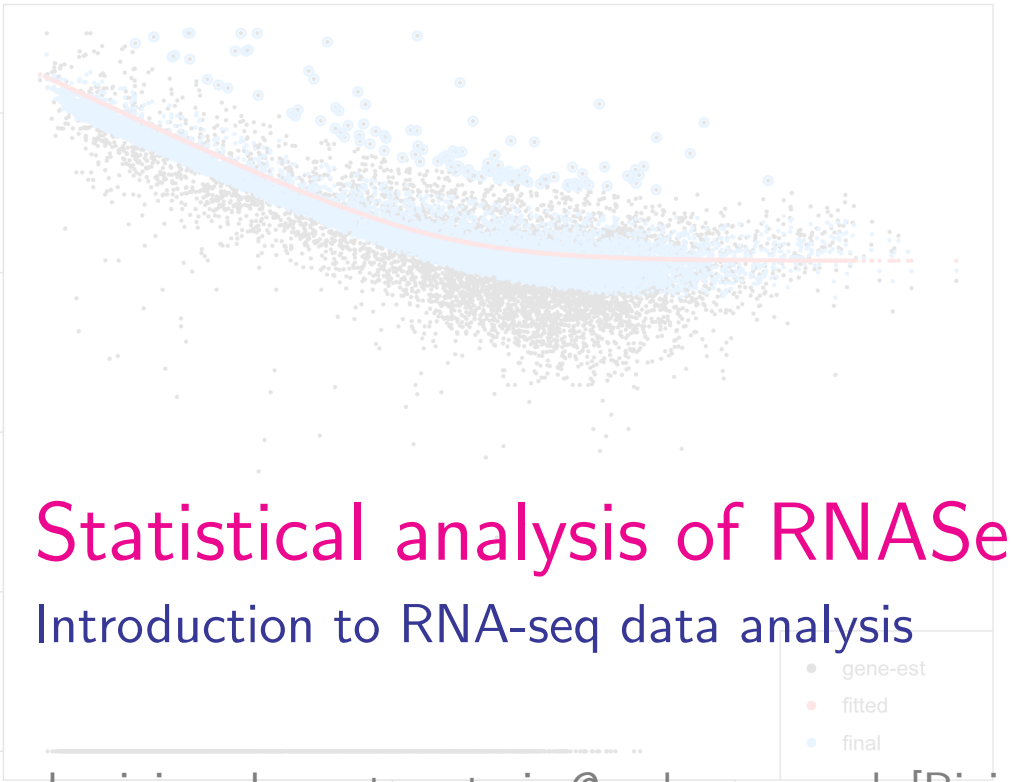


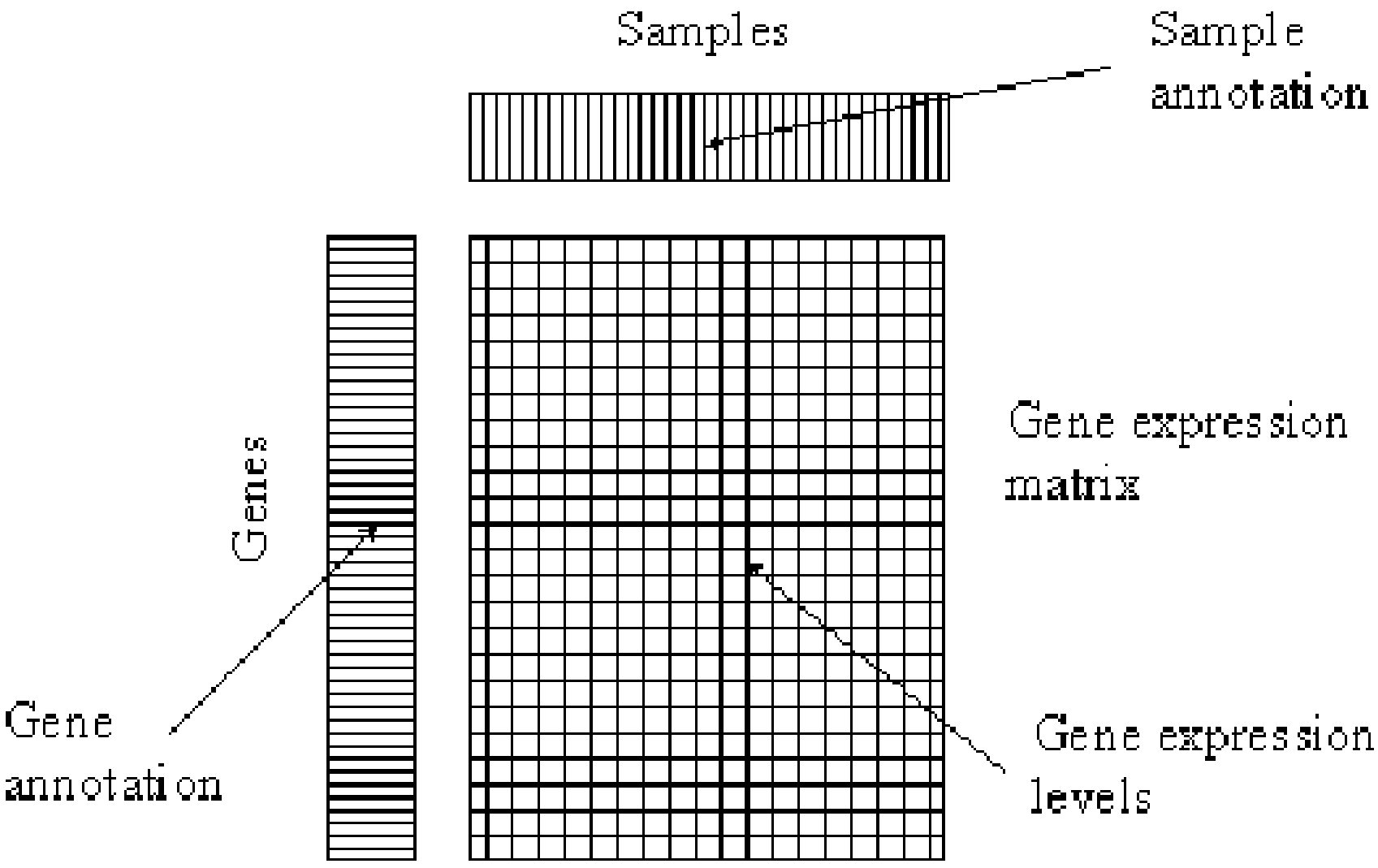
Statistical analysis of RNASeq Data

Introduction to RNA-seq data analysis

dominique-laurent.couturier@cruk.cam.ac.uk [Bioinformatics core]

(Source: O. Rueda, CRUK-CI; G. Marot, INRIA)

Introduction

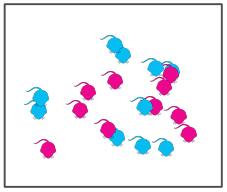


Grand Picture of Statistics

Statistical Hypotheses

H0: $\mu_B = \mu_L$

H1: $\mu_B \neq \mu_L$



Sample

Data: RNASeq counts

$(x_{B,1}; x_{B,2}; ...; x_{B,n_B})$

$(x_{L,1}; x_{L,2}; ...; x_{L,n_L})$

Point estimation

$\widehat{\mu}_B - \widehat{\mu}_L$

Inference

$$T_{obs} = \frac{\widehat{\mu}_B - \widehat{\mu}_L}{s_p \sqrt{\frac{1}{n_B} + \frac{1}{n_L}}} \sim St_{n_T + n_C - 2}$$

Idea:

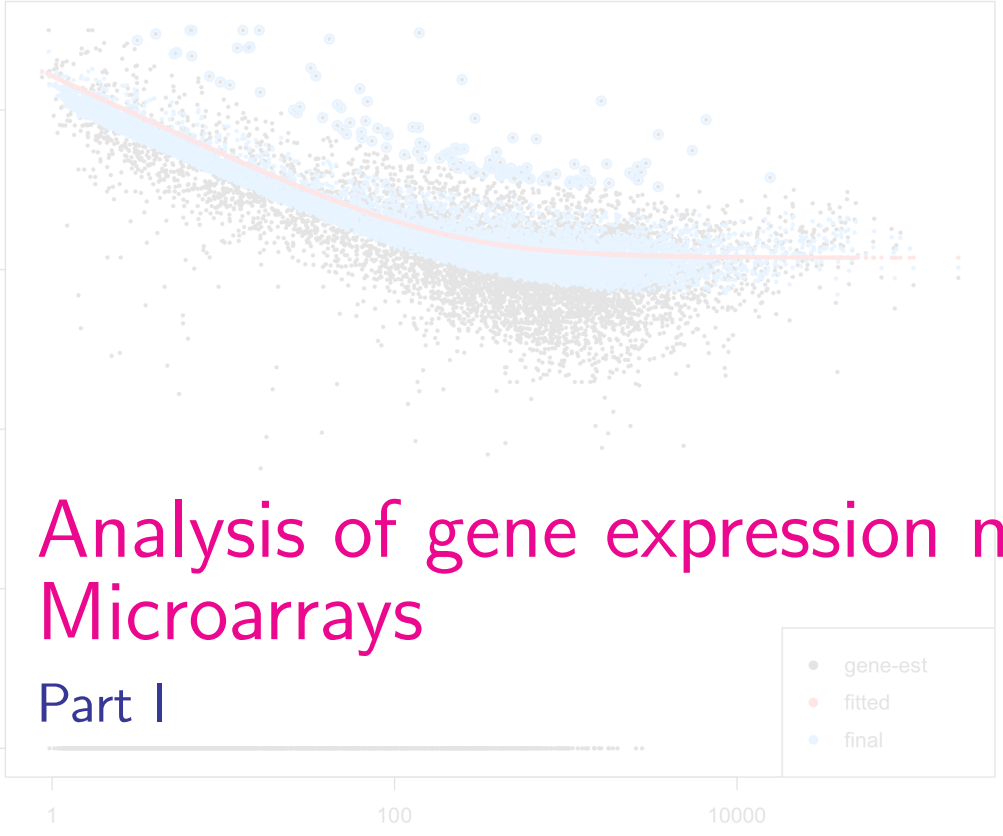
EGF is differentially expressed (DE) in luminal (L) and basal (B) cells

Outline

- 1/ Analysis of gene expression measured with Microarrays
  - 1a/ Normal distribution
  - 1b/ Test of equality of means for two samples: T-test
  - 1c/ Test of equality of means for > 2 samples: ANOVA
  - 1d/ Test of equality of means for 2 categorical predictors: ANOVA
  - 1e/ Test of equality of means for > 2 predictors: Linear model
  - 1f/ Confounding
- 2/ Analysis of gene expression measured by RNAseq
  - Generalisation of the linear model: Negative Binomial regression
    - 2a/ Negative Binomial distribution
    - 2b/ Nuisance parameter estimation: Shrinkage estimator
    - 2c/ Controlling for Library size: Offset
- 3/ Controlling for multiple testing
  - 3a/ Family-wise error rate
  - 3b/ False discovery rate

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

UNIVERSITY OF CAMBRIDGE



Analysis of gene expression measured with Microarrays

Part I

dominique-laurent.couturier@cruk.cam.ac.uk [Bioinformatics core]

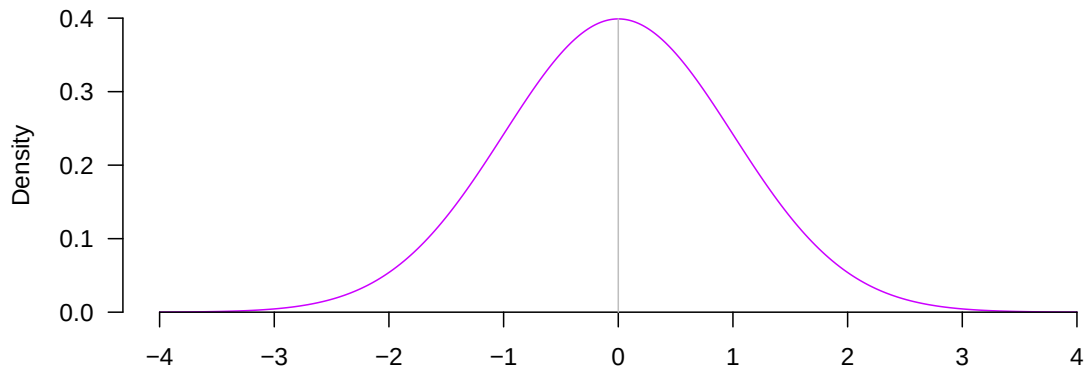
(Source: O. Rueda, CRUK-CI; G. Marot, INRIA)

