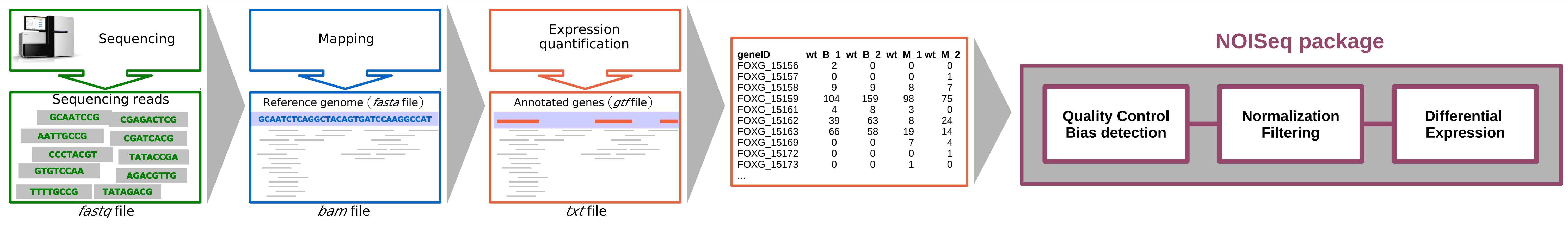
Quality-control, experimental design and FDR controlled differential expression of RNA-seq with the *NOISeq* R package

Sonia Tarazona[1,2], Pedro Furió[1], David Turrà[3], Antonio Di Pietro[3], Alberto Ferrer[2], Ana Conesa[1]

[1] Genomics of Gene Expression Lab, Centro de Investigación Príncipe Felipe, Valencia, Spain
[2] Department of Applied Statistics, Operations Research and Quality, Universitat Politècnica de València, Spain
[3] Department of Genetics, Universidad de Córdoba, Spain

