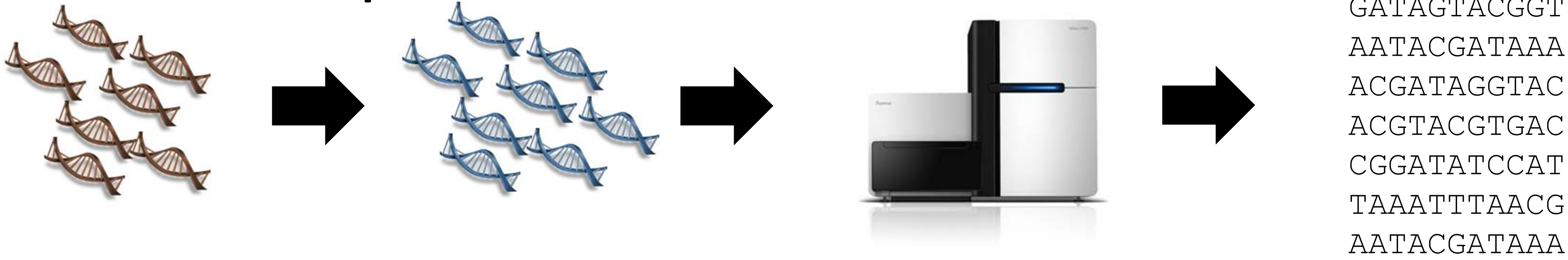
# Lecture 8

Introduction to bioinformatics (MVE510)

Autumn, 2020

Additional reading: Lecture notes – Linear models for RNA-seq analysis

# Repetition:the RNA-seq process



**Reverse transcription**

```
GATAGTACGGT
AATACGATAAA
ACGATAGGTAC
ACGTACGTGAC
CGGATATCCAT
TAAATTTAACG
AATACGATAAA
```

**Extracted RNA (total or mRNA)**

**Complementary DNA (cDNA)**

**Sequencer**

**Reads**

# Analysis of RNA seq data

**Three main steps**

1. Quantification of the gene expression
    - From reads to a (semi)quantitative measurements of gene expression
2. Normalization
    - Correction of systematic errors within and between samples
3. Identification of differentially abundant genes
    - Find genes with a significant difference in gene expression

# Three main approaches

**Methods based on normal assumptions**

**Methods based on non-parametric methods**

**Methods based on count distributions**

# Today’s agenda

- Introduction to linear models
- Linear model in R
- A first example: the cat dataset revisited
- Linear models for the analysis of RNA-seq data
- A second more comprehensive example: gene expression of SI-NETs

# Linear models in R

- Linear models can easily be fitted to any data using the **lm** function (stands for ‘linear model’).
- **lm** has two important argument
  **1.** A model formulation which is specified through an R ‘formula’.
  **2.** A data.frame with data used to fit the model.
- By default, **lm** always assumes independent and normally distributed errors and the model is fit using maximum likelihood.
- The names of the column of the data.frame needs to match the names of the dependent and independent variables

# Linear models in R: formulas

- The formula is written the form `y~model` where ‘y’ is the dependent variable and ‘model’ specifies the independent variables.
- The intercept does not need to be specified and is included by default.

| R formulation | Linear model |
|---|---|
| `y~x1` | $Y_i = \beta_0 + \beta_1 x_{i,1} + \varepsilon_i$ |
| `y~x1+x2` | $Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \varepsilon_i$ |
| `y~x1+x2-1` | $Y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \varepsilon_i$ |

# A first example: the cat dataset revisited

```
> library(MASS)
> data(cats)
> dim(cats)
[1] 144   3
```

```
> head(cats)
  Sex Bwt Hwt
1   F 2.0 7.0
2   F 2.0 7.4
3   F 2.0 9.5
4   F 2.1 7.2
5   F 2.1 7.3
6   F 2.1 7.6
```

# Linear models in R: the cat dataset revisited

Can we describe the hearth weight of a cat? Let $Y_j$ be the heart weight of cat *j* (*j*=1,…, 144). Let $x_{1,j}$ define the sex of cat *j*, i.e.

$$x_{1,j} = \begin{cases} 1 & \text{if cat } j \text{ is male,} \\ 0 & \text{if cat } j \text{ is female.} \end{cases}$$

The first model that we will use is

$$Y_j = \beta_0 + \beta_1 x_{1,j} + \varepsilon_j.$$

As before, $\varepsilon_j \sim \mathrm{Normal}(0, \sigma^2)$.

# Linear models in R: the cat dataset revisited

The model is specified into R using

```
> lm(Hwt~Sex, data=cats)

Call:
lm(formula = Hwt ~ Sex, data = cats)

Coefficients:
(Intercept)         SexM
      9.202        2.121
```

# Linear models in R: the cat dataset revisited

```
> summary(lm(Hwt~Sex, data=cats))

Call:
lm(formula = Hwt ~ Sex, data = cats)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.8227 -1.7227  0.0273  1.2273  9.1773 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   9.2021     0.3251  28.308  < 2e-16 ***
SexM          2.1206     0.3961   5.354 3.38e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.229 on 142 degrees of freedom
Multiple R-squared:  0.168,     Adjusted R-squared:  0.1621 
F-statistic: 28.66 on 1 and 142 DF,  p-value: 3.38e-07
```

# Linear models in R: the cat dataset revisited

Let $x_{2j}$ denote the body weight of cat *j* a refined model can be stated as

$$Y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \varepsilon_j.$$

```
> lm(Hwt~Sex+Bwt, data=cats)

Call:
lm(formula = Hwt ~ Sex + Bwt, data = cats)

Coefficients:
(Intercept)         SexM          Bwt
    -0.4150      -0.0821       4.0758
```

# Linear models in R: the cat dataset revisited

```
> summary(lm(Hwt~Sex+Bwt, data=cats))

Call:
lm(formula = Hwt ~ Sex + Bwt, data = cats)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.5833 -0.9700 -0.0948  1.0432  5.1016 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -0.4149     0.7273  -0.571    0.569    
SexM         -0.0821     0.3040  -0.270    0.788    
Bwt           4.0758     0.2948  13.826   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.457 on 141 degrees of freedom
Multiple R-squared:  0.6468,    Adjusted R-squared:  0.6418 
F-statistic: 129.1 on 2 and 141 DF,  p-value: < 2.2e-16
```

# Gene expression analysis using linear models

- **Challenge:** Data consists of many genes/transcripts (often >10,000)
- **Strategy:** Fit a linear model to each gene!
- This will result in
  - >10,000 linear models
  - >10,000 estimates of the parameter
  - >10,000 p-values
- It is practically impossible to examine the result from each fitted linear model individually!
- We can however use ‘clever plots’ to visualize **all** fitted parameter estimates and their p-values at the same time.

# Example: Gene expression in SI-NETs

**Data**: Gene expression of 45,015 transcripts in 33 patients. 23 were from less aggressive group and 10 from more aggressive group. For each patient, a biopsy was sampled, the RNA extracted, sequenced using standard protocols, preprocessed and transformed and normalized so that the resulting data is approximately normal distributed. For each patient, the age and gender was also recorded.

**Aim**: Identify genes that are differentially expressed between the groups

# Example: Gene expression in SI-NETs

The expression data is available as log CPM (counts per million mapped reads).