1a/ Normal distribution

$$X \sim N(\mu, \sigma^2), \qquad f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\mathsf{E}[X] = \mu, \qquad \mathsf{Var}[X] = \sigma^2,$$

Probability density function, $f_X(x|\mu = 0, \sigma = 1)$

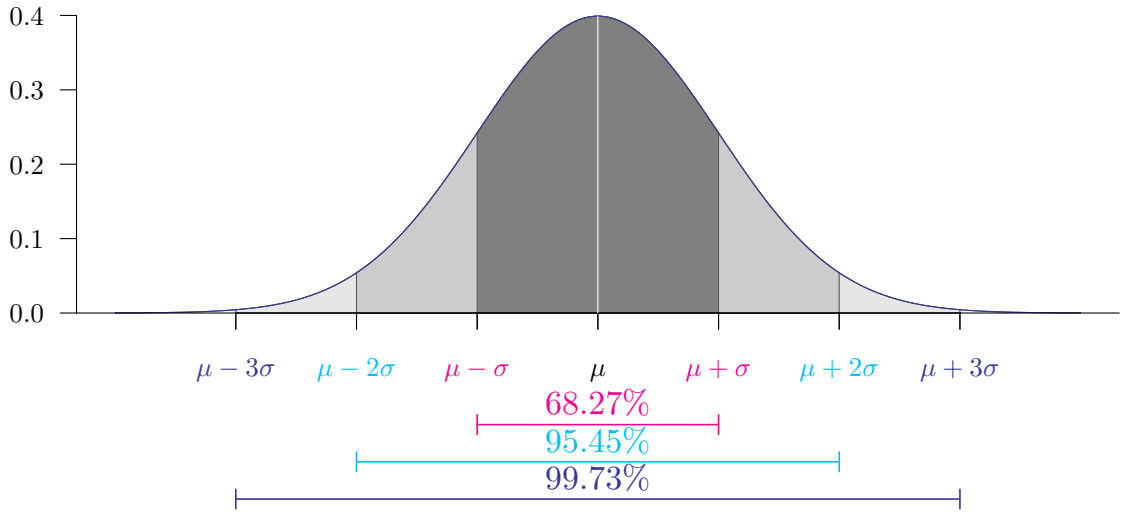


1a/ Normal distribution

$$X \sim N(\mu, \sigma^2), \qquad f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\mathsf{E}[X] = \mu, \qquad \mathsf{Var}[X] = \sigma^2,$$

Probability density function

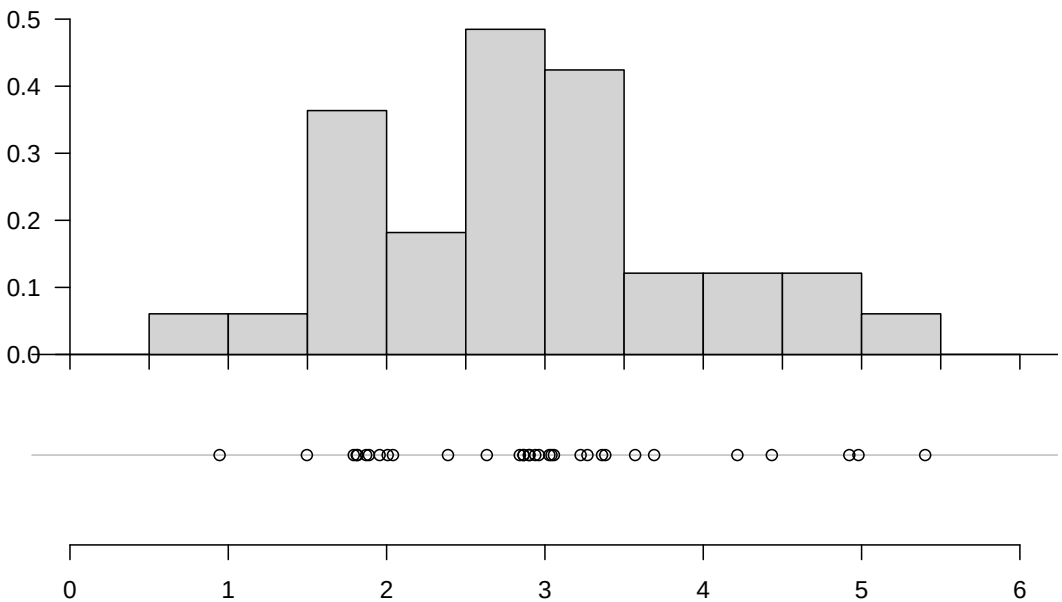


1a/ Normal distribution

$$X \sim N(\mu, \sigma^2), \qquad f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\mathsf{E}[X] = \mu, \qquad \mathsf{Var}[X] = \sigma^2,$$

- Suitable modelling for a lot of variables



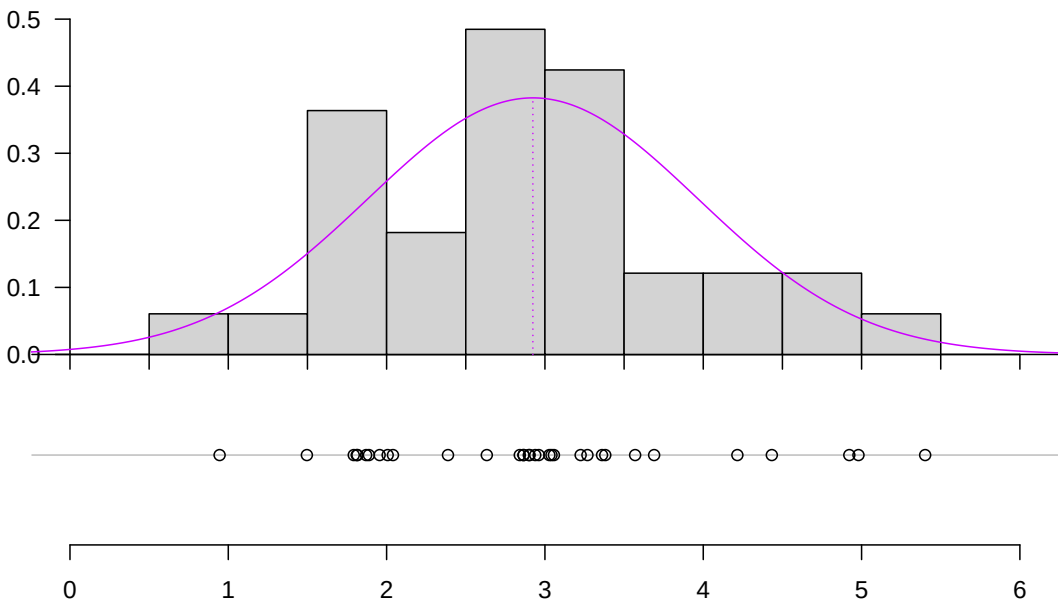
(Gene expression values of gene 'X' of basal cells of 33 mice)

1a/ Normal distribution

$$X \sim N(\mu, \sigma^2), \qquad f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

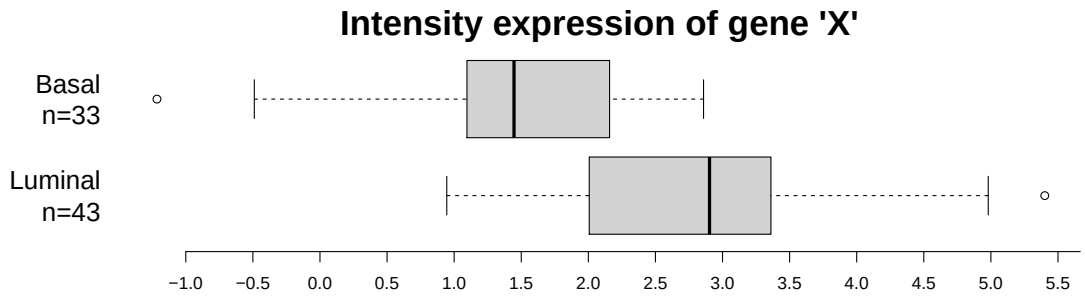
$$\mathsf{E}[X] = \mu, \qquad \mathsf{Var}[X] = \sigma^2,$$

- Suitable modelling for a lot of variables



(Gene expression values of gene 'X' of basal cells of 33 mice)

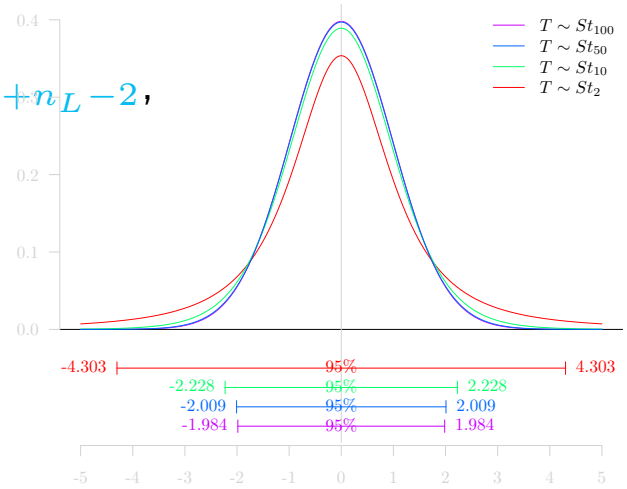
1b/ Test of equality of means for two samples



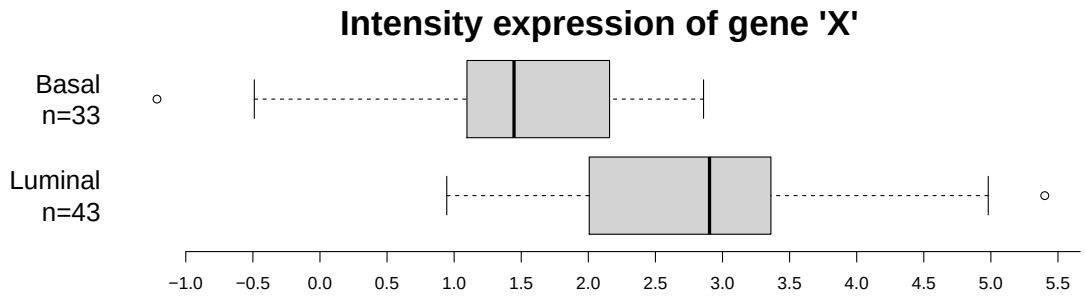
We test **H0**: $\mu_B - \mu_L = 0$ against **H1**: $\mu_B - \mu_L \neq 0$.

We know:

- Student's t-test [assume $\sigma_B^2 = \sigma_L^2$]: $\dfrac{\widehat{\mu}_B - \widehat{\mu}_L}{s_p \sqrt{\frac{1}{n_B} + \frac{1}{n_L}}} \sim t_{n_B + n_L - 2}$,
- $s_p = \sqrt{\dfrac{s_B^2 (n_B - 1) + s_L^2 (n_L - 1)}{n_B + N_L - 2}}$.



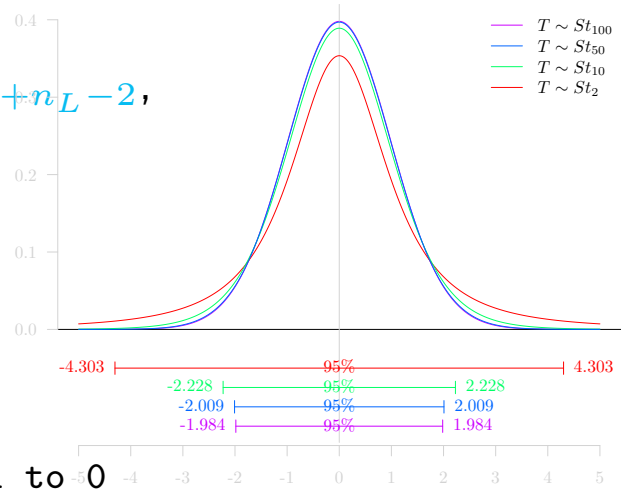
1b/ Test of equality of means for two samples



We test **H0**: $\mu_B - \mu_L = 0$ against **H1**: $\mu_B - \mu_L \neq 0$.

We know:

- Student's t-test [assume $\sigma_B^2 = \sigma_L^2$]: $\dfrac{\widehat{\mu}_B - \widehat{\mu}_L}{s_p \sqrt{\frac{1}{n_B} + \frac{1}{n_L}}} \sim t_{n_B + n_L - 2}$,
- $s_p = \sqrt{\dfrac{s_B^2 (n_B - 1) + s_L^2 (n_L - 1)}{n_B + N_L - 2}}$.