RNA-seq pipeline



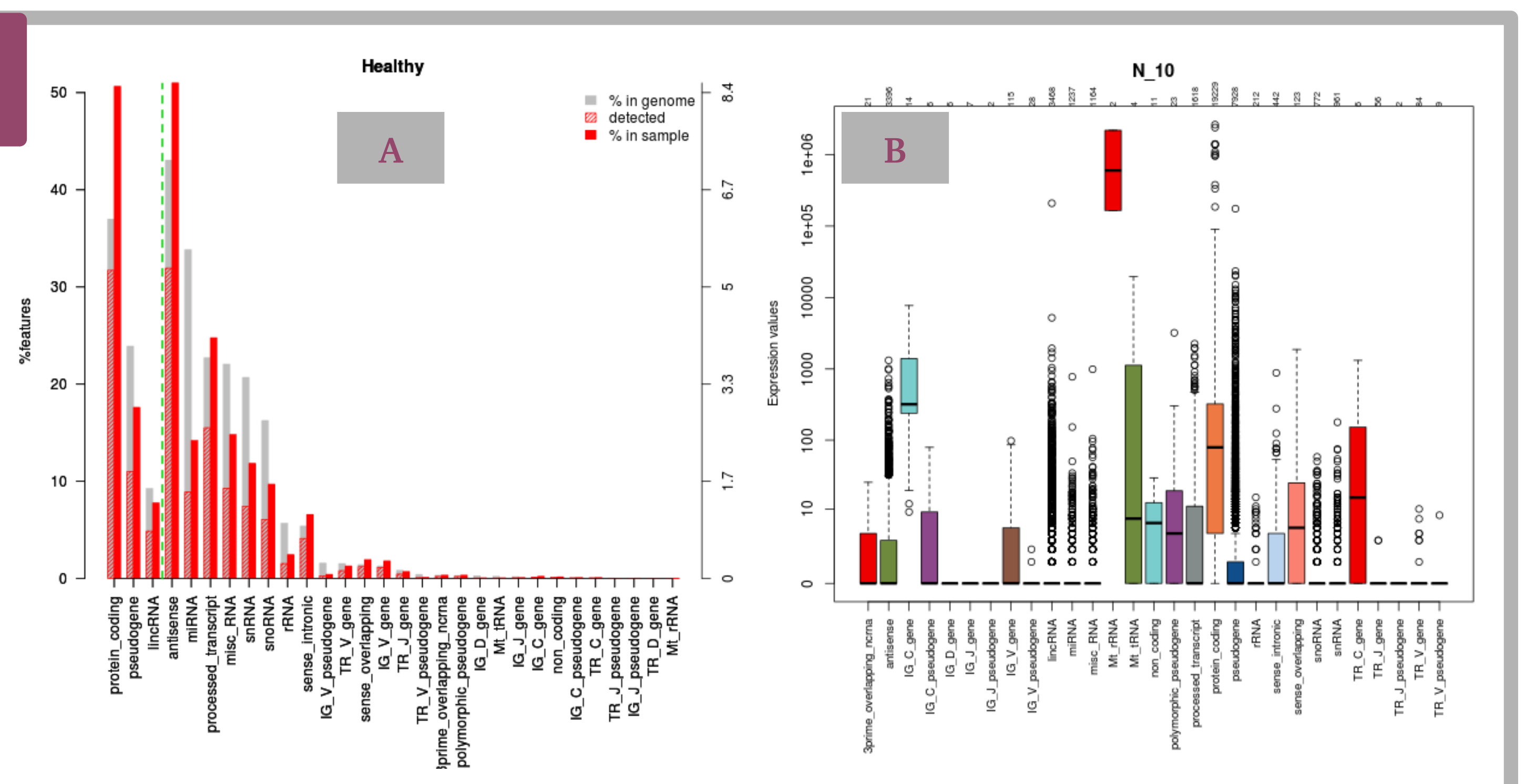
geneID	wt_B_1	wt_B_2	wt_M_1	wt_M_2
FOXG_15156	2	0	0	0
FOXG_15157	0	0	0	1
FOXG_15158	9	9	8	7
FOXG_15159	104	159	98	75
FOXG_15161	4	8	3	0
FOXG_15162	39	63	8	24
FOXG_15163	66	58	19	14
FOXG_15169	0	0	7	4
FOXG_15172	0	0	0	1
FOXG_15173	0	0	1	0
...				

RNA-seq is a powerful technique to study the transcriptional process in the cell. Despite the quality controls at each step of the procedure, technical biases may still be present in the data so it is essential to explore and process read count data prior to differential expression analysis to get reliable results.

Biotype detection

These are exploratory plots that show which type of features are being detected in our samples (**A**) and how expression values (counts per million) are distributed (log-scale) for each type of feature (**B**).

In these examples, we used the biotypes information provided by Ensembl and an experiment that compares healthy and tumoral prostate tissues (Ren *et al.*, 2012).



