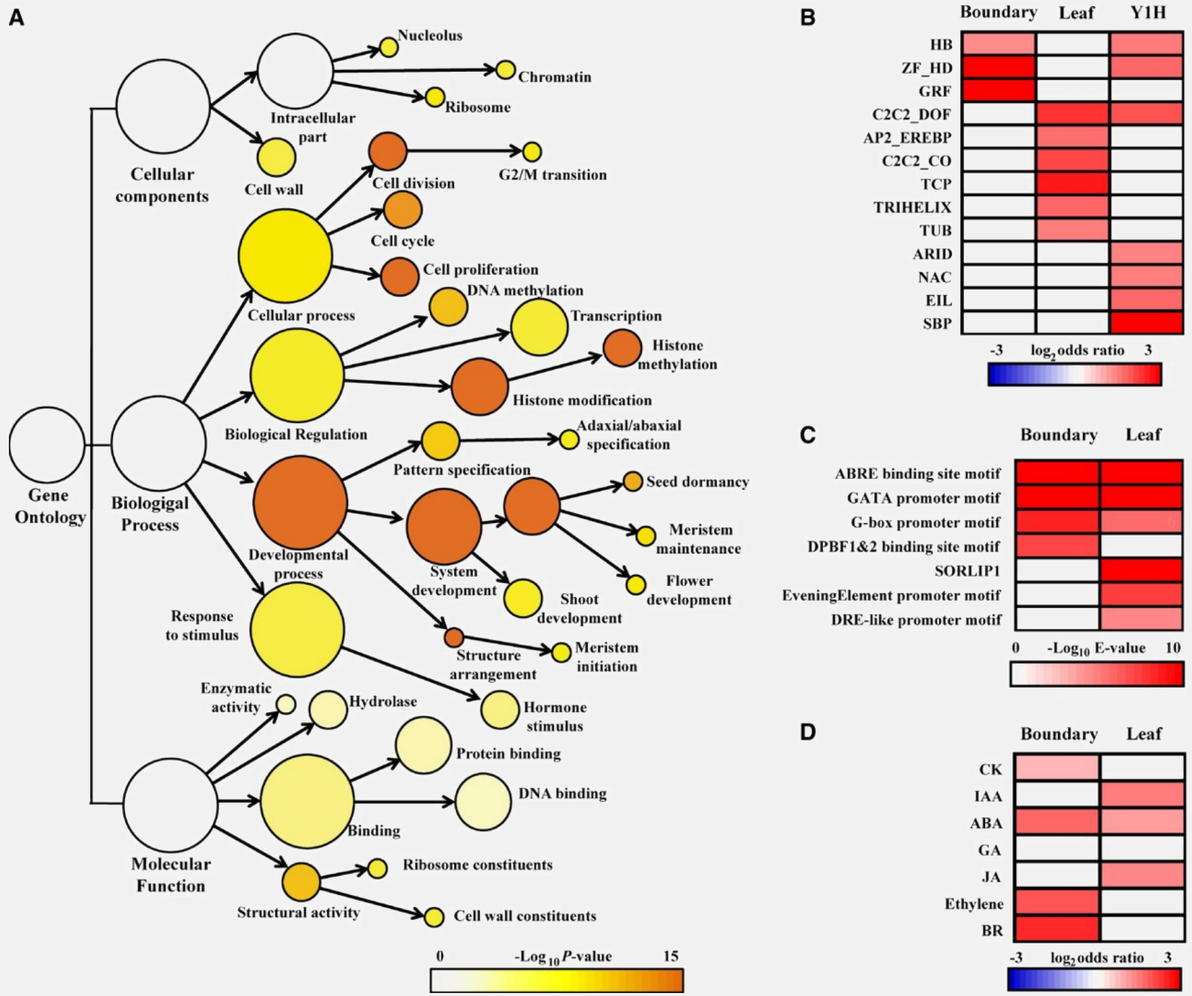
- **Gene Ontology**: contains three ontologies and a hierarchical structure reflecting the specificity of the entry (The Gene Ontology Consortium, NAR 2008).

Proteomics

- **UniProt**: comprehensive database for protein sequence and annotation data (The UniProt Consortium, NAR 2018).
- **STRING**: protein association database, including functional and physical associations; automatic and curated evidence (von Mering et al, NAR 2005).

Metabolomics

- **KEGG**: comprehensive and curated pathway database (Kanehisa, NAR 2000).
- **Reactome, BioCyc, WikiPathways, SMPDB**



An example of GO overrepresentation analysis (ORA) (Tian 2014).

Define a background of N genes that could have been selected. A pathway contains K of these genes; the selected list has n genes, with overlap k.