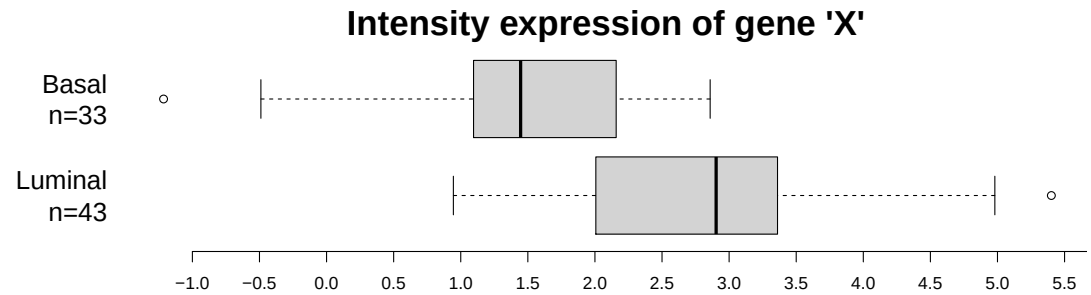



```
Two Sample t-test

data:  Basal and Luminal
t = 6.6751, df = 74, p-value = 3.941e-09
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.048457 1.940748
sample estimates:
mean of x mean of y
 2.923908  1.429305
```

1b/ Test of equality of means for two samples

- Modelling 1:

$$Y_{i(B)} = \mu_B + \epsilon_i$$
$$Y_{i(L)} = \mu_L + \epsilon_i$$



1b/ Test of equality of means for two samples

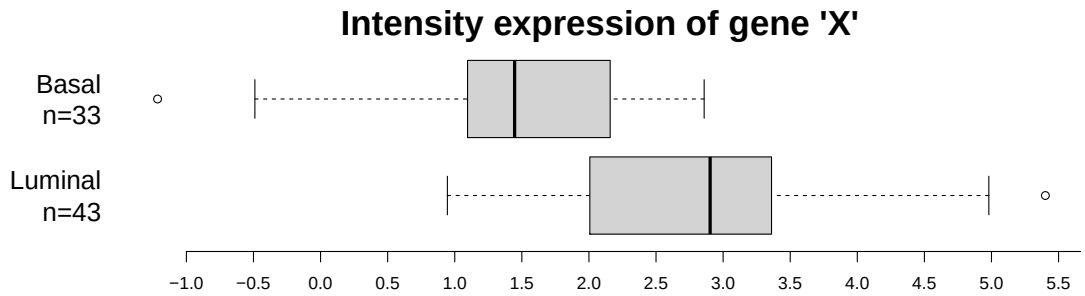
- Modelling 1:

$$Y_{i(B)} = \mu_B + \epsilon_i$$
$$Y_{i(L)} = \mu_L + \epsilon_i$$

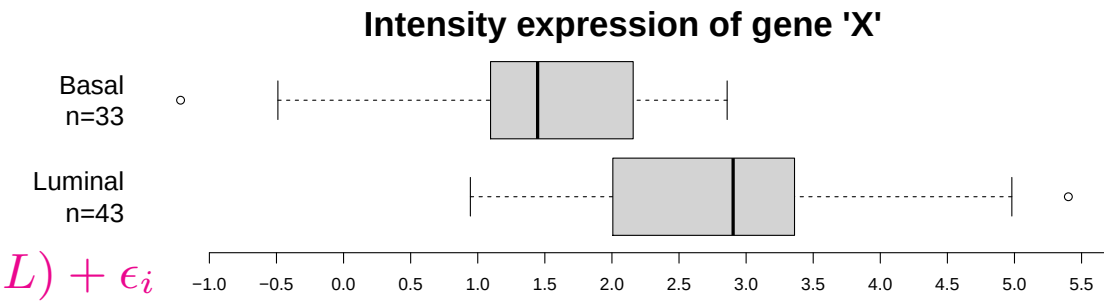
- Modelling 2:

$$Y_i = \mu_B + \delta_L \; I(i \in L)\epsilon_i$$
$$= \beta_0 + \beta_1 X_1 + \epsilon_i$$

where $i = 1, ..., n$; $\epsilon_i \sim N(0, \sigma^2)$.



1b/ Test of equality of means for two samples

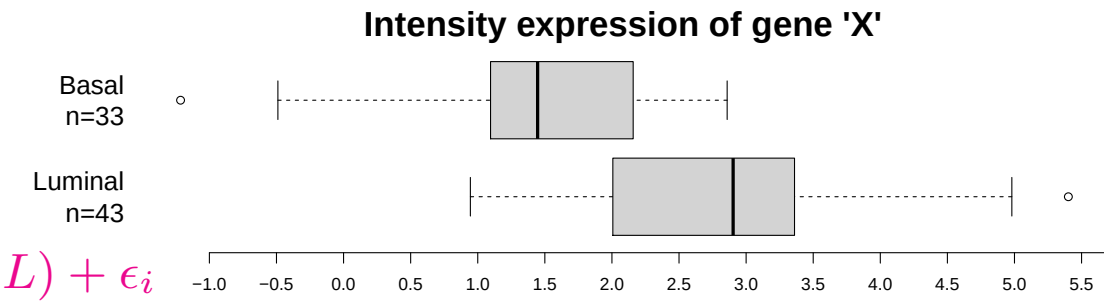


- Modelling 1:

$$Y_i = \mu_B I(i \in B) + \mu_L I(i \in L) + \epsilon_i$$
$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$$

where $i = 1, ..., n$; $\epsilon_i \sim N(0, \sigma^2)$.

1b/ Test of equality of means for two samples



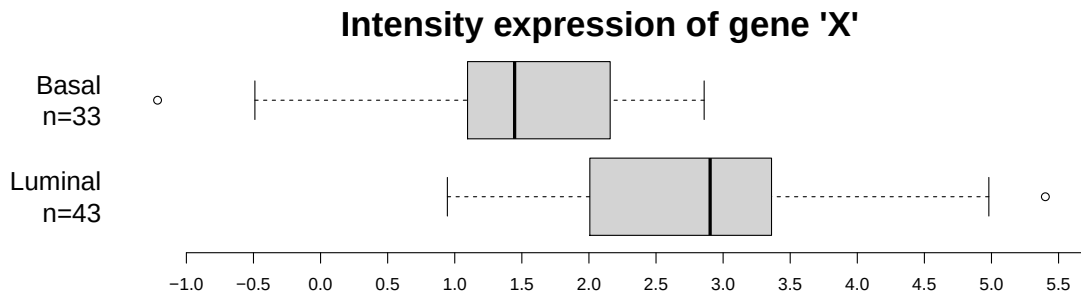
- Modelling 1:

$$Y_i = \mu_B I(i \in B) + \mu_L I(i \in L) + \epsilon_i$$
$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$$

where $i = 1, ..., n$; $\epsilon_i \sim N(0, \sigma^2)$.

```
Call:
lm(formula = expression ~ celltype - 1, data = microarrays)

Residuals:
     Min       1Q   Median       3Q      Max
-2.64401 -0.58586  0.01473  0.65051  2.47771

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
celltypeBasal     2.9239     0.1684  17.361  < 2e-16 ***
celltypeLuminal   1.4293     0.1475   9.687 8.47e-15 ***
---
Signif. codes:  0 ‚Äò***‚Äô 0.001 ‚Äò**‚Äô 0.01 ‚Äò*‚Äô 0.05 ‚Äò.‚Äô 0.1 ‚Äò ‚Äô 1

Residual standard error: 0.9675 on 74 degrees of freedom
Multiple R-squared:  0.8423,Adjusted R-squared:  0.838
F-statistic: 197.6 on 2 and 74 DF,  p-value: < 2.2e-16
```

1b/ Test of equality of means for two samples

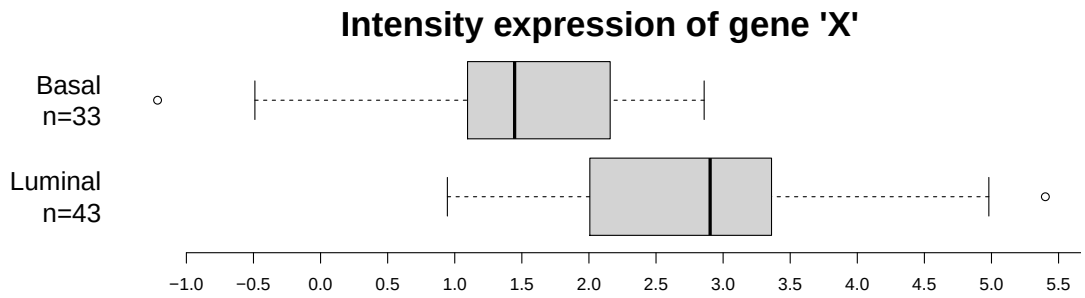


- Modelling 2:

$$Y_i = \mu_B + \delta_L \; I(i \in L)\epsilon_i$$
$$= \beta_0 + \beta_1 X_1 + \epsilon_i$$
$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$$

where $i = 1, ..., n$; $\epsilon_i \sim N(0, \sigma^2)$.

1b/ Test of equality of means for two samples



- Modelling 2:

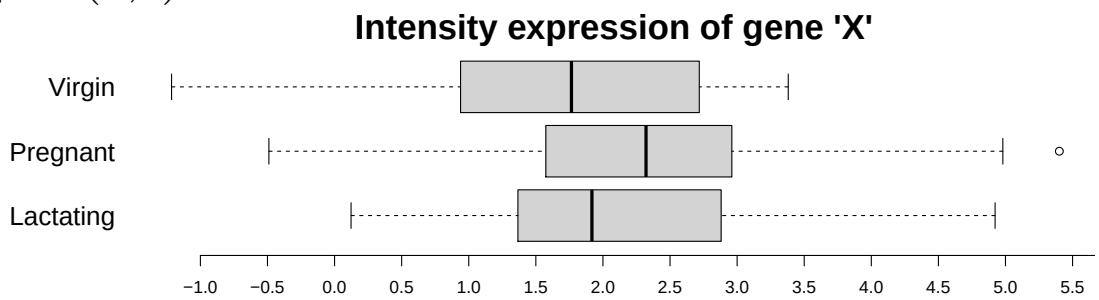
$$Y_i = \mu_B + \delta_L \; I(i \in L)\epsilon_i$$
$$= \beta_0 + \beta_1 X_1 + \epsilon_i$$
$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$$

where $i = 1, ..., n$; $\epsilon_i \sim N(0, \sigma^2)$.

```
Call:
lm(formula = expression ~ celltype, data = microarrays)

Residuals:
     Min       1Q   Median       3Q      Max
-2.64401 -0.58586  0.01473  0.65051  2.47771

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.9239     0.1684  17.361  < 2e-16 ***
celltypeLuminal  -1.4946     0.2239  -6.675 3.94e-09 ***
---
Signif. codes:  0 ‚Äò***‚Äô 0.001 ‚Äò**‚Äô 0.01 ‚Äò*‚Äô 0.05 ‚Äò.‚Äô 0.1 ‚Äò ‚Äô 1

Residual standard error: 0.9675 on 74 degrees of freedom
Multiple R-squared:  0.3758,Adjusted R-squared:  0.3674
F-statistic: 44.56 on 1 and 74 DF,  p-value: 3.941e-09
```

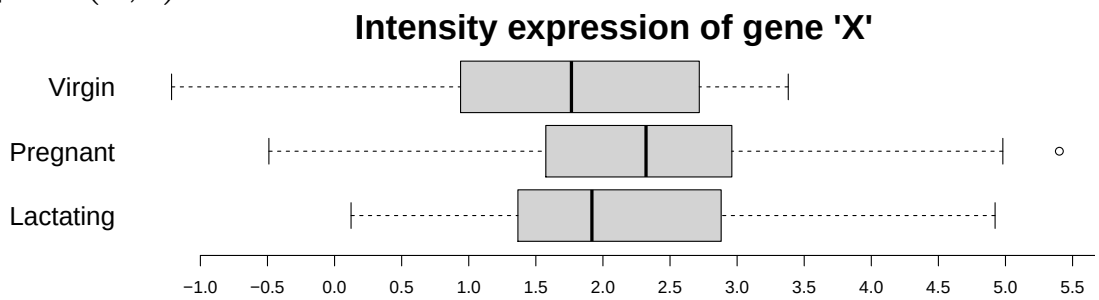
1c/ Test of equality of means for > 2 samples

- One-way ANOVA hypotheses
  - **H0:** $\mu_L = \mu_P = \mu_V$,
  - **H1:** $\mu_k \neq \mu_l$ for at least one pair (k, l).



1c/ Test of equality of means for > 2 samples

- One-way ANOVA hypotheses
  - **H0:** $\mu_L = \mu_P = \mu_V$,
  - **H1:** $\mu_k \neq \mu_l$ for at least one pair (k, l).