33 samples, one for each patient

45,015 transcripts

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **APOBEC3B** | -1.3080 | 1.2738 | -0.2291 | 0.1059 | -0.4974 | -0.9348 | -0.6349 | -0.6301 | -0.0095 | 1.1046 | -0.3840 | 2.3720 | 0.3894 | -0.4803 | 0.1640 | |
| **ATP11B** | 2.5493 | 3.2565 | 3.2085 | 3.1314 | 4.0809 | 3.8099 | 3.8516 | 4.0956 | 3.1072 | 2.7990 | 3.2238 | 2.1573 | 2.9087 | 2.9061 | 3.4756 | |
| **LOC100132006** | 1.5618 | 0.0585 | -1.0288 | -1.0245 | -0.7023 | -1.0489 | -0.3975 | -1.3643 | -0.0095 | -0.1328 | -0.6004 | -0.5999 | -0.1955 | -0.4803 | -0.4332 | - |
| **DNAJA1** | 5.3760 | 6.2587 | 7.4256 | 6.7160 | 6.3814 | 6.2466 | 6.9555 | 6.0039 | 6.2113 | 6.1688 | 6.0132 | 6.6026 | 6.0905 | 6.5462 | 6.1255 | |
| **EHMT2** | 0.4388 | 1.9512 | 2.4935 | 2.3967 | 1.6930 | 1.9324 | 2.7159 | 1.7267 | 2.3780 | 2.6070 | 2.3119 | 1.9179 | 0.9754 | 2.8758 | 1.9197 | |
| **RPL23** | 9.0068 | 9.4358 | 9.9730 | 10.0973 | 10.3703 | 9.6257 | 10.2602 | 10.4557 | 10.3915 | 9.7207 | 10.3020 | 9.2524 | 10.0280 | 10.0772 | 10.3294 | |
| **RPS13** | 7.0665 | 8.0025 | 8.8169 | 8.3768 | 8.9800 | 8.8094 | 8.6004 | 9.1958 | 8.2512 | 8.0732 | 8.4829 | 8.3298 | 8.7682 | 8.2037 | 8.4305 | |
| **HDDC3** | 4.4907 | 5.8964 | 5.6970 | 6.1755 | 4.7235 | 5.5553 | 5.8848 | 5.1051 | 5.6407 | 5.6031 | 5.5775 | 5.0305 | 5.1183 | 6.0040 | 5.0338 | |
| **ITPRIPL2** | 4.7796 | 5.3413 | 5.5084 | 5.3690 | 6.3345 | 6.4191 | 4.9050 | 7.0215 | 5.4606 | 5.9255 | 6.8179 | 6.1367 | 6.8954 | 5.3338 | 5.2335 | |
| **MEGF11** | 3.8322 | 3.5196 | 4.4392 | 4.4959 | 3.0314 | 4.6823 | 5.3699 | 3.4580 | -1.9481 | -0.8646 | 3.3412 | -2.0039 | -0.0593 | 4.3704 | 4.9336 | |
| **APBA3** | 0.8934 | 2.4857 | 2.2194 | 1.6381 | 0.5120 | 1.6215 | 1.6203 | 1.1843 | 1.0574 | 2.1153 | 1.0123 | 2.6806 | -0.0593 | 1.8369 | 0.8602 | |
| **CRCP** | -1.3080 | 1.6620 | 2.0027 | 1.3050 | 1.1986 | 1.1388 | 1.4457 | 1.5825 | 1.4785 | 2.1255 | 0.4539 | 0.7771 | 0.7500 | 2.8043 | 2.3283 | |
| **CATSPERG** | 0.4388 | 1.6620 | -1.4684 | -0.8890 | -1.0377 | -1.1605 | -0.5497 | -1.9087 | -0.2434 | -1.0133 | -0.9930 | -2.4231 | -1.5598 | -1.1634 | -1.0569 | |
| **KBTBD4** | 3.0410 | 4.9136 | 5.2735 | 4.7079 | 4.1695 | 4.5461 | 5.1626 | 4.2997 | 5.2048 | 3.5967 | 4.0339 | 3.0177 | 2.5556 | 4.7077 | 5.0728 | |
| **SLED1** | -1.3080 | 0.0585 | -1.2969 | -1.8432 | -1.7387 | -1.5791 | -0.0676 | -1.0893 | -1.2196 | 0.0656 | -2.1872 | 0.7771 | -1.7203 | -1.9351 | -0.8524 | |
| **LRP1** | 2.8198 | 2.6055 | 1.6173 | 2.5457 | 3.3422 | 3.2787 | 1.7357 | 3.2244 | 1.6452 | 1.7286 | 2.6295 | 3.4818 | 2.3179 | 1.9766 | 1.3362 | |
| **TSC1** | 4.8925 | 5.3413 | 6.7008 | 5.6007 | 5.3680 | 6.3200 | 6.5464 | 5.7273 | 6.5691 | 5.3013 | 5.9183 | 5.8130 | 4.3613 | 7.0232 | 6.7578 | |
| **ADORA3** | -1.3080 | 0.3075 | -1.9988 | -1.9626 | -1.5419 | -1.1605 | -1.5308 | 0.9833 | 0.0927 | -1.3080 | -1.2574 | -2.2403 | 1.2852 | -1.7367 | -1.5083 | |
| **GOLGA3** | 7.1910 | 8.2709 | 8.2451 | 8.2910 | 7.8938 | 7.9649 | 8.1578 | 7.9703 | 8.3507 | 7.5201 | 8.0828 | 7.5201 | 7.6011 | 8.2854 | 7.8768 | |
| **CDH6** | 2.4089 | 2.6055 | 3.5993 | 3.3648 | 2.2291 | 2.8344 | 3.5798 | 2.4037 | 2.5515 | 2.0331 | 3.0900 | 2.0780 | 1.0436 | 3.7041 | 3.3674 | |
| **CD99L2** | 7.7646 | 8.7518 | 8.9573 | 9.0448 | 8.8229 | 9.0057 | 9.0378 | 9.0860 | 9.2734 | 8.7090 | 8.6807 | 8.6743 | 8.3349 | 9.0227 | 8.9655 | |
| **SMG7** | 4.3290 | 5.5150 | 4.9404 | 5.8186 | 4.5539 | 4.8396 | 5.5633 | 4.3482 | 5.7882 | 3.5967 | 3.8761 | 4.3709 | 3.4132 | 5.3875 | 5.0413 | |
| **C9orf30** | 1.5618 | 3.2942 | 2.6625 | 2.3509 | 2.7460 | 2.1913 | 3.2096 | 2.9466 | 2.3494 | 2.8388 | 1.8588 | 2.0780 | 2.6055 | 2.6301 | 2.9089 | |
| **CXorf57** | -2.1763 | 1.6620 | -1.0288 | -2.1751 | -1.7888 | -2.3085 | -0.4356 | -1.7751 | -2.6751 | -1.7578 | -2.4766 | -2.2403 | -2.2579 | -1.5974 | -0.1983 | - |

# Example: Gene expression in SI-NETs

We have also some data about the patients (‘metadata’).

33 samples, one for each patient

```
Sample Group Gender Age
     1    G2   Male  49
     2    G1 Female  81
     3    G1 Female  64
     4    G1 Female  42
     5    G1 Female  74
     6    G1 Female  58
     7    G1   Male  63
     8    G1   Male  59
     9    G1 Female  53
    10    G2 Female  61
    11    G2 Female  70
    12    G2   Male  74
    13    G2 Female  75
    14    G1   Male  64
    15    G2   Male  71
```

# Example: Gene expression in SI-NETs

Let $Y_{i,j}$ is the expression of gene *i* in patient *j* and define the independent variable $x_{1,j}$ as

$$x_{1,j} = \begin{cases} 1 & \text{if patient } j \text{ has a more aggressive tumor,} \\ 0 & \text{if patient } j \text{ has a less aggresive tumor,} \end{cases}$$

A simple linear model can be formulated as

log fold-change

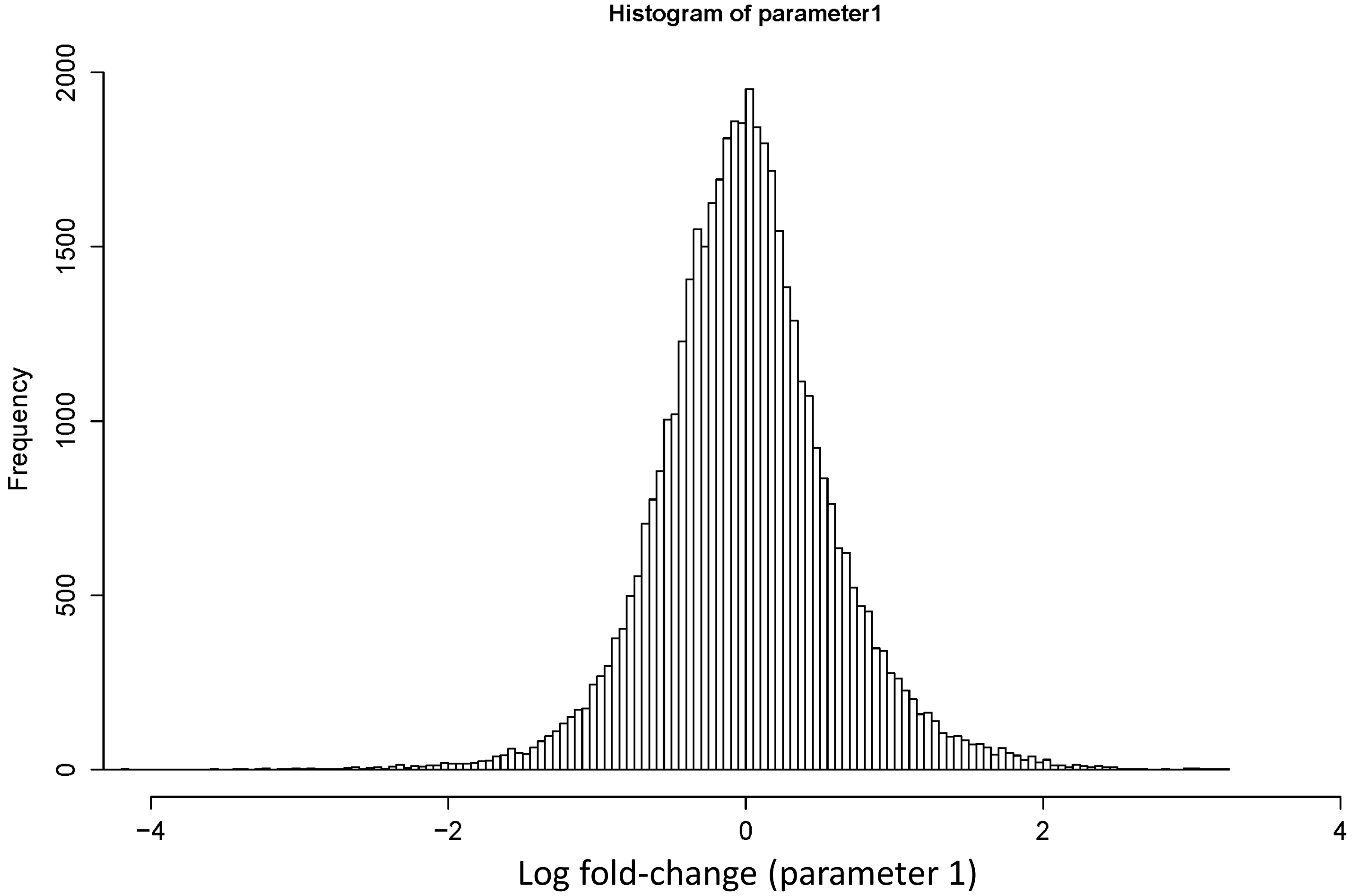
$$Y_{i,j} = \beta_{0,i} + \beta_{1,i} x_{1,j} + \varepsilon_{i,j}$$