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}.$$

For enrichment, use the **upper-tail** probability:

$$p = P(X \geq k) = \sum_{j=k}^{\min(K,n)} P(X = j).$$

Adjust across the tested gene sets. The background and selection process determine the null model.

Gene data enrichment: over representation

Advantages

- Relates genes with biological processes.
- Assesses statistical significance.
- Eases exploratory analysis.

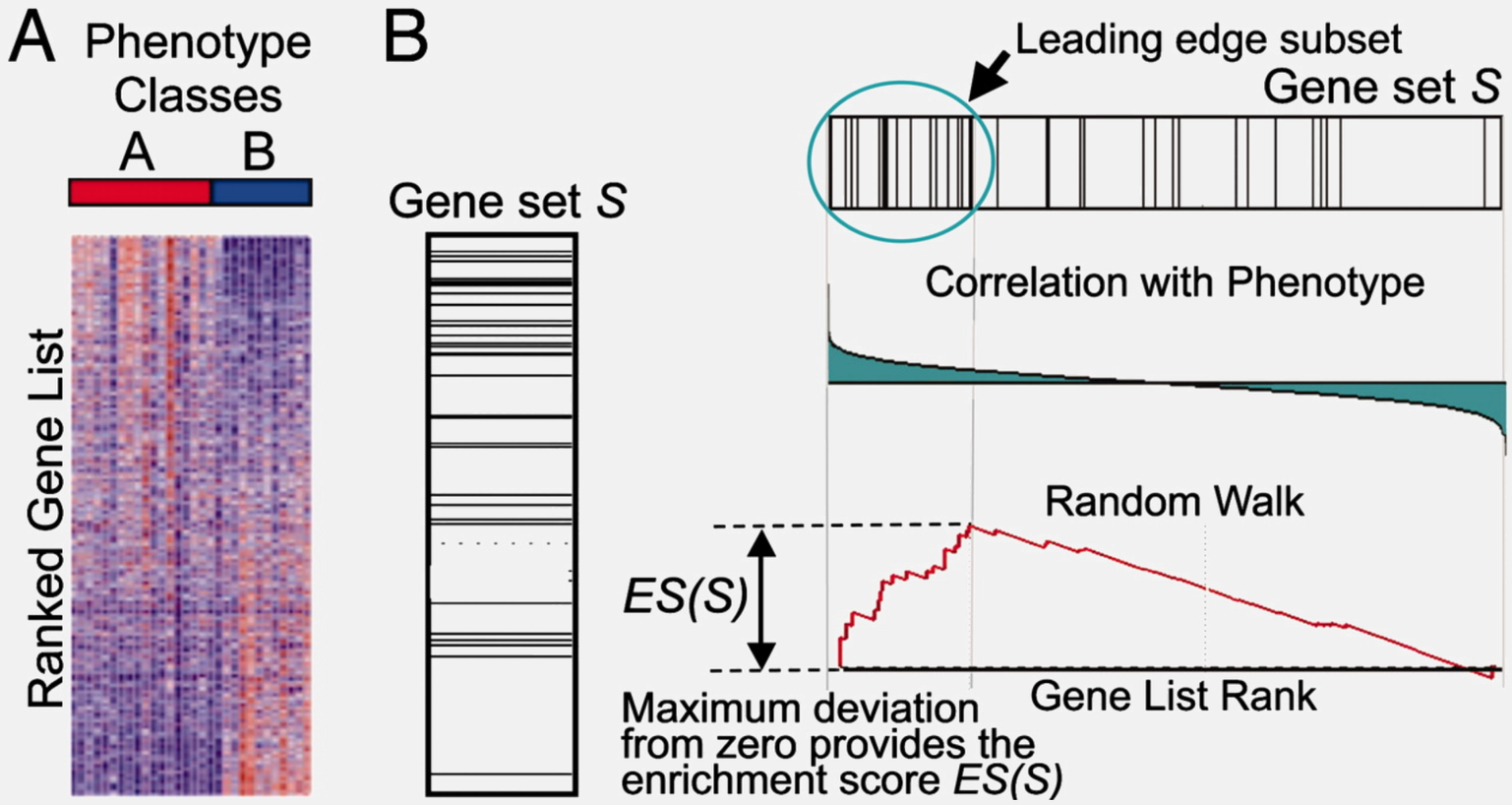
Limitations

- Relies on a prior (arbitrary) cutoff method.
- Sensitive on the input size.
- Ignores magnitude of the gene affectation.
- Simplistic context representation in a binary fashion.

Motivation

The second generation of methods intend to take advantage of **quantitative** data about gene expression.

Gene Set Enrichment Analysis



GSEA procedure from original publication.

Gene data enrichment: set enrichment analysis (II)

Advantages

- Takes into account quantitative measures.
- Has reported to capture subtle but consistent changes in expression data.