- Modelling 1:

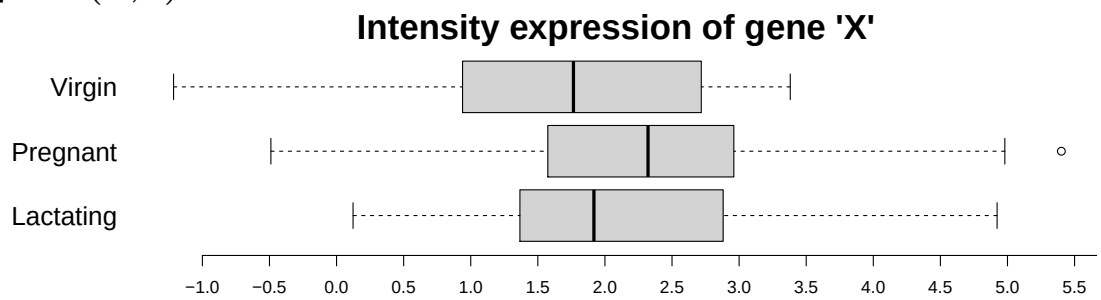
$$Y_{i(L)} = \mu_L + \epsilon_i$$
$$Y_{i(P)} = \mu_P + \epsilon_i$$
$$Y_{i(V)} = \mu_V + \epsilon_i$$
$$Y_i = \mu_L I(i \in L) + \mu_P I(i \in P) + \mu_V I(i \in V) + \epsilon_i$$
$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$$

1c/ Test of equality of means for > 2 samples

- One-way ANOVA hypotheses
  - **H0:** $\mu_L = \mu_P = \mu_V$,
  - **H1:** $\mu_k \neq \mu_l$ for at least one pair (k, l).

- Modelling 1:

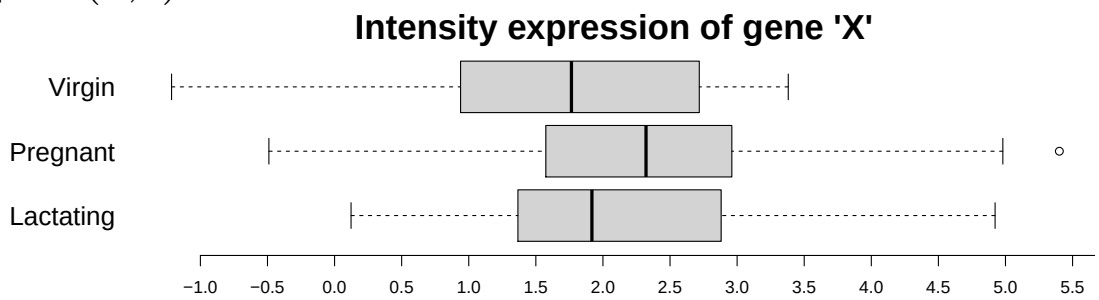
$$Y_{i(L)} = \mu_L + \epsilon_i$$
$$Y_{i(P)} = \mu_P + \epsilon_i$$
$$Y_{i(V)} = \mu_V + \epsilon_i$$
$$Y_i = \mu_L I(i \in L) + \mu_P I(i \in P) + \mu_V I(i \in V) + \epsilon_i$$
$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$$




```
          Df Sum Sq Mean Sq F value Pr(>F)
mousetype  3  334.8  111.61   78.03 <2e-16 ***
Residuals 73  104.4    1.43
---
Signif. codes:  0 ‚Äò***‚Äô 0.001 ‚Äò**‚Äô 0.01 ‚Äò*‚Äô 0.05 ‚Äò.‚Äô 0.1 ‚Äò ‚Äô 1

Call:
lm(formula = expression ~ mousetype - 1, data = microarrays)

Residuals:
     Min       1Q   Median       3Q      Max
-2.91070 -0.78893 -0.09926  0.80387  2.98027

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
mousetypeLactating   2.1051     0.2302   9.146 9.90e-14 ***
mousetypePregnant    2.4213     0.2392  10.123 1.51e-15 ***
mousetypeVirgin      1.6907     0.2441   6.926 1.43e-09 ***
---
Signif. codes:  0 ‚Äò***‚Äô 0.001 ‚Äò**‚Äô 0.01 ‚Äò*‚Äô 0.05 ‚Äò.‚Äô 0.1 ‚Äò ‚Äô 1

Residual standard error: 1.196 on 73 degrees of freedom
Multiple R-squared:  0.7623,Adjusted R-squared:  0.7525
F-statistic: 78.03 on 3 and 73 DF,  p-value: < 2.2e-16
```

1c/ Test of equality of means for > 2 samples

- One-way ANOVA hypotheses
  - **H0:** $\mu_V = \mu_P = \mu_L$,
  - **H1:** $\mu_k \neq \mu_l$ for at least one pair (k, l).



- Modelling 2:

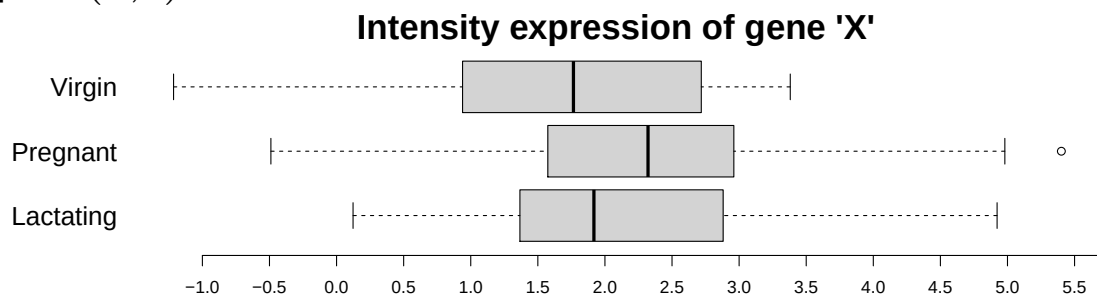
$$Y_{i(L)} = \mu_L + \epsilon_i$$
$$Y_{i(P)} = \mu_L + \delta_P + \epsilon_i$$
$$Y_{i(V)} = \mu_L + \delta_V + \epsilon_i$$
$$Y_i = \mu_L + \delta_P \; I(i \in P) + \delta_V \; I(i \in V) + \epsilon_i$$
$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$$

1c/ Test of equality of means for > 2 samples

- One-way ANOVA hypotheses
  - **H0:** $\mu_V = \mu_P = \mu_L$,
  - **H1:** $\mu_k \neq \mu_l$ for at least one pair (k, l).

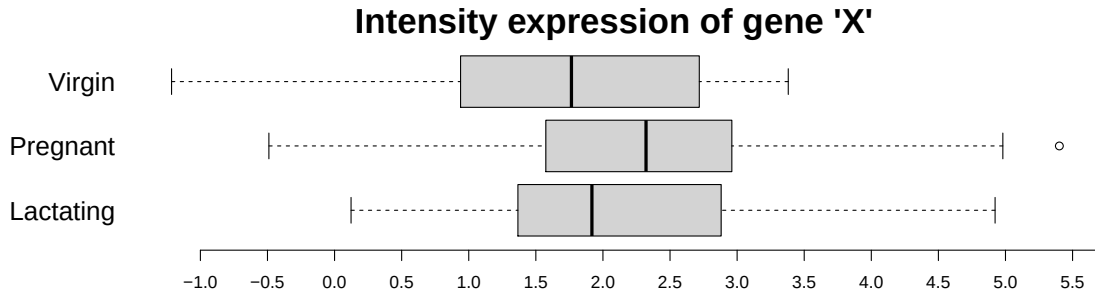
- Modelling 2:

$$Y_{i(L)} = \mu_L + \epsilon_i$$
$$Y_{i(P)} = \mu_L + \delta_P + \epsilon_i$$
$$Y_{i(V)} = \mu_L + \delta_V + \epsilon_i$$
$$Y_i = \mu_L + \delta_P \; I(i \in P) + \delta_V \; I(i \in V) + \epsilon_i$$
$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$$




```
            Df Sum Sq Mean Sq F value Pr(>F)
mousetype    2   6.57   3.283   2.296  0.108
Residuals   73 104.41   1.430

Call:
lm(formula = expression ~ mousetype, data = microarrays)

Residuals:
     Min       1Q   Median       3Q      Max
-2.91070 -0.78893 -0.09926  0.80387  2.98027

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        2.1051     0.2302   9.146  9.9e-14 ***
mousetypePregnant  0.3162     0.3319   0.953    0.344
mousetypeVirgin   -0.4144     0.3355  -1.235    0.221
---
Signif. codes:  0 ‚Äò***‚Äô 0.001 ‚Äò**‚Äô 0.01 ‚Äò*‚Äô 0.05 ‚Äò.‚Äô 0.1 ‚Äò ‚Äô 1

Residual standard error: 1.196 on 73 degrees of freedom
Multiple R-squared:  0.05917,Adjusted R-squared:  0.0334
F-statistic: 2.296 on 2 and 73 DF,  p-value: 0.1079
```

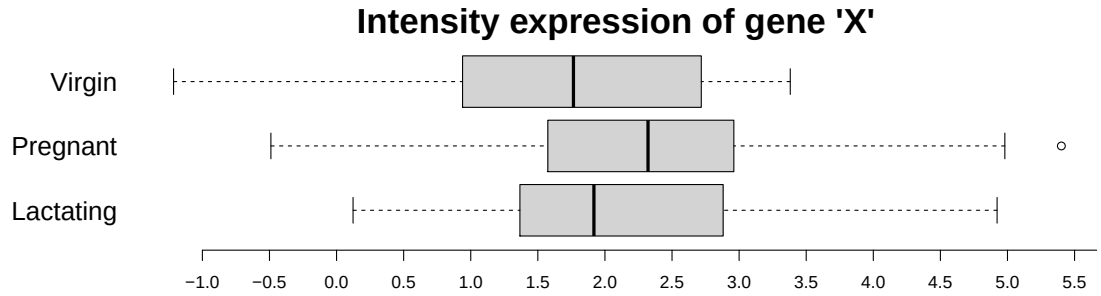
1c/ Test of equality of means for > 2 samples



- One-way ANOVA hypotheses
  - **H0:** $\mu_V = \mu_P = \mu_L$,
  - **H1:** $\mu_k \neq \mu_l$ for at least one pair (k, l).
- Modelling 3:

$$Y_i = \mu + \delta_V \; I(i \in V) + \delta_P \; I(i \in P) + \delta_L \; I(i \in L) + \epsilon_i$$
$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$$

1c/ Test of equality of means for > 2 samples

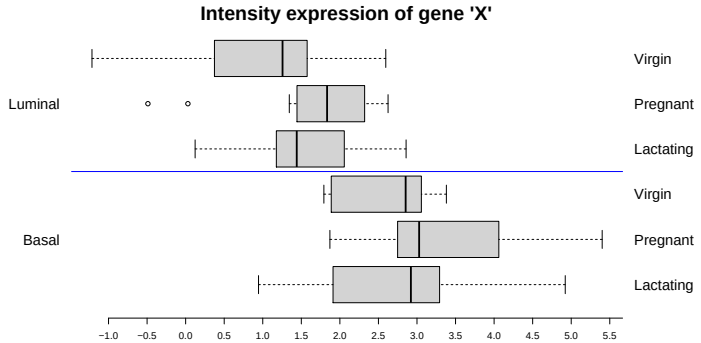


- One-way ANOVA hypotheses
  - **H0:** $\mu_V = \mu_P = \mu_L$,
  - **H1:** $\mu_k \neq \mu_l$ for at least one pair (k, l).
- Modelling 3:

$$Y_i = \mu + \delta_V \; I(i \in V) + \delta_P \; I(i \in P) + \delta_L \; I(i \in L) + \epsilon_i$$
$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$$

```
Call:
lm(formula = expression ~ mousetype.sum, data = microarrays)

Residuals:
     Min       1Q   Median       3Q      Max
-2.91070 -0.78893 -0.09926  0.80387  2.98027

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.07239    0.13735  15.089   <2e-16 ***
mousetype.sum1  0.03272    0.19111   0.171   0.8645
mousetype.sum2  0.34895    0.19477   1.792   0.0773 .
---
Signif. codes:  0 ‚Äò***‚Äô 0.001 ‚Äò**‚Äô 0.01 ‚Äò*‚Äô 0.05 ‚Äò.‚Äô 0.1 ‚Äò ‚Äô 1

Residual standard error: 1.196 on 73 degrees of freedom
Multiple R-squared:  0.05917,Adjusted R-squared:  0.0334
F-statistic: 2.296 on 2 and 73 DF,  p-value: 0.1079
```