where $\beta_{0,i}$ is the base line expression of gene $i$ and $\beta_{1,i}$ is the difference in expression in the more aggressive compared to the less aggressive tumor form and $\varepsilon_{i,j}$ is the error for gene *i* and patient *j* (normal).

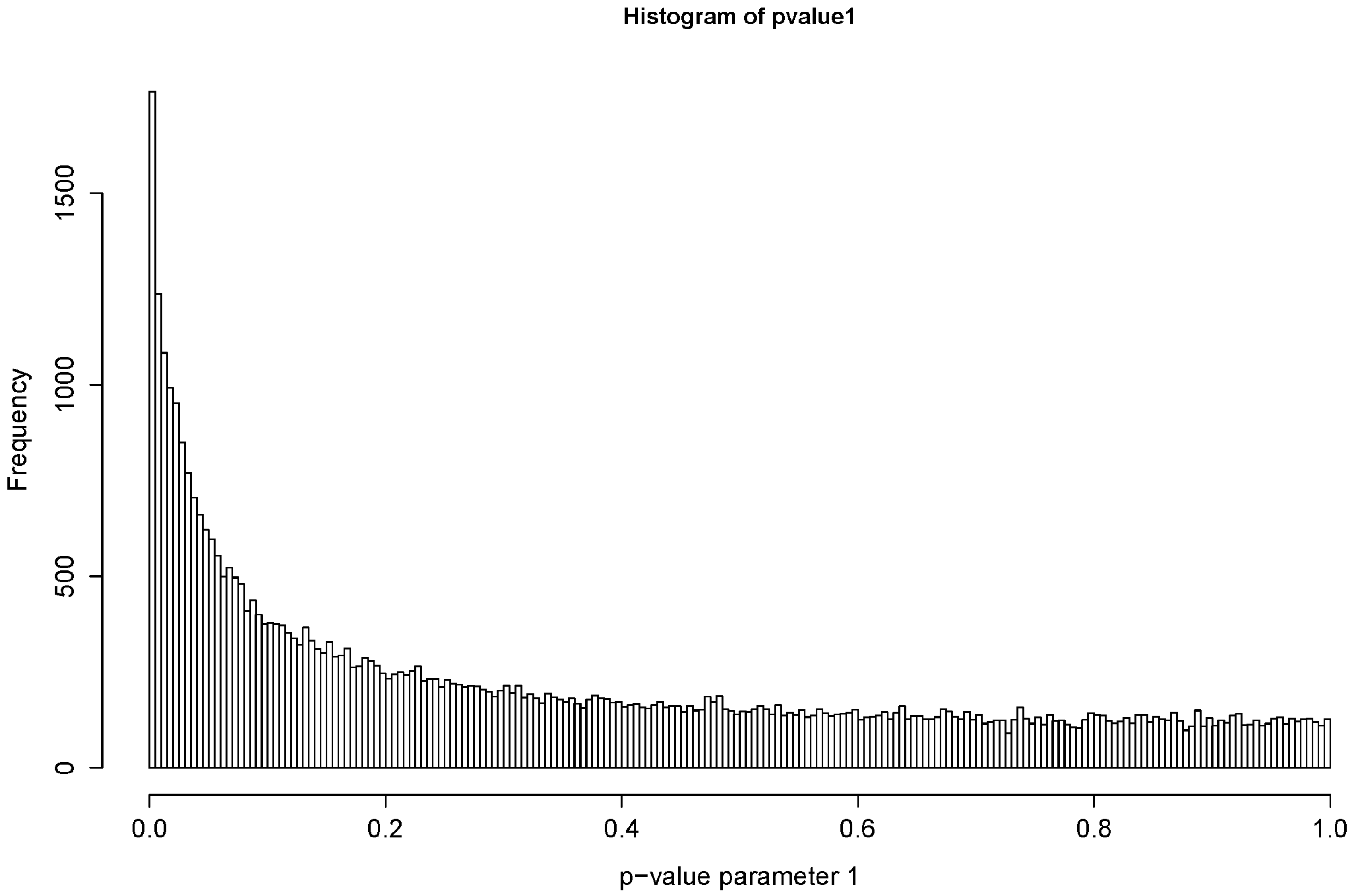
# We will fit a linear model to each gene by looping over the data.

```
# R code for fitting a linear model to each transcript
ngenes=nrow(data.exp) # Number of rows (transcripts) in the expression data

parameter=rep(NA, ngenes) # Allocate a vector for parameter values
pvalue=rep(NA, ngenes) # Allocate a vector for p-values

for(i in 1:ngenes){ # Loop over the number of transcripts

  # Create a data.frame for lm for transcript i
  data.cur=data.frame(data.exp[i,], annotation[,2], annotation[,3], annotation[,4])

  # Add column names to the data.frame
  colnames(data.cur)=c("Expression", "Group", "Gender", "Age")

  # Fit a linear model
  res.lm=lm(Expression~Group, data=data.cur)

  # Save the parameter 1 ("Group") from the result. Note that res.lm$coef[1] is the intercept.
  parameter1[i]=res.lm$coefficient[2]

  # Calculate p-values (and other stuffs for the model)
  res.summary.lm=summary(res.lm)

  # Extact the p-value. Note that $coefficient after running summar
  # is a matrix with a row for each parameter.
  pvalue1[i]=res.summary.lm$coefficient[2,4]

}
```