Limitations

- Genes have complex interactions that could be accounted.
- Biological pathways exhibit substantial overlapping.

Motivation

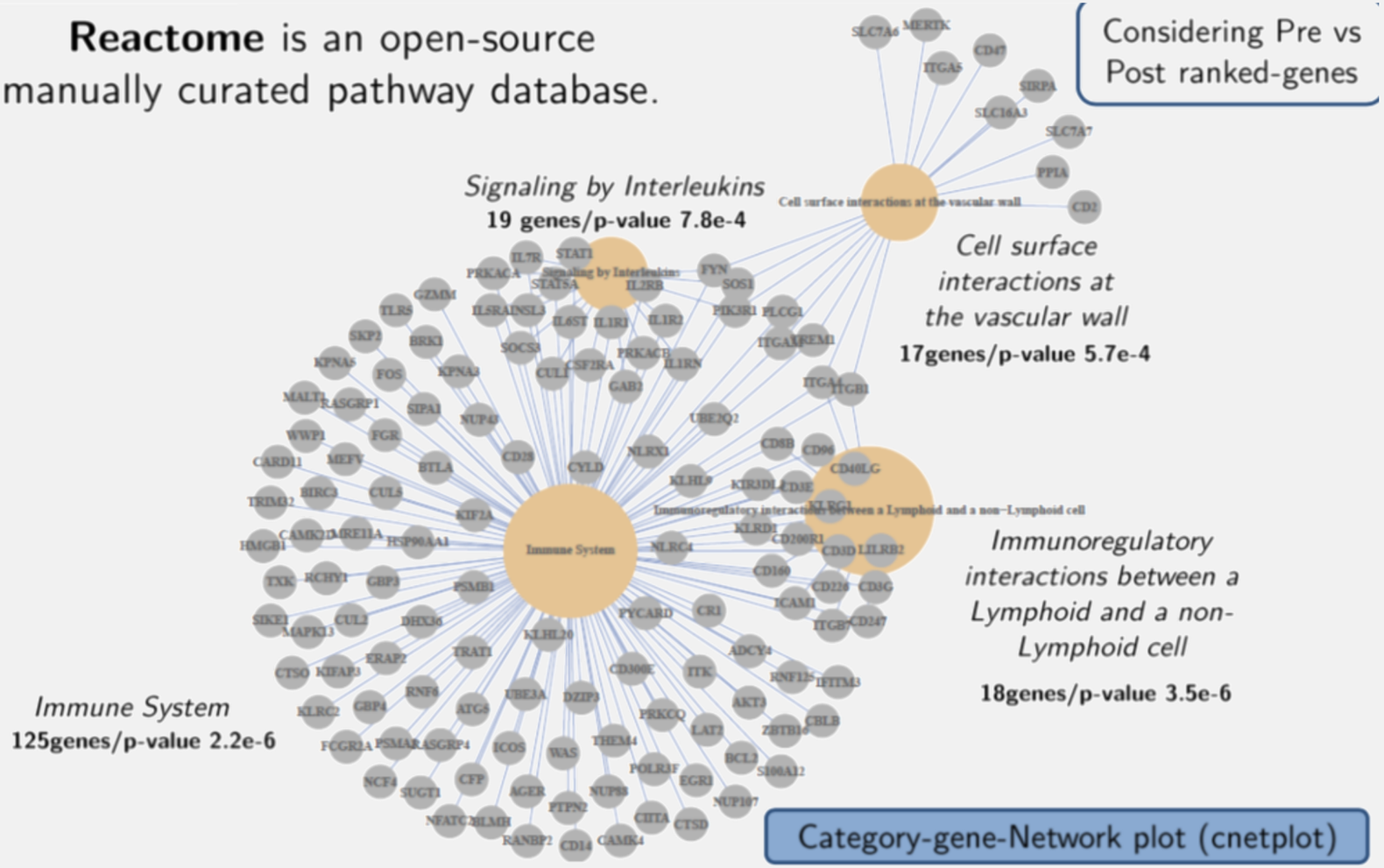
Network-based methods take into account the **topological features** of biological networks in order to test pathways.

Pathway-Express Amin 2007

$$IF(P) = \log \frac{1}{p} + \frac{\sum_{g \in P} |PF(g)|}{|\bar{\Delta E}| \cdot N_{de}(P)}$$

- $\log \frac{1}{p}$ incorporates a p-value from over-representation analysis in the original Pathway-Express formulation
- $PF(g) = \Delta E(g) + \sum_{u \in US_g} \beta_{ug} \cdot \frac{PF(u)}{N_{ds}(u)}$ considers the topology of the genes within the pathway, and is similar to PageRank scores.

Results Sample



Interactive result service (historical URL, requires campus network):
http://eko.upc.edu:7777/sergi/FELLA