1d/ Two-way ANOVA without interaction

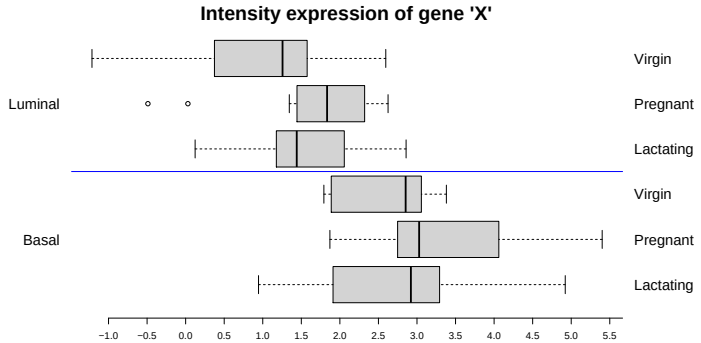


- Linear model with base 'Lactating, Basal'

$$Y_i = \mu_{L,B} + \delta_P \; I(i \in P) + \delta_V \; I(i \in V) + \theta_{L'} I(i \in L') + \epsilon_i$$

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$$

1d/ Two-way ANOVA without interaction



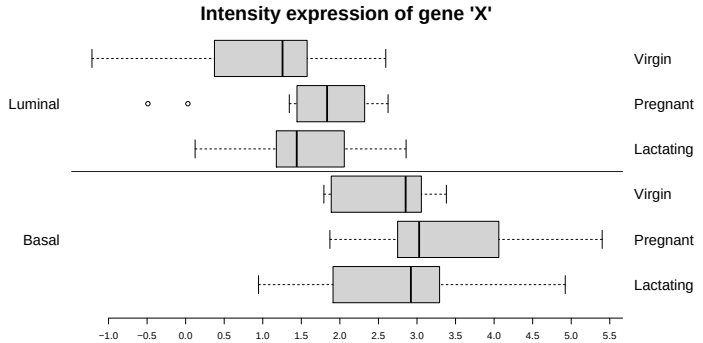
- Linear model with base 'Lactating, Basal'

$$Y_i = \mu_{L,B} + \delta_P \; I(i \in P) + \delta_V \; I(i \in V) + \theta_{L'} I(i \in L') + \epsilon_i$$

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$$

```
            Df Sum Sq Mean Sq F value   Pr(>F)
mousetype    2   6.57    3.28   3.733   0.0287 *
celltype     1  41.08   41.08  46.698 2.24e-09 ***
Residuals   72  63.33    0.88
---
Signif. codes:  0 ‚Äò***‚Äô 0.001 ‚Äò**‚Äô 0.01 ‚Äò*‚Äô 0.05 ‚Äò.‚Äô 0.1 ‚Äò ‚Äô 1

Call:
lm(formula = expression ~ mousetype + celltype, data = microarrays)

Residuals:
     Min       1Q   Median       3Q      Max
-2.28721 -0.47310  0.00495  0.50585  2.14941

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.9294     0.2171  13.494  < 2e-16 ***
mousetypePregnant  0.3228     0.2603   1.240    0.219
mousetypeVirgin   -0.3732     0.2632  -1.418    0.161
celltypeLuminal   -1.4837     0.2171  -6.834 2.24e-09 ***
---
Signif. codes:  0 ‚Äò***‚Äô 0.001 ‚Äò**‚Äô 0.01 ‚Äò*‚Äô 0.05 ‚Äò.‚Äô 0.1 ‚Äò ‚Äô 1

Residual standard error: 0.9379 on 72 degrees of freedom
Multiple R-squared:  0.4293,Adjusted R-squared:  0.4055
F-statistic: 18.05 on 3 and 72 DF,  p-value: 7.754e-09
```

1d/ Two-way ANOVA with interaction



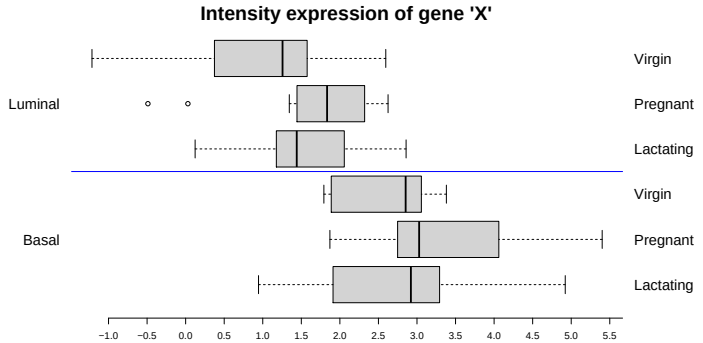
- Linear model with base 'Lactating, Basal'

$$Y_i = \mu_{L,B} + \delta_P\ I(i \in P) + \delta_V\ I(i \in V) + \theta_{L'} I(i \in L') + \epsilon_i$$
$$+ \eta_{PL'} I(i \in P\ \&\ i \in L') + \eta_{VL'} I(i \in V\ \&\ i \in L') + \epsilon_i$$
$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$$

- Hypotheses:

  - **H0**$_1$: $\delta_P = \delta_{L'} = 0$,
  - **H1**$_1$: **H0**$_1$ is false.
  - **H0**$_2$: $\theta_{L'} = 0$,
  - **H1**$_2$: **H0**$_2$ is false.
  - **H0**$_3$: $\eta_{PL'} = \eta_{VL'} = 0$,
  - **H1**$_3$: **H0**$_3$ is false.

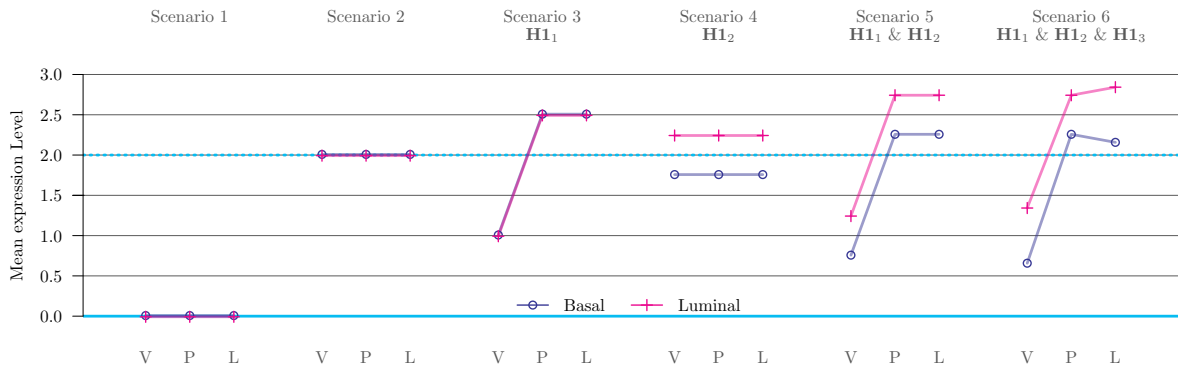
1d/ Two-way ANOVA with interaction



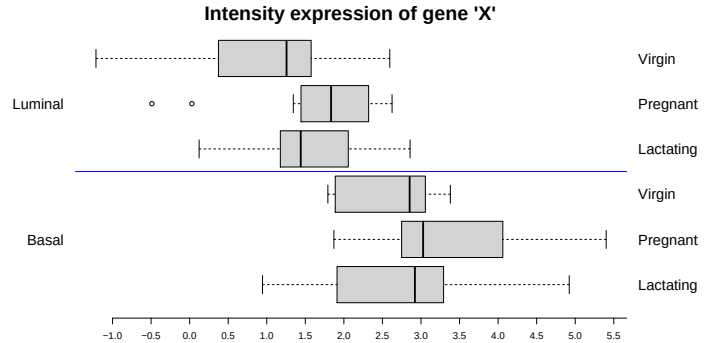
- Linear model with base 'Lactating, Basal'

$$Y_i = \mu_{L,B} + \delta_P\ I(i \in P) + \delta_V\ I(i \in V) + \theta_{L'} I(i \in L') + \epsilon_i$$
$$+ \eta_{PL'} I(i \in P\ \&\ i \in L') + \eta_{VL'} I(i \in V\ \&\ i \in L') + \epsilon_i$$
$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$$

- Hypotheses:
  - **H0**$_1$: $\delta_P = \delta_{L'} = 0$,
  - **H1**$_1$: **H0**$_1$ is false.
  - **H0**$_2$: $\theta_{L'} = 0$,
  - **H1**$_2$: **H0**$_2$ is false.
  - **H0**$_3$: $\eta_{PL'} = \eta_{VL'} = 0$,
  - **H1**$_3$: **H0**$_3$ is false.



1d/ Two-way ANOVA with interaction



- Linear model with base 'Virgin,Luminal'

$$Y_i = \mu_{L,B} + \delta_P\ I(i \in P) + \delta_V\ I(i \in V) + \theta_{L'} I(i \in L') + \epsilon_i$$
$$+ \eta_{PL'} I(i \in P\ \&\ i \in L') + \eta_{VL'} I(i \in V\ \&\ i \in L') + \epsilon_i$$
$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$$

```
Call:
lm(formula = expression ~ mousetype * celltype, data = microarrays)

Residuals:
    Min      1Q  Median      3Q     Max
-2.2424 -0.5921  0.1583  0.6059  2.1799

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                     2.7427     0.2719  10.088 2.77e-15 ***
mousetypePregnant               0.6562     0.3931   1.669  0.09956 .
mousetypeVirgin                -0.1237     0.4033  -0.307  0.75991
celltypeLuminal                -1.1476     0.3648  -3.146  0.00243 **
mousetypePregnant:celltypeLuminal -0.5980     0.5264  -1.136  0.25980
mousetypeVirgin:celltypeLuminal   -0.4437     0.5340  -0.831  0.40885
---
Signif. codes:  0 ‚Äò***‚Äô 0.001 ‚Äò**‚Äô 0.01 ‚Äò*‚Äô 0.05 ‚Äò.‚Äô 0.1 ‚Äò ‚Äô 1

Residual standard error: 0.9418 on 70 degrees of freedom
Multiple R-squared:  0.4405,Adjusted R-squared:  0.4005
F-statistic: 11.02 on 5 and 70 DF,  p-value: 7.529e-08

Analysis of Variance Table

Response: expression
                   Df Sum Sq Mean Sq F value    Pr(>F)
mousetype           2  6.567   3.283  3.7017   0.02964 *
celltype            1 41.077  41.077 46.3093 2.825e-09 ***
mousetype:celltype  2  1.243   0.622  0.7007   0.49969
Residuals          70 62.090   0.887
```

Statistical models

– We want to model the expected result of an outcome (dependent variable) under given values of other variables (independent variables)

Arbitrary function (any shape)

Expected value of variable y

A set of k independent variables (also called factors)

$$E(Y) = f(X)$$

$$Y = f(X) + \varepsilon$$

This is the variability around the expected mean of y

53

Linear models

– The observed value of Y is a linear combination of the effects of the independent variables

Arbitrary number of independent variables

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

Polynomials are valid

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \ldots + \beta_p X_1^p$$

$$E(Y) = \beta_0 + \beta_1 \log(X_1) + \beta_2 f(X_2) + \ldots + \beta_k X_k$$

We can use functions of the variables if the effects are linear

Smooth functions: not exactly the same as the so-called ***additive models***

– If we include categorical variables the model is called **General Linear Model**

54

Model Estimation

We can use **maximum likelihood estimation**

Find the set of values that maximizes the likelihood of the observed data

$$MLE : \hat{\beta} = \arg\max\{L(\beta \mid x)\}$$

$$L(\beta \mid y) = \prod f_{\beta}(y)$$

It is easier to work with the log-likelihood

In the case of errors normally distributed, the least squares and the MLE estimators are the same

57

Model Estimation

$$Y = \beta X + \varepsilon$$

β → Parameter of interest (effect of X on Y)

$\hat{\beta}$ → **Estimator** of the parameter of interest

$se(\hat{\beta})$ → **Standard Error** of the estimator of the parameter of interest

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

$$se(\hat{\beta}_i) = \sigma \sqrt{c_i}$$

where c_i is the i^{th} diagonal element of $\left(X^T X\right)^{-1}$

$\hat{y} = X\hat{\beta}$ → Fitted values (predicted by the model)

$e = y - \hat{y}$ → Residuals (observed errors)

58

1f/ Be Clever!

Not a recent idea !