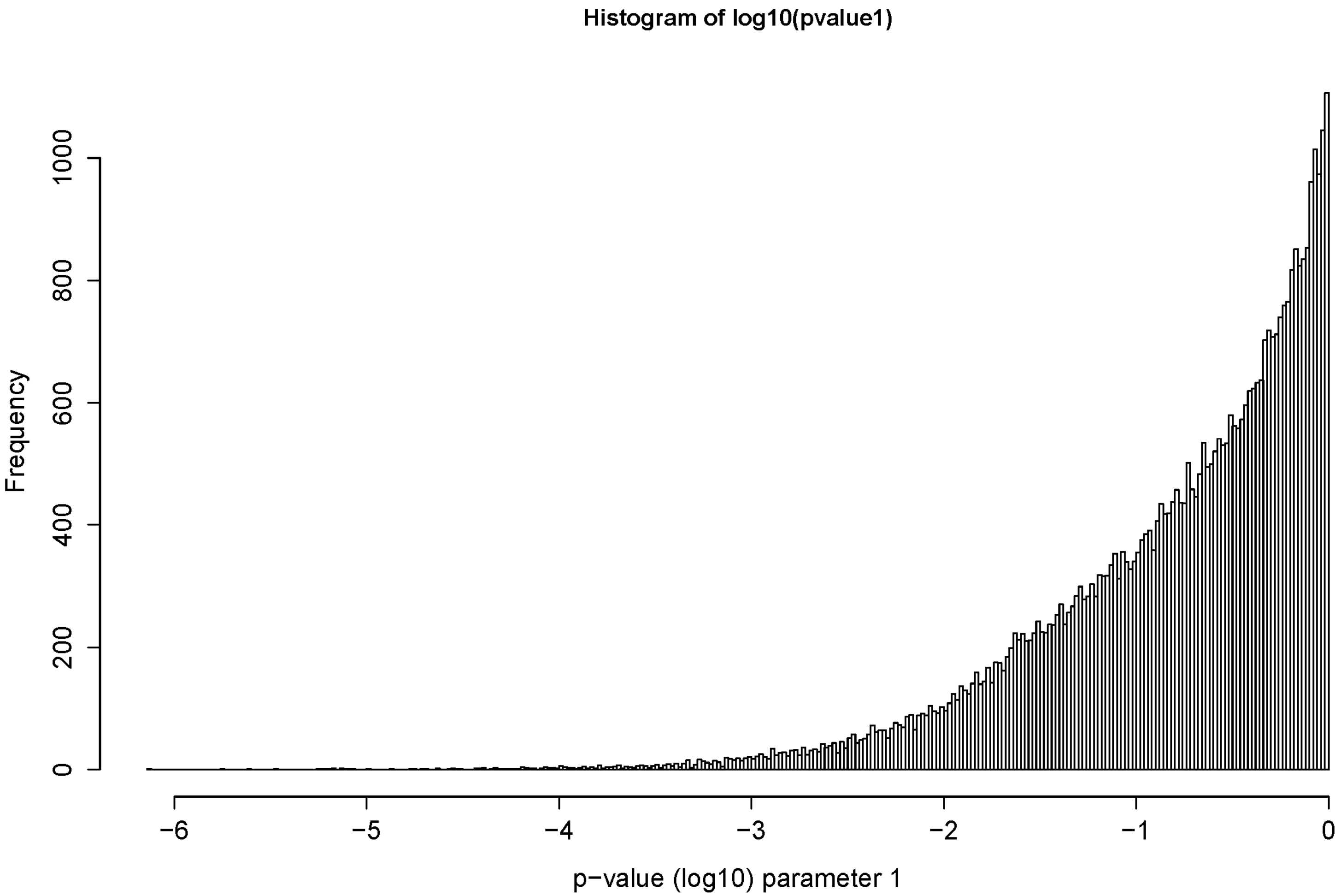
# Example: Gene expression in SI-NETs

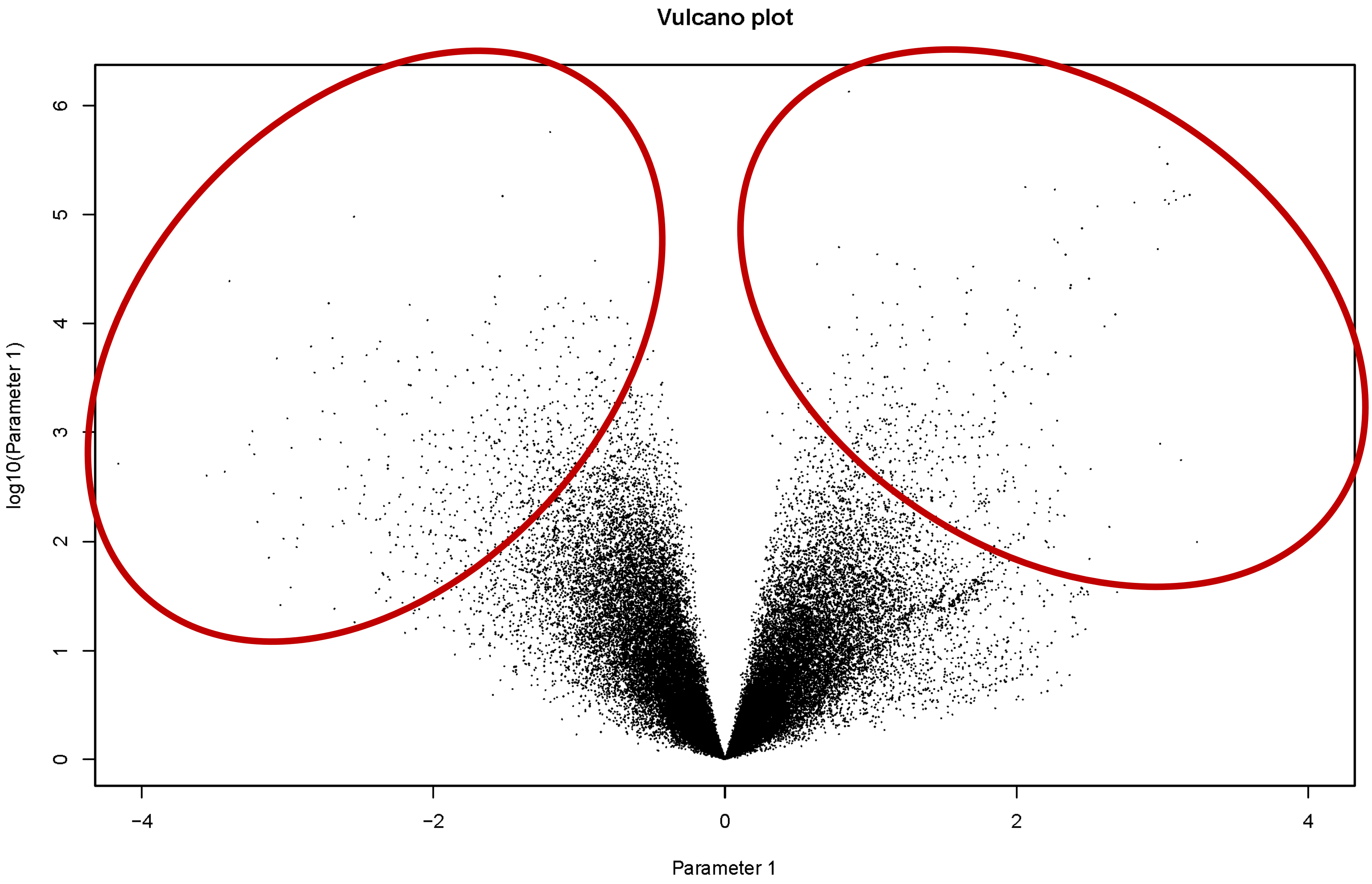


# Example: Gene expression in SI-NETs



# Example: Gene expression in SI-NETs




Vulcano plot
log10(Parameter 1)
0
1
2
3
4
5
6
−4
−2
0
2
4
Parameter 1

# Example: Gene expression in SI-NETs

The extended model is

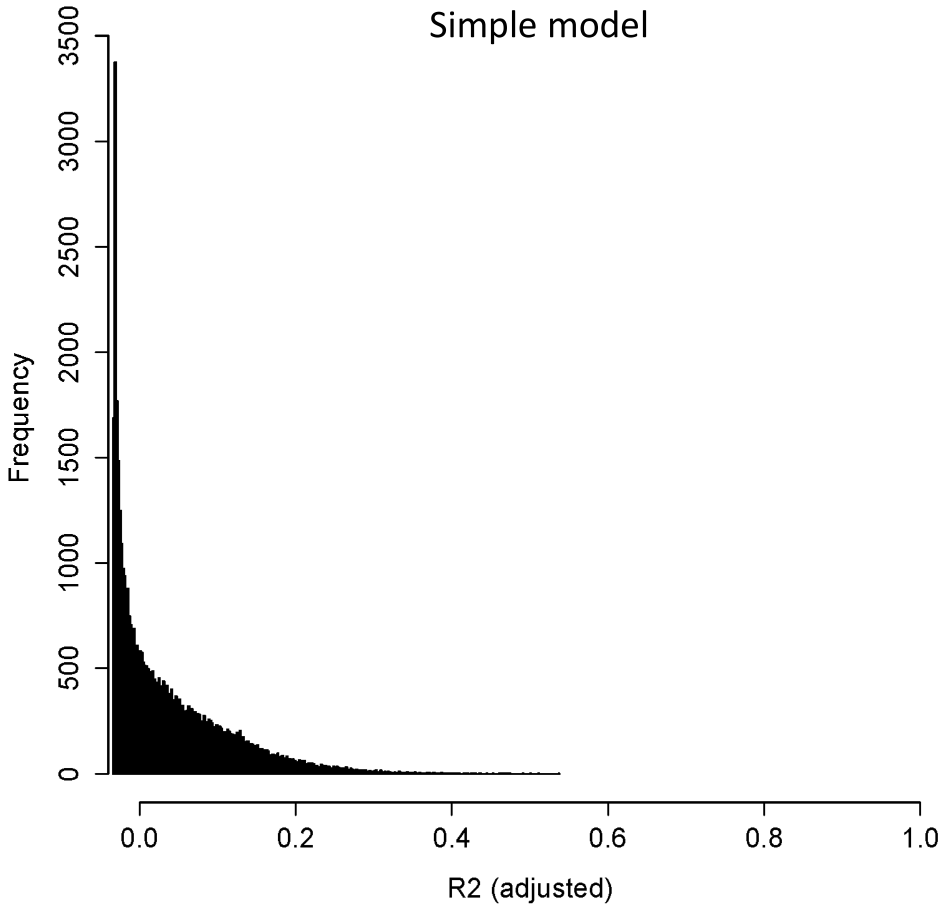
$$Y_{i,j} = \beta_{0,i} + \beta_{1,i}x_{1,j} + \beta_{2,i}x_{2,j} + \beta_{3,i}x_{3,j} + \varepsilon_{i,j}.$$