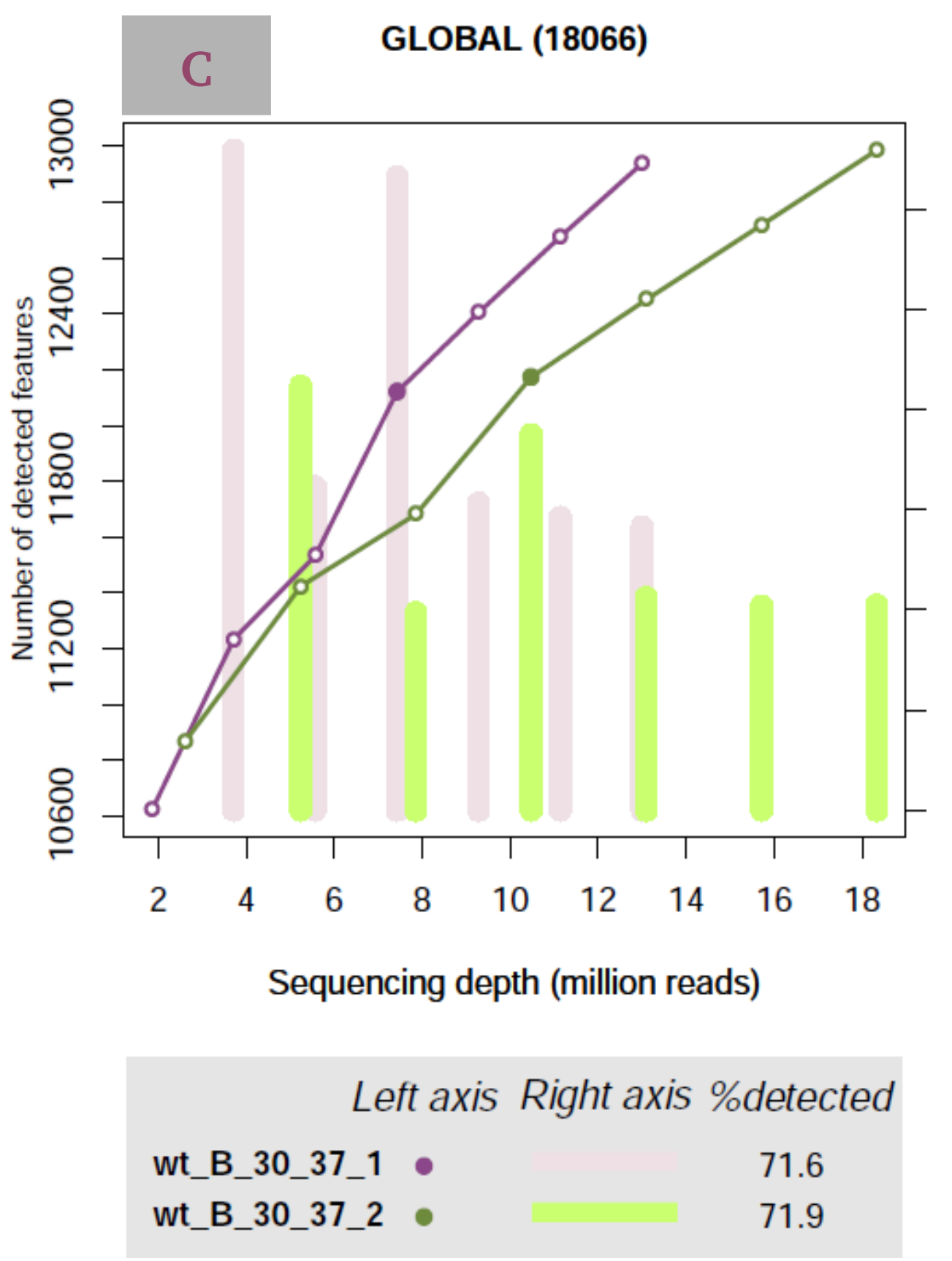
Sequencing depth & Expression quantification

One of the most recurrent questions in RNA-seq is if the number of sequencing reads is enough to study our genome and properly quantify the feature expression. The following plots will help to answer these questions. Data from an experiment with *Fusarium oxysporum* fungi to compare gene expression in blood and minimal medium culture have been used:

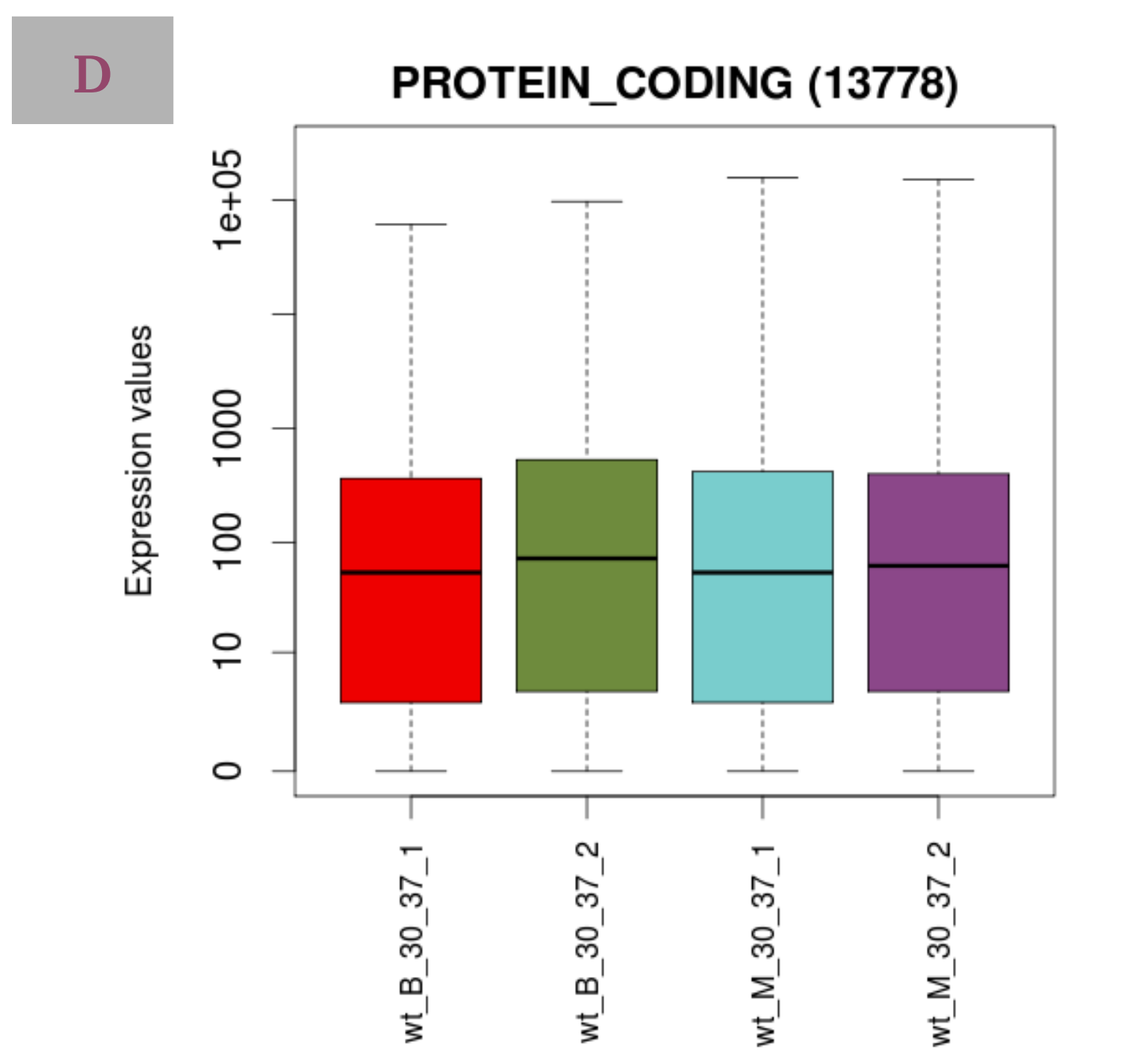
Plot (**C**): The number of detected genes (with counts>0) for different simulated sequencing depths are plotted in the left axis (lines). Solid points represent the real sequencing depth. Right axis (bars) show the number of new detections per each additional million of sequenced reads.