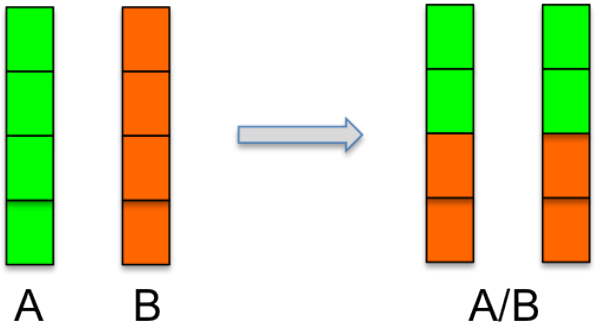
To consult a statistician after an experiment is finished is often merely to ask him to conduct a post-mortem examination. He can perhaps say what the experiment died of (Ronald A. Fisher, Indian statistical congress, 1938, vol. 4, p 17).

While a good design does not guarantee a successful experiment, a suitably bad design guarantees a failed experiment (Kathleen Kerr, Inserm workshop 145, 2003)

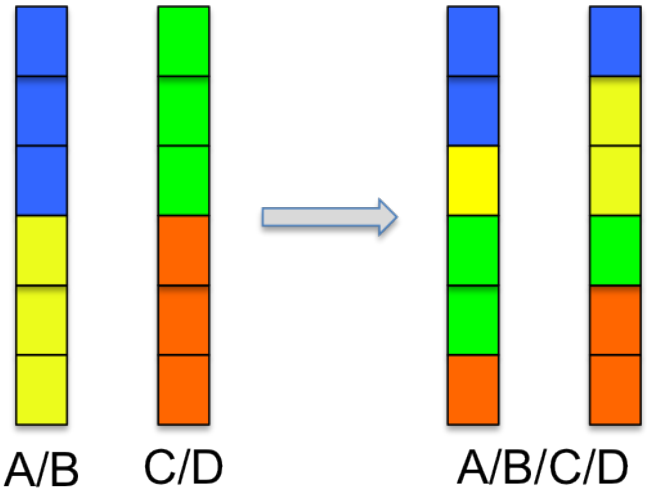
1f/ Be Clever! Confounding I

Experimental design

AVOID CONFUSION between the biological variability of interest and a biological or technical source of variation



Problem : Confusion between lane and condition
Solution : Distribute the conditions evenly on both lanes



Problem : Partial confusion between lane and condition
Solution : Distribute the conditions "evenly" on both lanes

1f/ Be Clever! Confounding II

Experimental design

Find genes that are differentially expressed between a normal skin and a damaged skin on mouse

Sample	Condition	RNA extraction date
S1	control	July 12th, 2016
S2	control	July 12th, 2016
S3	control	July 12th, 2016
S4	wound	July 20th, 2016
S5	wound	July 20th, 2016
S6	wound	July 20th, 2016

Confusion between skin status and RNA extraction date : comparing healthy and damaged skin is comparing RNAs extracted July 12th and 20th

1f/ Be Clever! Type of replicates (sample size)

Experimental design

Biological vs technical replicate

Biological replicate : Repetition of the same experimental protocol but independent data acquisition (several samples).
Technical replicate : Same biological material but independent replications of the technical steps (several extracts from the same sample).

Sequencing technology does not eliminate biological variability.
(Nature Biotechnology Correspondence, 2011)

$$\text{lane effect} < \text{run effect} < \text{library prep effect} << \text{biological effect}$$

[Marioni et al., 2008],[Bullard et al., 2010]

Include at least three biological replicates in your experiments !
Technical replicates are not necessary.

1f/ Be Clever! Number of replicates (sample size)

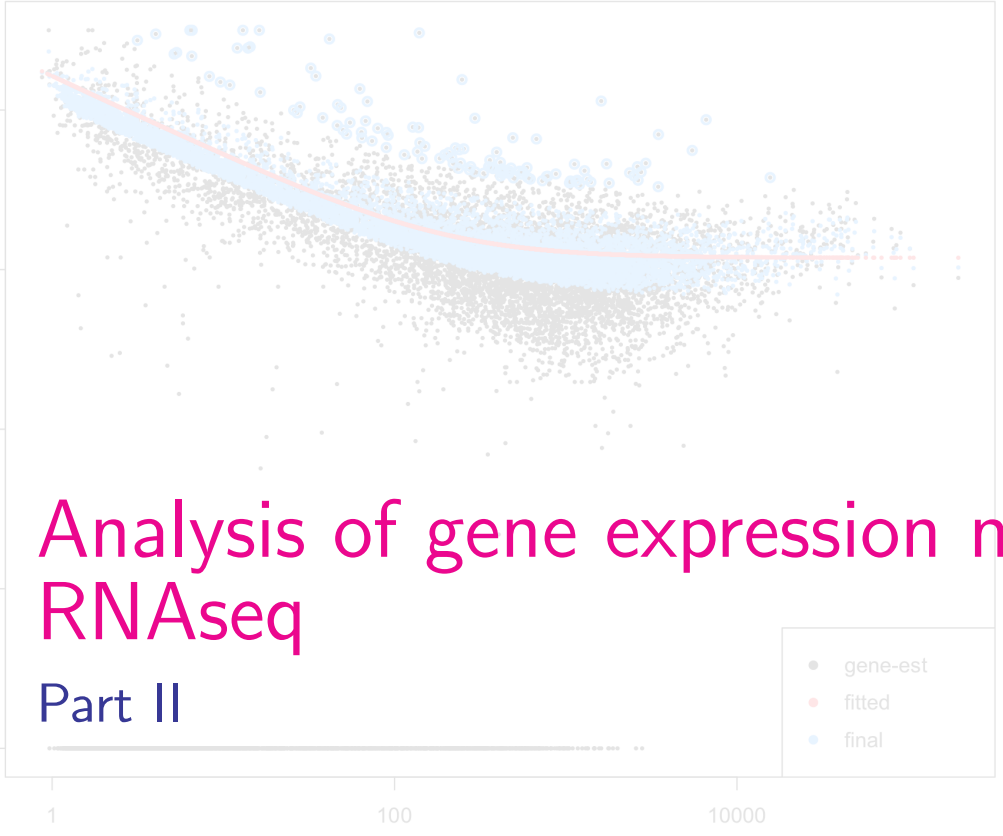
Experimental design

Why increasing the number of biological replicates ?

- To generalize to the population level
- To estimate with a higher degree of accuracy variation in individual transcript [Hart et al., 2013]
- To improve detection of DE transcripts and control of false positive rate [Soneson and Delorenzi, 2013]
- To focus on detection of low mRNAs, inconsistent detection of exons at low levels ($\leq$ 5 reads) of coverage [McIntyre et al., 2011]

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

UNIVERSITY OF CAMBRIDGE



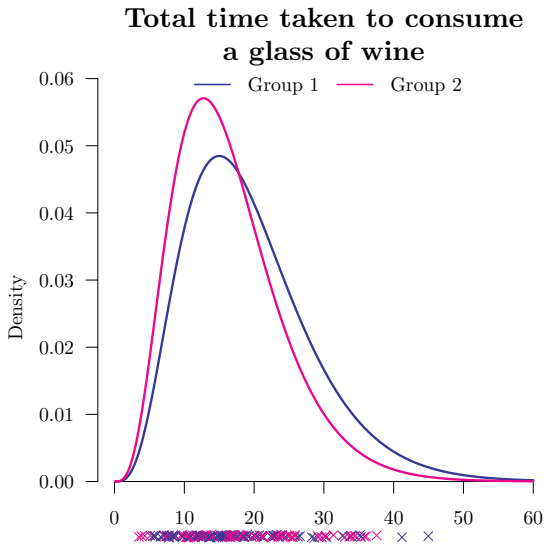
Analysis of gene expression measured with RNAseq

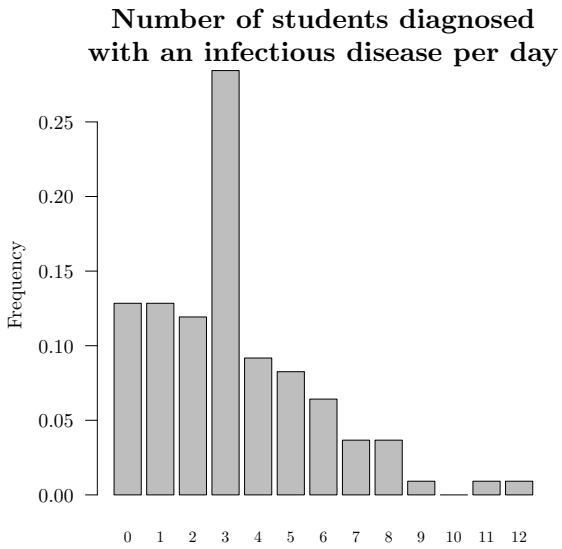
Part II

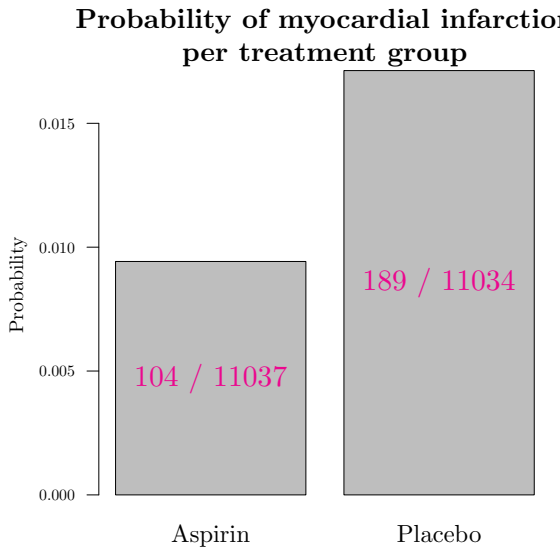
dominique-laurent.couturier@cruk.cam.ac.uk [Bioinformatics core]

(Source: O. Rueda, CRUK-CI; G. Marot, INRIA)

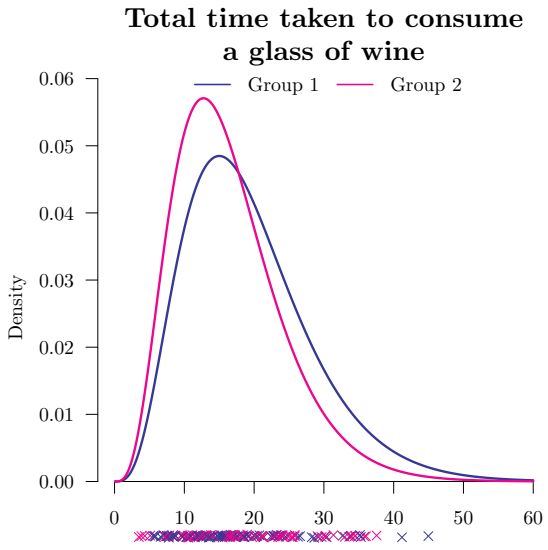
Examples of data with non-normal conditional distributions

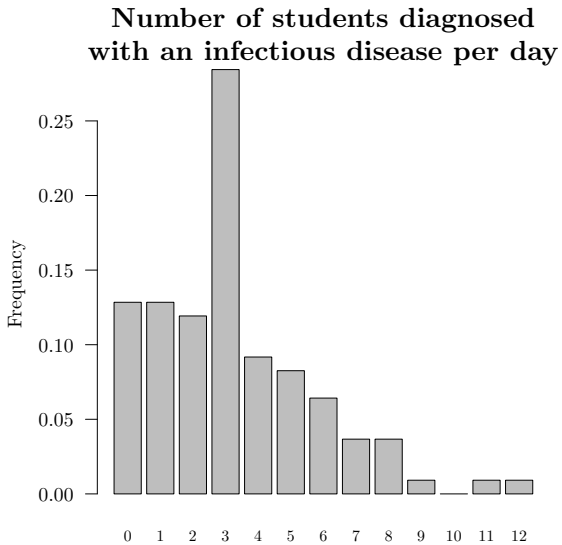


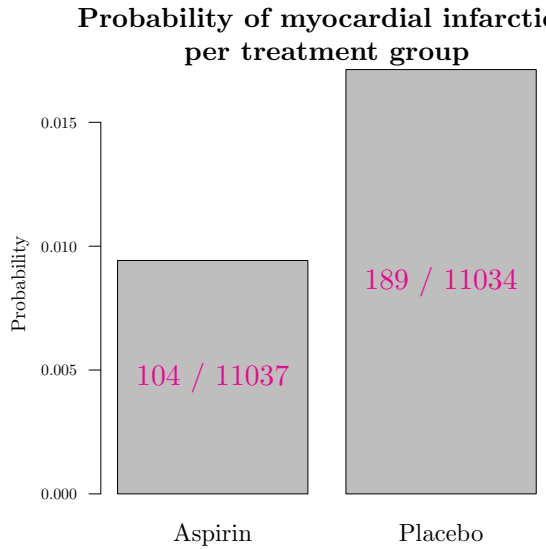




Examples of data with non-normal conditional distributions







Linear model not suitable:

- Assumed model:

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i \text{ where } \epsilon_i \sim N(0, \sigma^2),$$

$$Y_i | (\mathbf{x}_i, \boldsymbol{\beta}) \sim N(\mu_i, \sigma^2).$$

  - theoretical range of $\epsilon_i = [-\infty, +\infty]$,
  - $\mathbf{x}_i^T \boldsymbol{\beta}$ not bounded to $[0, \infty]$ or $[0, 1]$,
  - Var$[Y_i]$ independent of E$[Y_i]$.