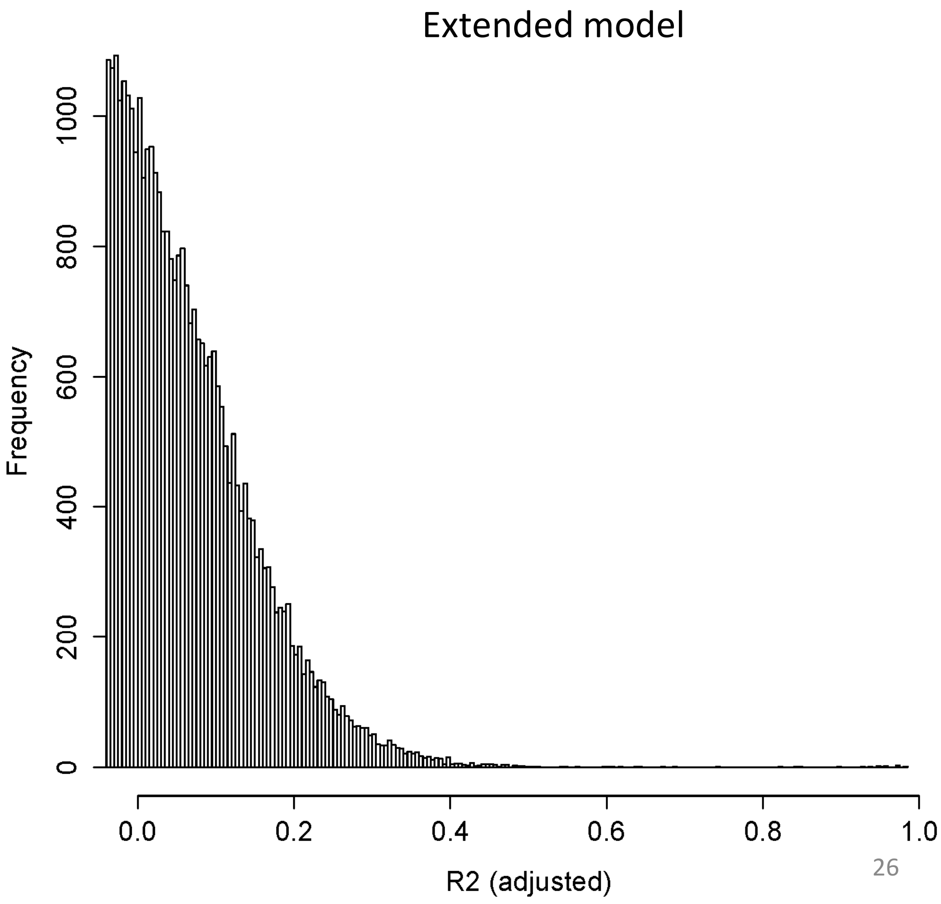
**Questions**

- Is the extended model better in describing the gene expression?
- Does the extended model result in more genes that sigificantly differentially expressed betweenthe ’more aggressive’ and ’less aggressive’ patient groups?
- Are there any genes where the expression is significantly associated with gender and/or age.

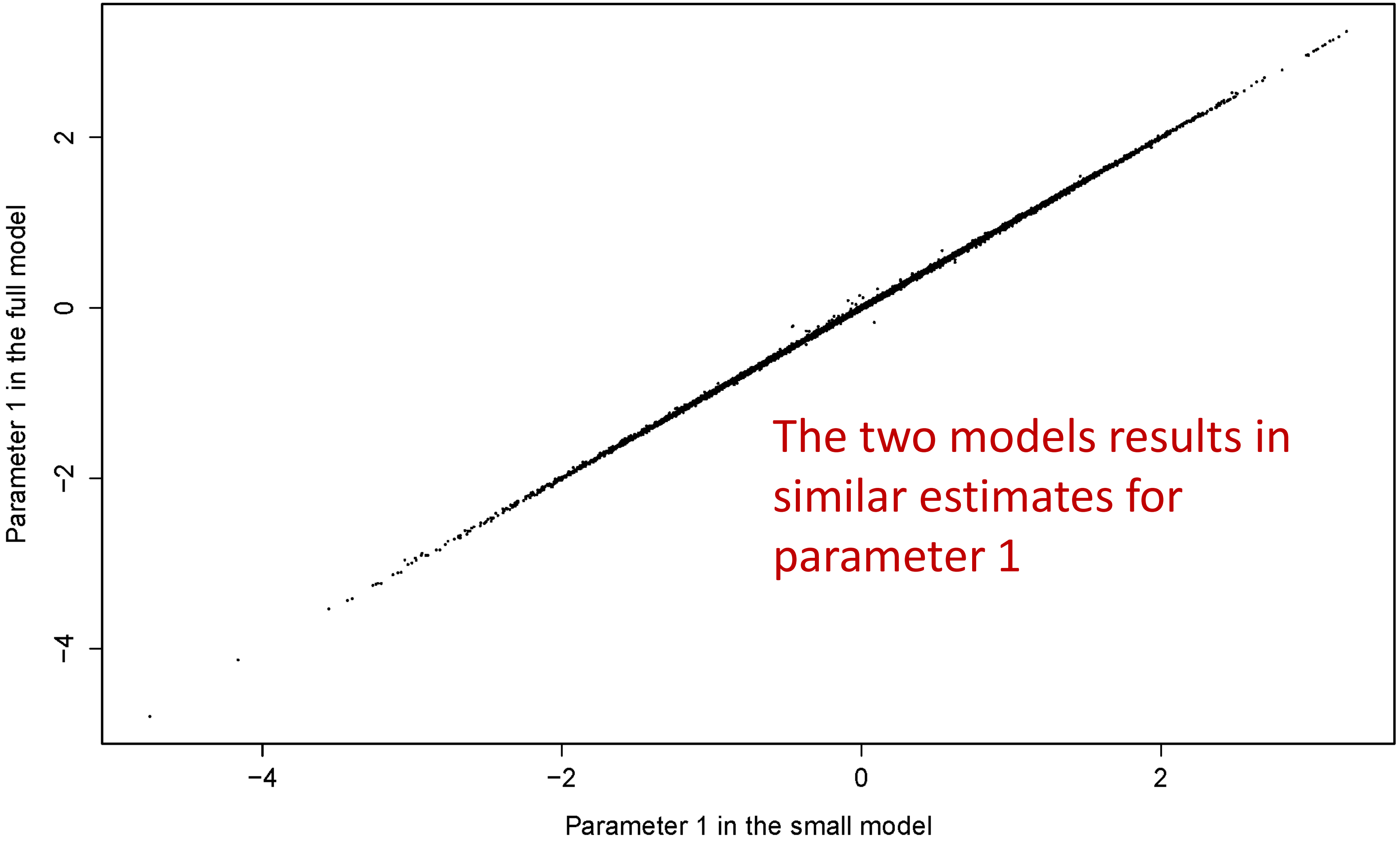
# Is the extended model better in describing the gene expression?