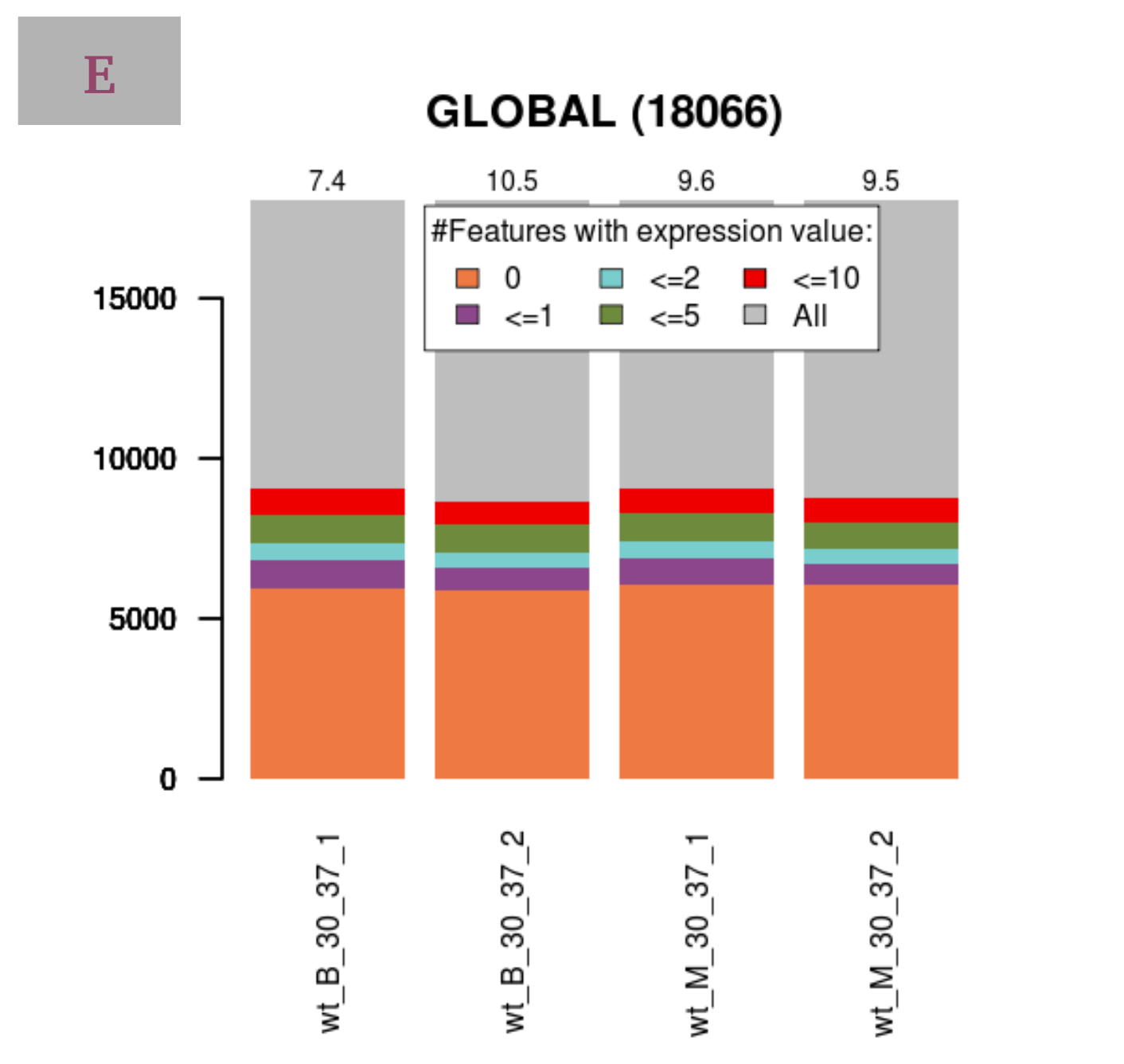
Plot (**D**): Count distribution for protein-coding genes with more than 0 counts in each sample (log-scale).

Plot (**E**): For each sample, it shows the number of features with 0 counts, 1 count or less, etc.