- Solution:

$$Y_i | (\mathbf{x}_i, \boldsymbol{\beta}, \phi) \sim distribution(function(\mathbf{x}_i^T \boldsymbol{\beta}), \phi),$$

GLM: conditional distributions

$$Y_i|(\mathbf{x}_i, \boldsymbol{\beta}, \phi) \sim distribution(function(\mathbf{x}_i^T \boldsymbol{\beta}), \phi),$$

- Some possible conditional *distributions* :
  statistical probability mass functions & density functions
  - ▷ Within the exponential family [‘classical’ GLM framework]

normal	chi-squared	Poisson	Inverse Wishart
exponential	beta	Negative Binomial	
gamma	Dirichlet	Bernoulli	...

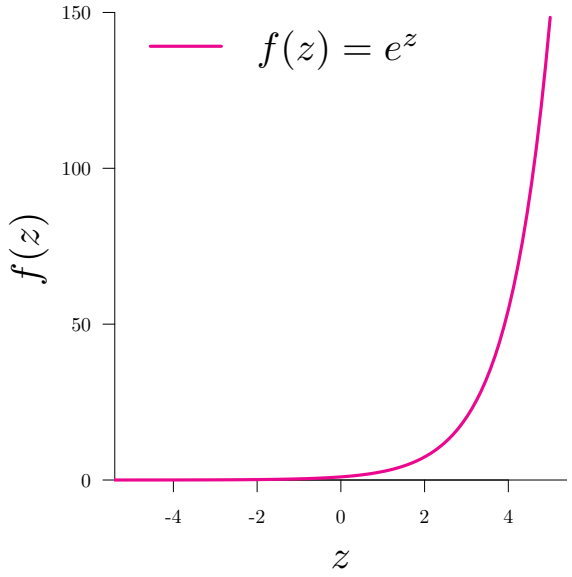
  - ▷ Outside the exponential family [‘extended’ GLM framework]

Box-Cox power	Gaussian	Weibull
exponential	inverse Gaussian	Pareto type I, II, III
exponential Gaussian	logistic	Poisson inverse Gaussian
generalized beta	power exponential	
generalized gamma	reverse Gumbel	...
generalized inverse	skew power exponential	

GLM: link functions

$$Y_i|(\mathbf{x}_i, \boldsymbol{\beta}, \phi) \sim distribution(function(\mathbf{x}_i^T \boldsymbol{\beta}), \phi),$$

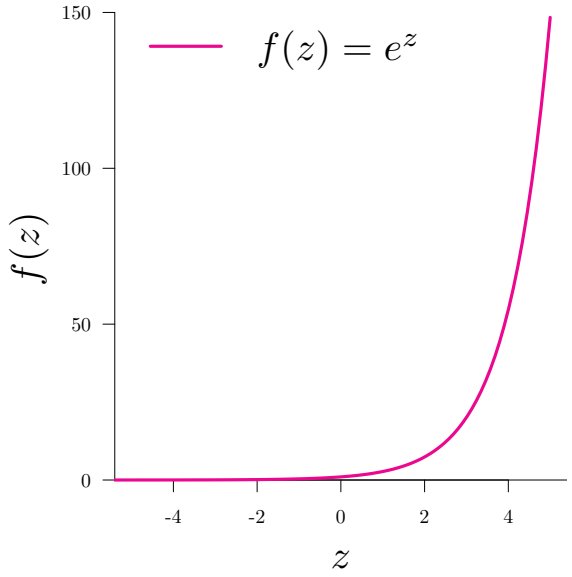
- Most used link $functions$:
  connection between Y_i and $\mathbf{x}_i^T \boldsymbol{\beta}$
  - to restrict $f(\mathbf{x}_i^T \boldsymbol{\beta})$ to belong to $[0, \infty[$:
    - log link: $f(z) = e^z$

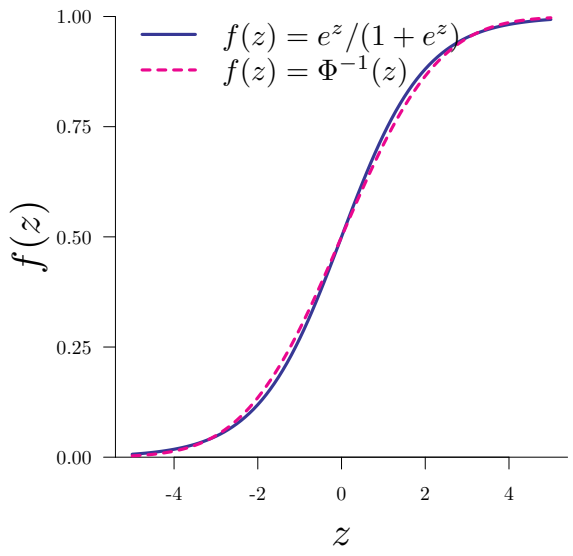


GLM: link functions

$$Y_i|(\mathbf{x}_i, \boldsymbol{\beta}, \phi) \sim distribution(function(\mathbf{x}_i^T\boldsymbol{\beta}), \phi),$$

- Most used link $functions$:
  connection between Y_i and $\mathbf{x}_i^T\boldsymbol{\beta}$
  - to restrict $f(\mathbf{x}_i^T\boldsymbol{\beta})$ to belong to $[0, \infty[$:
    - log link: $f(z) = e^z$
  - to restrict $f(\mathbf{x}_i^T\boldsymbol{\beta})$ to belong to $[0, 1]$:
    - logit link: $f(z) = e^z/(1 + e^z) = 1/(1 + e^{-z})$ where z is positive
    - probit link: $f(z) = \Phi(z)$, where Φ denotes the $N(0, 1)$.





Distribution for count data: Poisson

Example:
Interest for the number of reads/counts for gene 'X' for a sample basal cells of n mice

Sample of n mice:	$i = 1$	$i = 2$	$i = 3$	$\cdots$	$i = 115$
y_i	607	873	1218	$\cdots$	2715

If, during a time interval or in a given area,

- events occur independently,
- at the same rate,
- and the probability of an event to occur in a small interval (area) is proportional to the length of the interval (size of the area),

then,

- a count occurring in a fixed time interval or in a given area, Y, may be modelled by means of a Poisson distribution with parameter μ:

$$Y \sim Poisson(\mu) \text{ where } \mu = \mathsf{E}[Y] = \mathsf{Var}[Y],$$

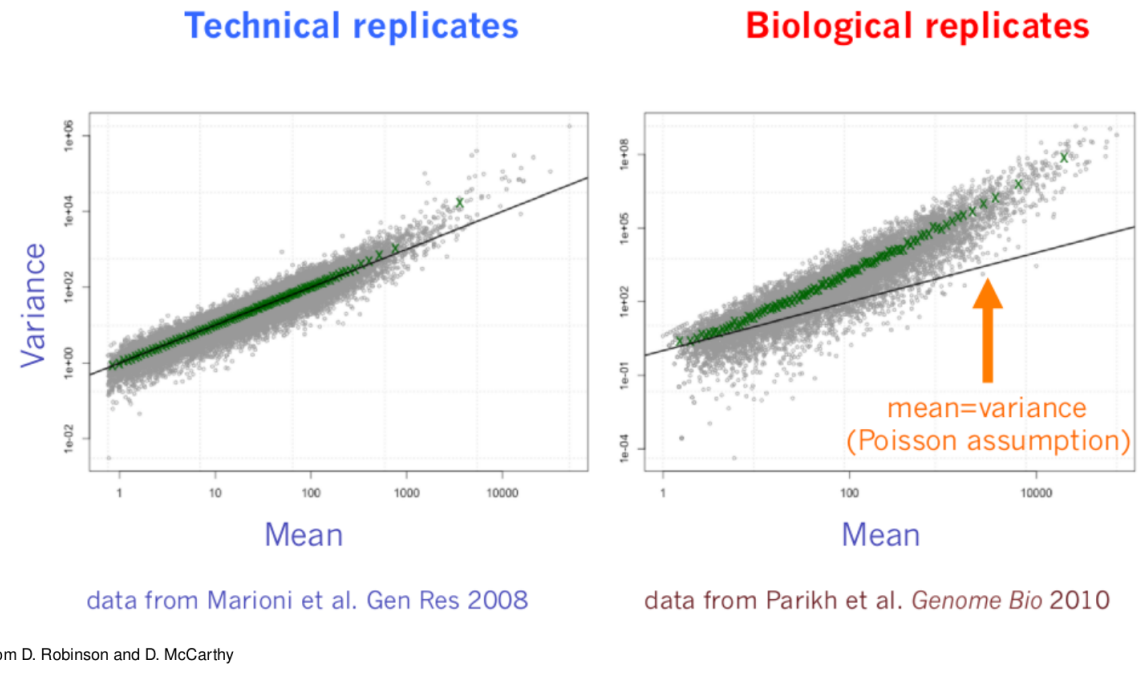
- the probability of observing x events during a fixed time interval or in a given area is given by

$$P(Y = y|\mu) = \frac{\mu^y e^{-\mu}}{y!}.$$

Distribution for count data: Poisson vs Neg. Bin.

Exploratory data analysis

scores between 0 and 1 $\Rightarrow$ underdispersion (variance smaller than mean)



From D. Robinson and D. McCarthy

scores greater than 1 : overdispersion $\Rightarrow$ adapted to biological replicates

Distribution for count data: Poisson vs Neg. Bin.

Available tests

Models of count data

- Data transformation and gaussian-based model : limma - voom
- Poisson : TSPM
- Negative Binomial : edgeR, DESeq(2), NBPSeq, baySeq, ShrinkSeq, ...

Statistical approaches

- Frequentist Approach : edgeR, DESeq(2), NBPSeq, TSPM, ...
- Bayesian Approach : baySeq, ShrinkSeq, EBSeq, ...
- Non-parametric approach : SAMSeq, NOISeq, ...

2a/ Negative binomial

- General form:

$$Y_i \sim \text{NB}(\mu_i, \phi)$$

$$f_{Y_i}(y_i|\mu_i, \phi) = \frac{\Gamma(y + \frac{1}{\phi})}{\Gamma(\frac{1}{\phi})\Gamma(y+1)} \left(\frac{\phi\mu_i}{1+\phi\mu_i}\right)^y \left(\frac{1}{1+\phi\mu_i}\right)^{\frac{1}{\phi}}$$

with expectation and variance given by

  - $\text{E}[Y_i] = \mu_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta})$
  - $\text{Var}[Y_i] = \mu_i(1 + \phi\mu_i)$

and a coefficient of variation (CV) of given by

  - $\text{CV}^2 = \frac{1}{\mu_i} + \phi_i$

2b/ Negative binomial: Estimation

Empirical bayesian approaches

Principles

- Bayes theorem : $P(A/B) = P(B/A)P(A)$
- "empirical" $\Rightarrow$ priors from the observed data

$$\tilde{\boldsymbol{\theta}}_{\mathbf{g}} = \hat{\boldsymbol{\theta}}_{\mathbf{c}} + b(\widehat{\boldsymbol{\theta}_{\mathbf{g}}} - \hat{\boldsymbol{\theta}}_{\mathbf{c}})$$

with $\tilde{\boldsymbol{\theta}}_{\mathbf{g}}$ = shrinkage estimator
$\hat{\boldsymbol{\theta}}_{\mathbf{c}}$ = estimator of the mean population
$\widehat{\boldsymbol{\theta}_{\mathbf{g}}}$ = usual empirical estimator gene by gene
b =shrinkage factor

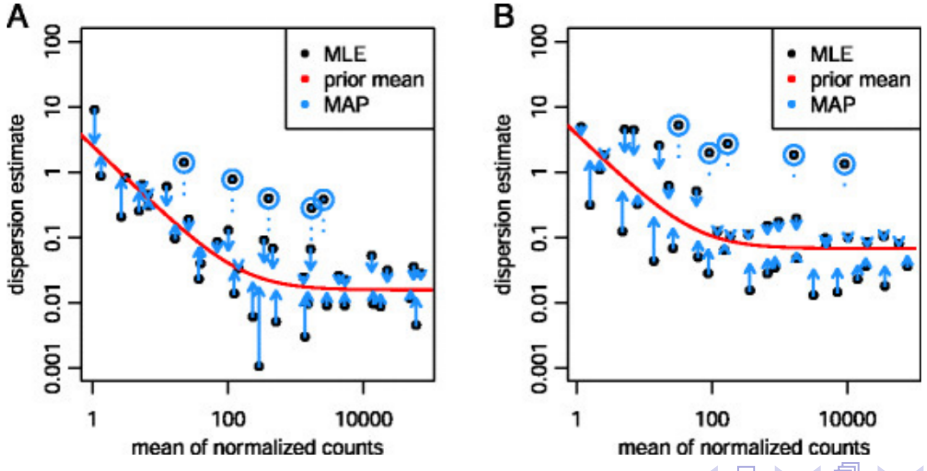
$b = 1 \Rightarrow \tilde{\boldsymbol{\theta}}_{\mathbf{g}} = \widehat{\boldsymbol{\theta}_{\mathbf{g}}}$
$b = 0 \Rightarrow \tilde{\boldsymbol{\theta}}_{\mathbf{g}} = \hat{\boldsymbol{\theta}}_{\mathbf{c}}$

2b/ Negative binomial: Estimation

Dispersion estimation with DESeq2

Hypothesis : genes of similar average expression strength have similar dispersion

1. Estimate gene-wise dispersion estimates using maximum likelihood (ML) (black dots)
2. Fit a smooth curve (red line)
3. Shrink the gene-wise dispersion estimates (empirical Bayes approach) toward the values predicted by the curve to obtain final dispersion values (blue arrow heads).