# Comparison of the adjusted $R^2$

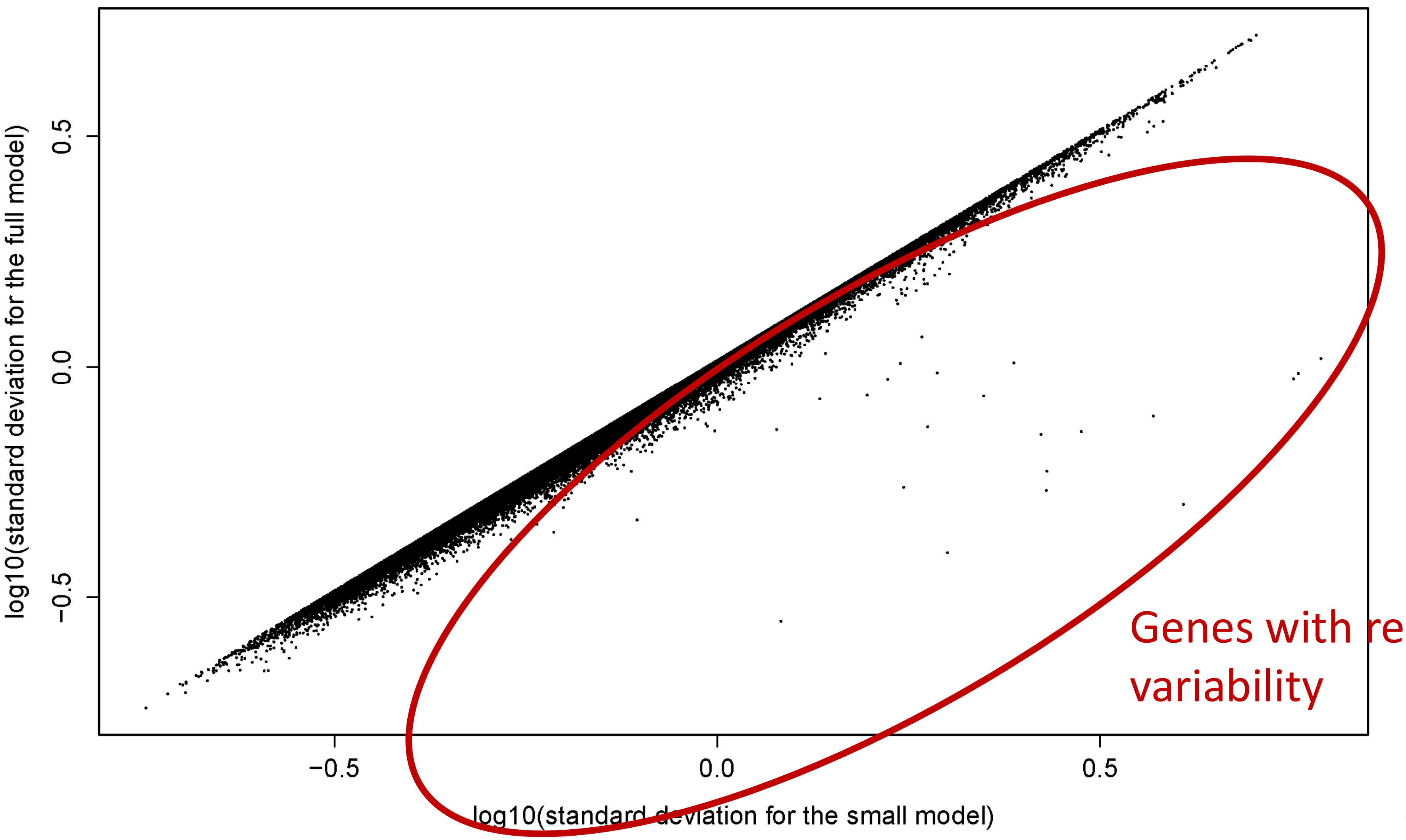




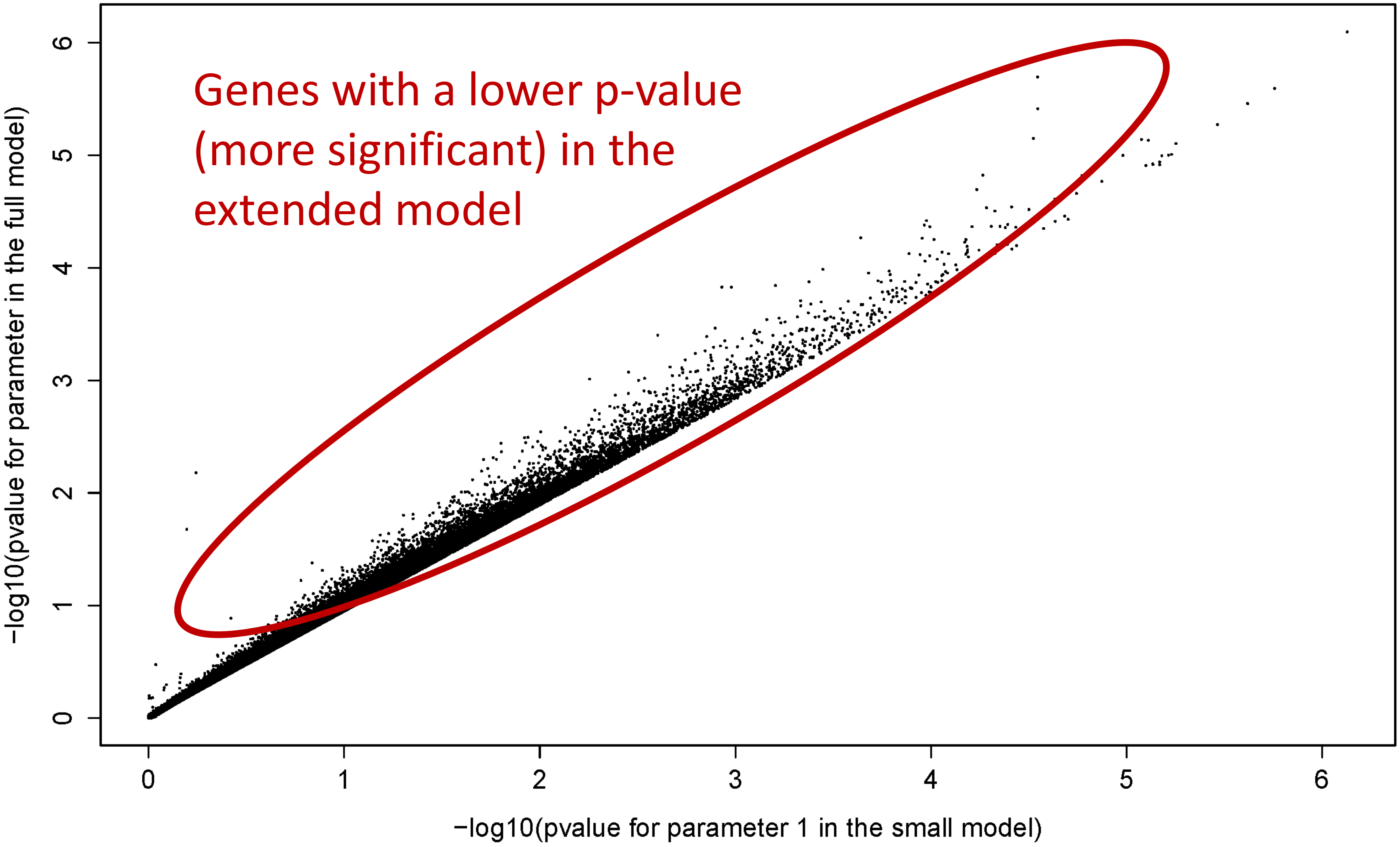
# Does the extended model result in more significantly differentially expressed genes?