	Left axis	Right axis	%detected
wt_B_30_37_1	●	▬	71.6
wt_B_30_37_2	●	▬	71.9



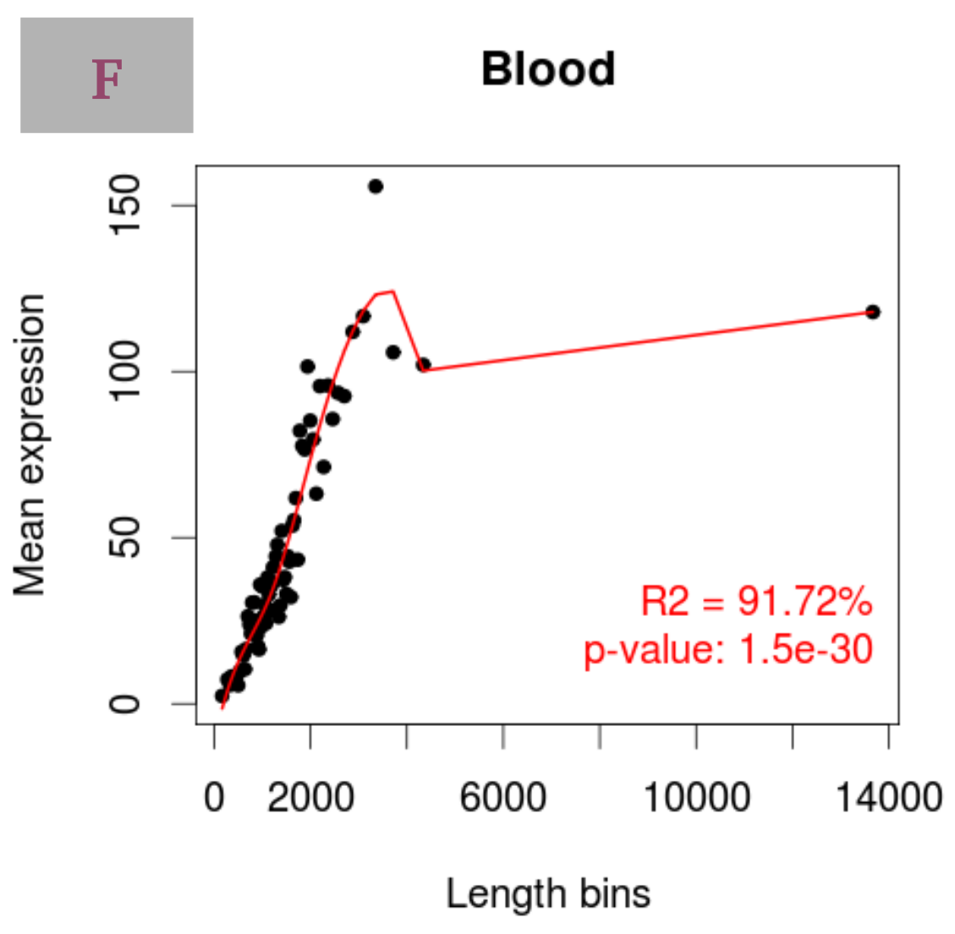


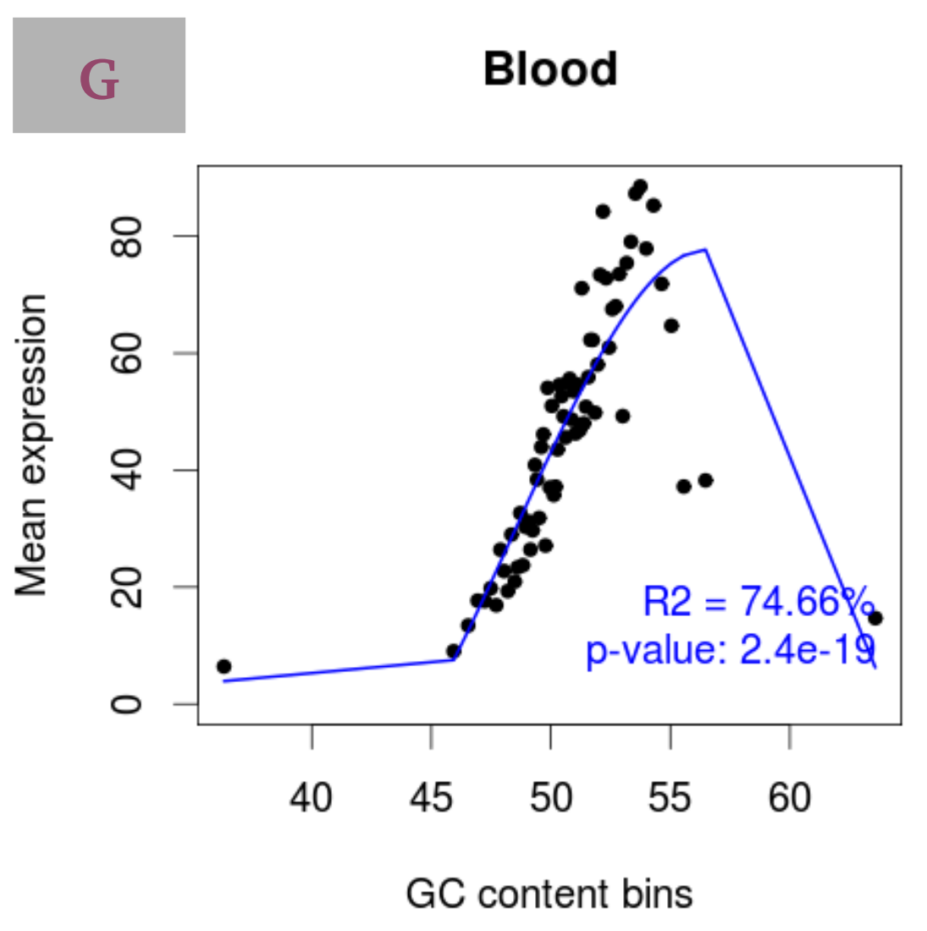
Sequencing bias detection

NOISeq package allows for the detection of the most common RNA-seq bias: length (**F**), GC content (**G**) and RNA composition (**H**), so the user may choose appropriate normalization methods. Both plots and diagnostic tests are available for this purpose.