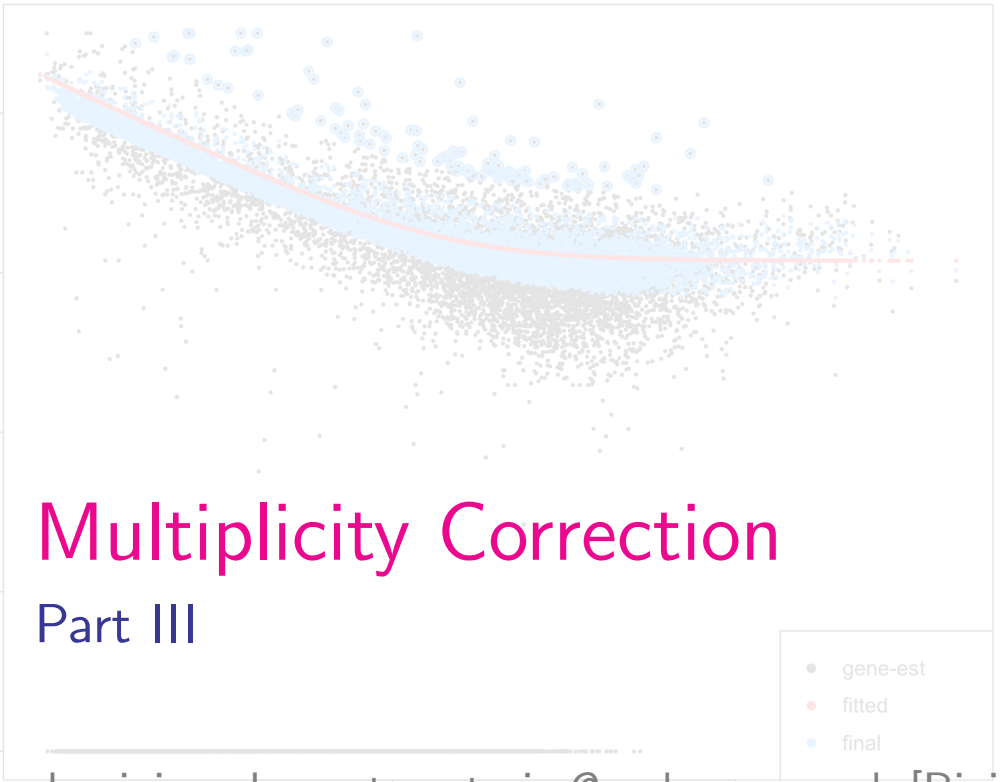
2b/ Negative binomial: Controlling for library size

- For a given gene, the variance of the Negative Binomial for the ith sample is given by

$$\text{Var}(Y_i) = \mu_i(1 + \phi\mu_i)$$

- To control for the library size S_i of the ith sample, DESeq2 uses

$$\text{Var}(Y_i) = S_i\mu_i(1 + \phi S_i\mu_i)$$



Multiplicity Correction

Part III

dominique-laurent.couturier@cruk.cam.ac.uk [Bioinformatics core]

(Source: O. Rueda, CRUK-CI; G. Marot, INRIA)

3/ Multiplicity correction

Simultaneous tests of G null hypotheses

Reality	Declared non diff. exp.	Declared diff. exp.
G_0 non DE genes	**True Negatives** (*TN*)	**False Positives** (*FP*)
G_1 DE genes	**False Negatives** (*FN*)	**True Positives** (*TP*)
G Genes	N Negatives	P Positives

Aim : minimize *FP* and *FN*.

3/ Multiplicity correction

Multiple Testing

False positive (FP) : A non differentially expressed (DE) gene which is declared DE.

For all 'genes', we test H_0 (gene i is not DE) vs H_1 (the gene is DE) using a statistical test

Problem

Let assume all the G genes are not DE. Each test is realized at α level
Ex : $G = 10000$ genes and $\alpha = 0.05 \rightarrow E(FP) = 500$ genes.

3/ Multiplicity correction

The Family Wise Error Rate (FWER)

Definition

Probability of having at least one Type I error (false positive), of declaring DE at least one non DE gene.

$$FWER = \mathbb{P}(FP \leq 1)$$

The Bonferroni procedure

Either each test is realized at $\alpha = \alpha^*/G$ level
or use of adjusted pvalue $pBonf_i = min(1, p_i * G)$ and FWER $\leq \alpha^*$.
For $G = 2000$, $\leq \alpha^* = 0.05$, $\alpha = 2.510^{-5}$.

Easy but conservative and not powerful.

The False Discovery Rate (FDR)

Idea : Do not control the error rate but the proportion of error
$\Rightarrow$ less conservative than control of the FWER.

Definition

The false discovery rate of [Benjamini and Hochberg, 1995] is the expected proportion of Type I errors among the rejected hypotheses

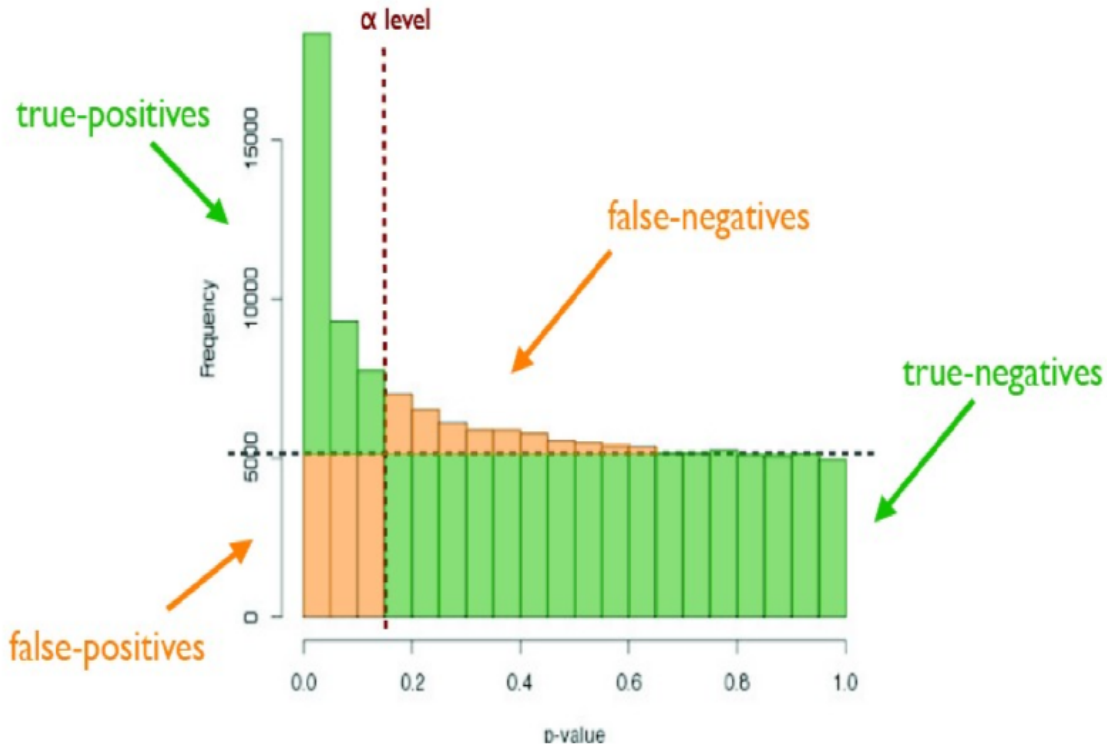
$$\text{FDR} = \mathbb{E}(FP/P) \text{ if } P > 0 \text{ and } 0 \text{ if } P = 0$$

Prop

$$\text{FDR} \leq \text{FWER}$$

3/ Multiplicity correction

Standard assumption for p-value distribution

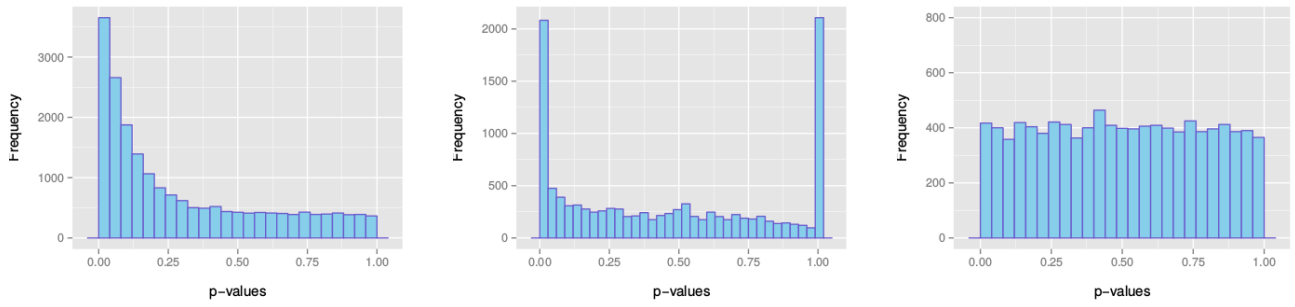


Source : M. Guedj, Pharnext

3/ Multiplicity correction

p-values histograms for diagnosis

Examples of expected overall distribution



(a) : the most desirable shape

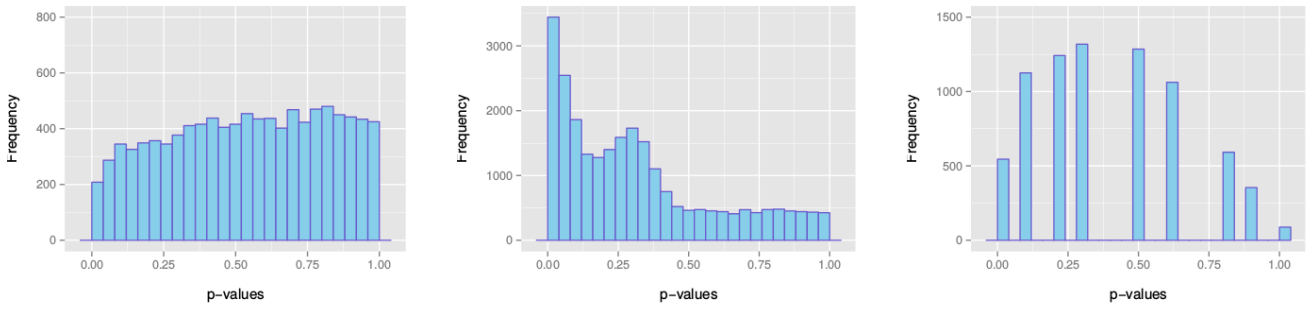
(b) : very low counts genes usually have large p-values

(c) : do not expect positive tests after correction

3/ Multiplicity correction

p-values histograms for diagnosis

Examples of not expected overall distribution



(a) : indicates a batch effect (confounding hidden variables)

(b) : the test statistics may be inappropriate (due to strong correlation structure for instance)

(c) : discrete distribution of p-values : unexpected

3/ Multiplicity correction

Multiple testing : key points

- Important to control for multiple tests
- FDR or FWER depends on the cost associated to FN and FP

Controlling the FWER :

Having a great confidence on the DE elements (strong control). Accepting to not detect some elements (lack of sensitivity $\Leftrightarrow$ a few DE elements)

Controlling the FDR :

Accepting a proportion of FP among DE elements. Very interesting in exploratory study.