The two models results in similar estimates for parameter 1

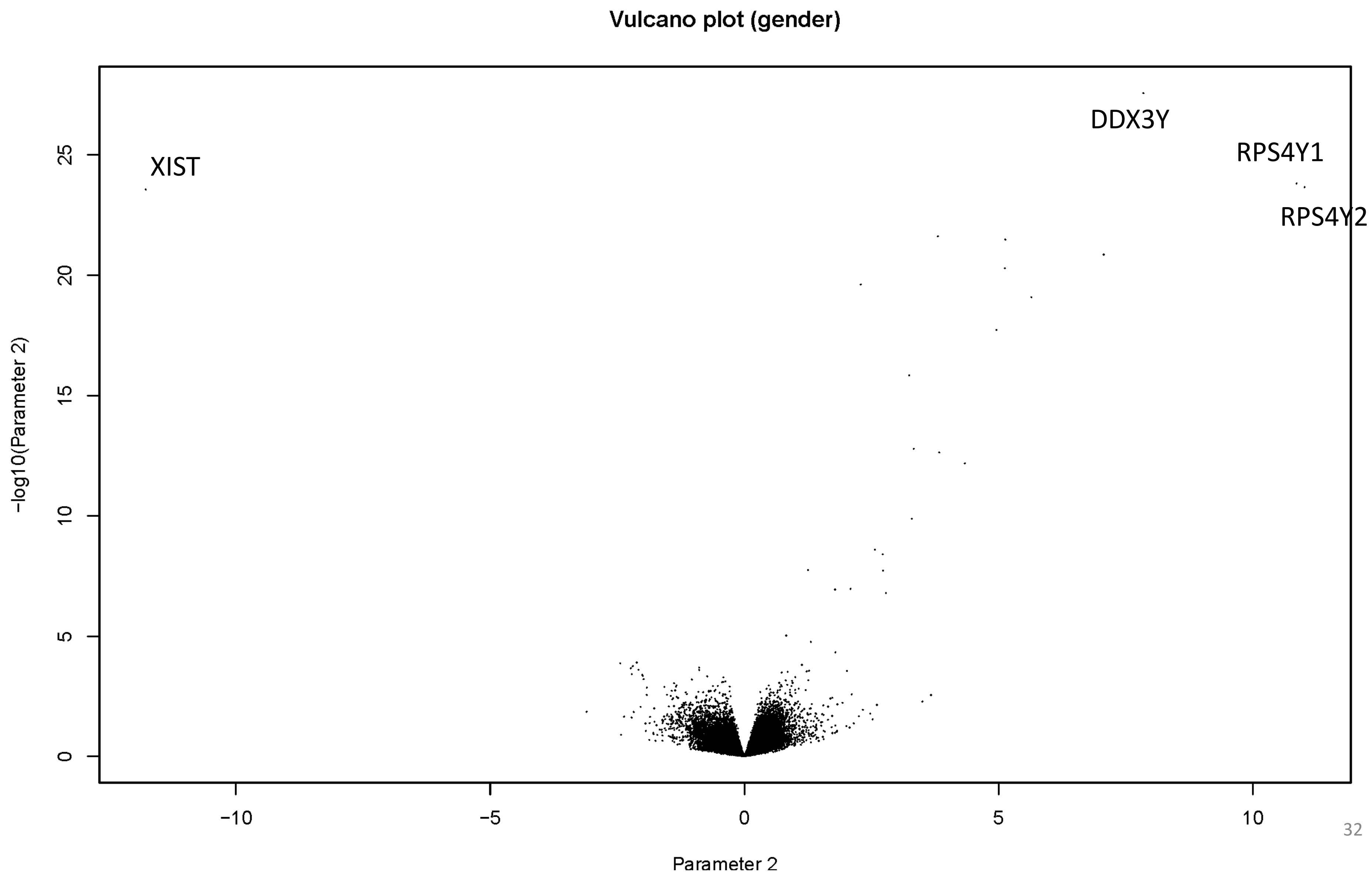


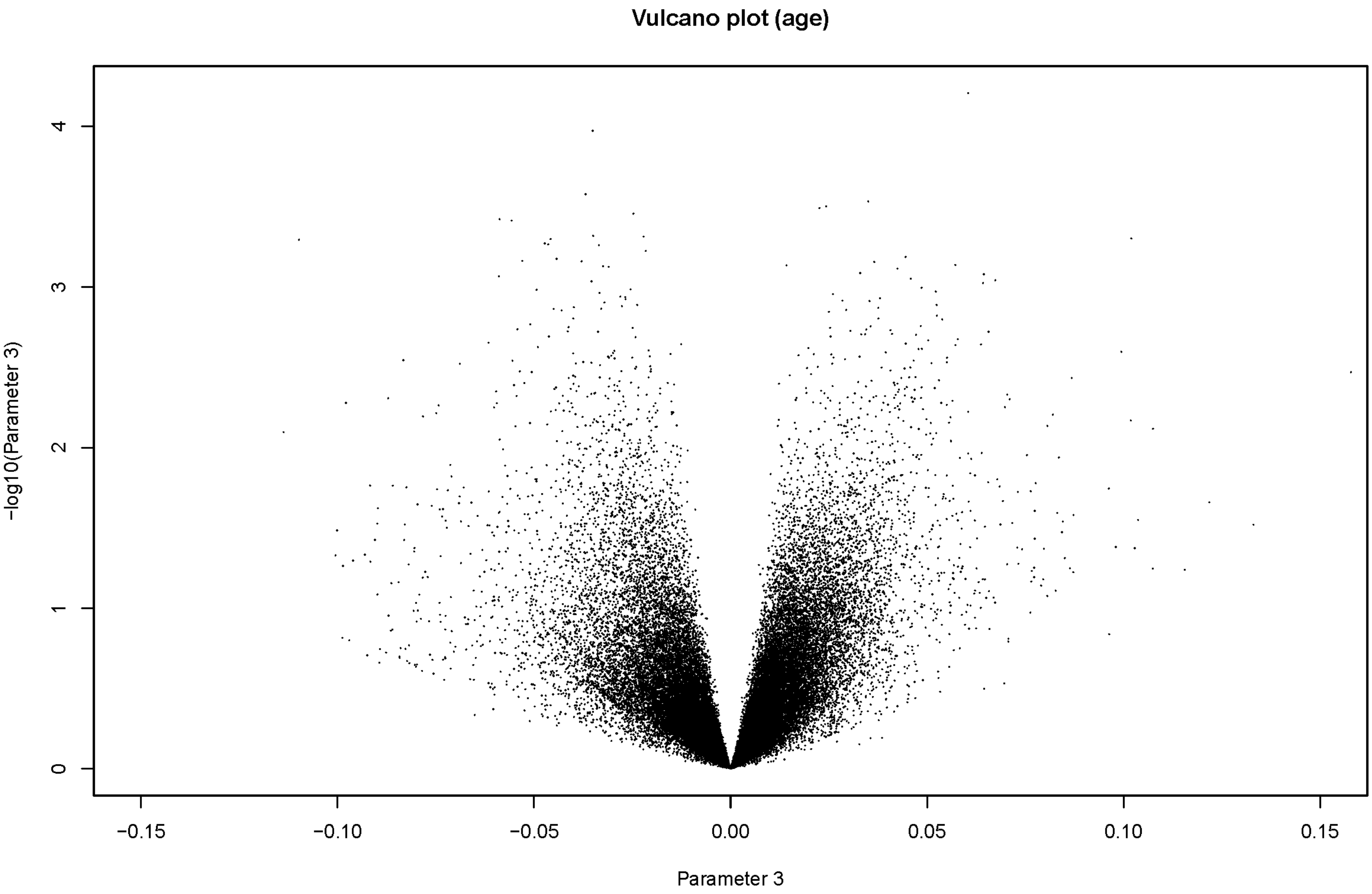
Genes with reduced variability



Genes with a lower p-value (more significant) in the extended model

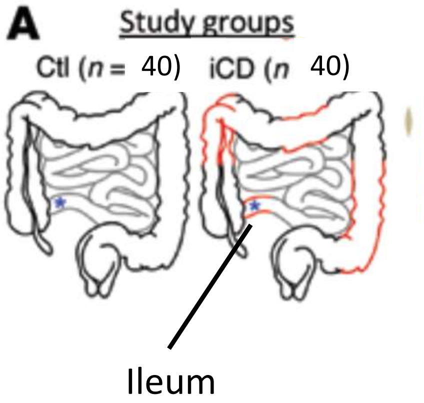
# Are there any genes where the expression is significantly associated with gender and/or age?

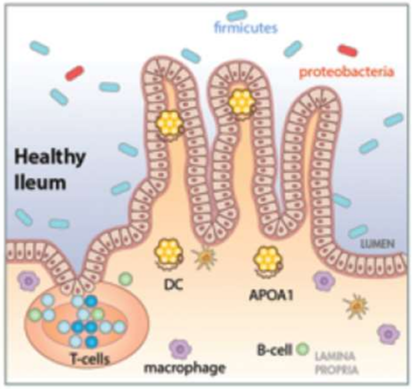

Vulcano plot (gender)
XIST
DDX3Y
RPS4Y1
RPS4Y2
-log10(Parameter 2)
0
5
10
15
20
25
Parameter 2
−10
−5
0
5
10

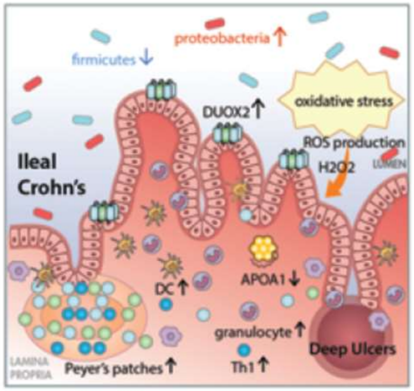

Vulcano plot (age)
-log10(Parameter 3)
0
1
2
3
4
−0.15
−0.10
−0.05
0.00
0.05
0.10
0.15
Parameter 3

# Introduction to computer exercise 3

Gene expression analysis of patients with and without irritated bowel disease (IBD). The aim is to understand the mechanisms behind Crohn's disease among children.



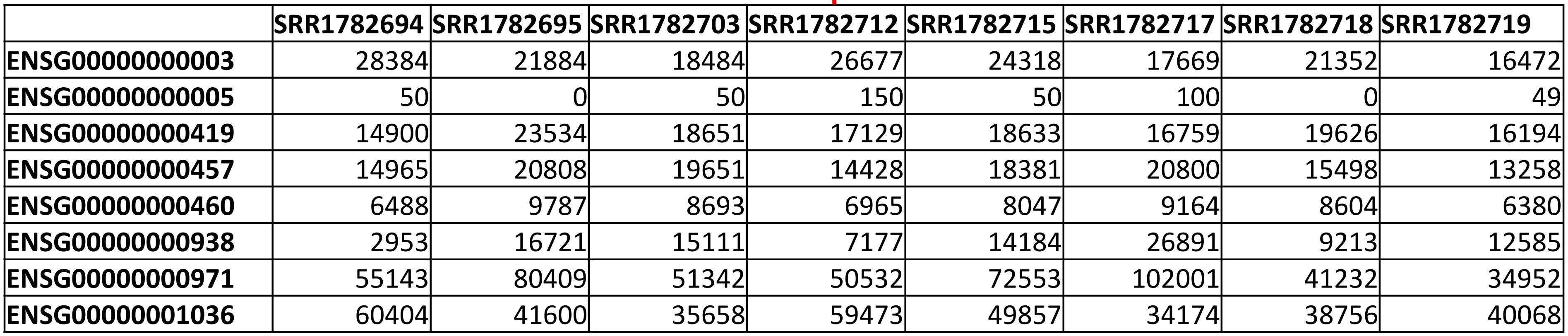




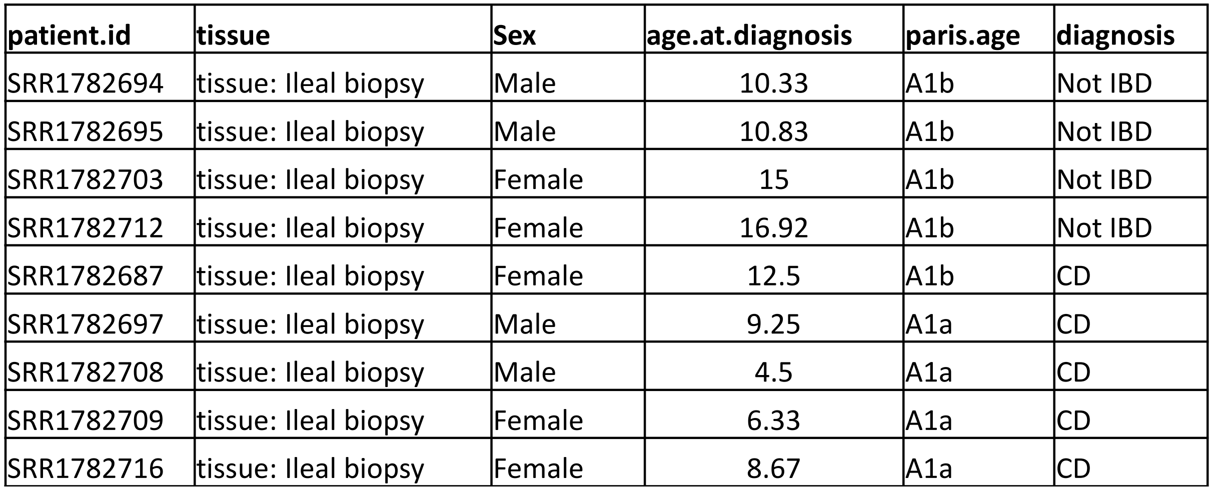
# Introduction to computer exercise 3

Samples

Genes



| | SRR1782694 | SRR1782695 | SRR1782703 | SRR1782712 | SRR1782715 | SRR1782717 | SRR1782718 | SRR1782719 |
|---|---|---|---|---|---|---|---|---|
| **ENSG00000000003** | 28384 | 21884 | 18484 | 26677 | 24318 | 17669 | 21352 | 16472 |
| **ENSG00000000005** | 50 | 0 | 50 | 150 | 50 | 100 | 0 | 49 |
| **ENSG00000000419** | 14900 | 23534 | 18651 | 17129 | 18633 | 16759 | 19626 | 16194 |
| **ENSG00000000457** | 14965 | 20808 | 19651 | 14428 | 18381 | 20800 | 15498 | 13258 |
| **ENSG00000000460** | 6488 | 9787 | 8693 | 6965 | 8047 | 9164 | 8604 | 6380 |
| **ENSG00000000938** | 2953 | 16721 | 15111 | 7177 | 14184 | 26891 | 9213 | 12585 |
| **ENSG00000000971** | 55143 | 80409 | 51342 | 50532 | 72553 | 102001 | 41232 | 34952 |
| **ENSG00000001036** | 60404 | 41600 | 35658 | 59473 | 49857 | 34174 | 38756 | 40068 |



| patient.id | tissue | Sex | age.at.diagnosis | paris.age | diagnosis |
|---|---|---|---|---|---|
| SRR1782694 | tissue: Ileal biopsy | Male | 10.33 | A1b | Not IBD |
| SRR1782695 | tissue: Ileal biopsy | Male | 10.83 | A1b | Not IBD |
| SRR1782703 | tissue: Ileal biopsy | Female | 15 | A1b | Not IBD |
| SRR1782712 | tissue: Ileal biopsy | Female | 16.92 | A1b | Not IBD |
| SRR1782687 | tissue: Ileal biopsy | Female | 12.5 | A1b | CD |
| SRR1782697 | tissue: Ileal biopsy | Male | 9.25 | A1a | CD |
| SRR1782708 | tissue: Ileal biopsy | Male | 4.5 | A1a | CD |
| SRR1782709 | tissue: Ileal biopsy | Female | 6.33 | A1a | CD |
| SRR1782716 | tissue: Ileal biopsy | Female | 8.67 | A1a | CD |

# Summary

- Linear models are highly flexible statistical tools that can be used to analyze gene expression data from RNA-seq experiments.
- Linear models are implemented in R under the **lm** function.
- Each gene is analyzed separately – this results in a large number of linear models, coefficients and p-values.
- Visualization of the results for all gene simultaneously can be used to verify the model and identify genes that are differentially expressed.
- Adding additional factors may increase the fit of the model, reduce the variability and increase the number of significant genes.

# Example 1: Gene expression in SI-NETs

RNA-seq was used to characterize the expression of 45,015 transcripts in small intestinal neuroendocrine tumors (SI-NET). SI-NETs are divided in to two groups depending on its aggressiveness.

**Aim**: Identify genes that are differentially expressed between the groups

**Data**: 33 patients were included in the study of which 23 were from the less aggressive group and 10 from the more aggressive group. For each patient, a biopsy was sampled, the RNA extracted, sequenced using standard protocols, preprocessed and transformed and normalized so that the resulting data is approximately normal distributed. For each patient, the age and gender was also recorded.

# Example 1: Gene expression in SI-NETs

Let $Y_{i,j}$ is the expression of gene *i* in patient *j* and define the independent categorical variable $x_{1,j}$ as

$$x_{1,j} = \begin{cases} 1 & \text{if patient } j \text{ has a more aggressive tumor,} \\ 0 & \text{if patient } j \text{ has a less aggresive tumor,} \end{cases}$$

A simple linear model can be formulated as

$$Y_{i,j} = \beta_{0,i} + \beta_{1,i} x_{1,j} + \varepsilon_{i,j}$$

where $\beta_{0,i}$ is the base line expression of gene *I* and $\beta_{1,i}$ is the difference in expression in the more aggressive compared to the less aggressive tumor form and $\varepsilon_{i,j}$ is the error for gene *i* and patient *j* (normal).

# Example 1: Gene expression in SI-NETs

The scientists suspected that the age and gender may be influencing the gene expression and an extended model was therefore formulated. Let $x_{2,j}$ be the age of patient *j* and let

$$x_{3,j} = \begin{cases} 1 & \text{if patient } j \text{ is a male,} \\ 0 & \text{if patient } j \text{ is a female.} \end{cases}$$

The extended model can then be formulated as

$$Y_{i,j} = \beta_{0,i} + \beta_{1,i}x_{1,j} + \beta_{2,i}x_{2,j} + \beta_{3,i}x_{3,j} + \varepsilon_{i,j}.$$