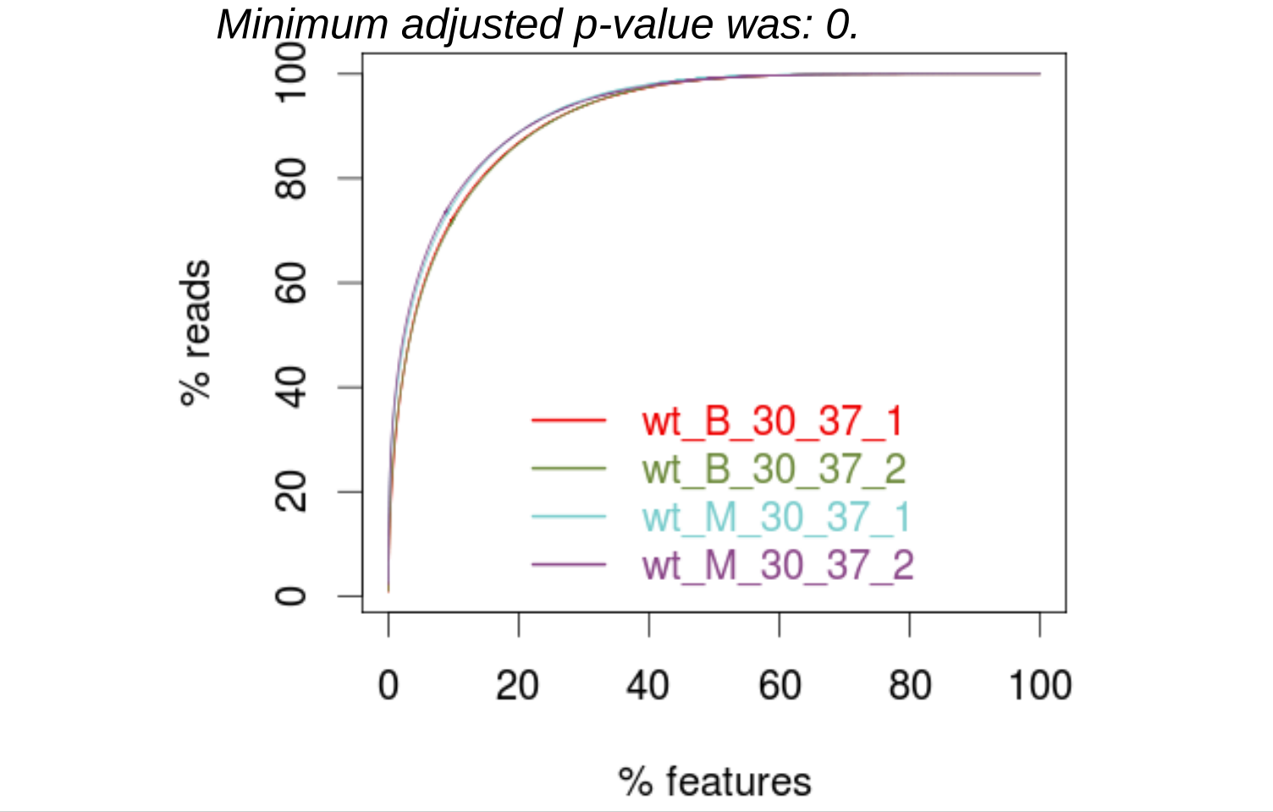
Length (**F**) and GC content (**G**) bias plots show the mean expression (5% trimmed values) per length or GC content bin. Each bin contains 200 values. A cubic spline regression model is fitted and both R^2 value and model p-value are returned. If any tendency is observed, normalization is required.

Figure (**H**) shows whether RNA composition is different for each sample. A Kolmogorov-Smirnov test for each pair of samples is done. If any of the adjusted p-values is significant, it indicates that data should be normalized (e.g. TMM method).





Diagnostic test: FAILED.
Normalization is required to remove this bias.
According to Kolmogorov-Smirnov tests, at least a pair of samples have significantly different distributions.
Minimum adjusted p-value was: 0.



Normalization

NOISeq accepts both raw and normalized data.

The package also includes three normalization methods:

- ➔ **RPKM** (Mortazavi *et al.*, 2008)
- ➔ **Upper Quartile** (Bullard *et al.*, 2010)
- ➔ **TMM** (Robinson & Oshlack, 2010)

Filtering out features with low counts

Excluding features with low counts improves, in general, differential expression results, no matter the method being used, since noise in the data is reduced. However, the best procedure to filter these low count features has not been yet decided nor implemented in the differential expression packages. NOISeq proposes three different methods to filter out features with low counts:

- ➔ **CPM**: A value for the counts per million (CPM) under which a feature is considered to have low counts is set. The cutoff or a condition with *s* samples is CPM x s. Features with sum of expression values below the condition cutoff in all conditions are removed. Also a cutoff for the coefficient of variation per condition may be established to eliminate features with inconsistent expression values.
- ➔ **Proportion test**: For each feature and condition, H_0: $p=p_0$ versus H_1: $p>p_0$ is tested, where p is the feature relative expression and p_0 = CPM/10^6. Features with p-value > 0.05 in all conditions are filtered out.
- ➔ **Wilcoxon test**: Similar procedure but testing H_0: m=0 versus H_1: m>0 (for *s*≥5).

Differential Expression

The differential expression methods included in the package are:

- ➔ Technical replicates: **NOISeq-real** (Tarazona *et al.*, 2011)
- ➔ No replicates: **NOISeq-sim** (Tarazona *et al.*, 2011)
- ➔Biological replicates: **NOISeqBIO**

NOISeqBIO

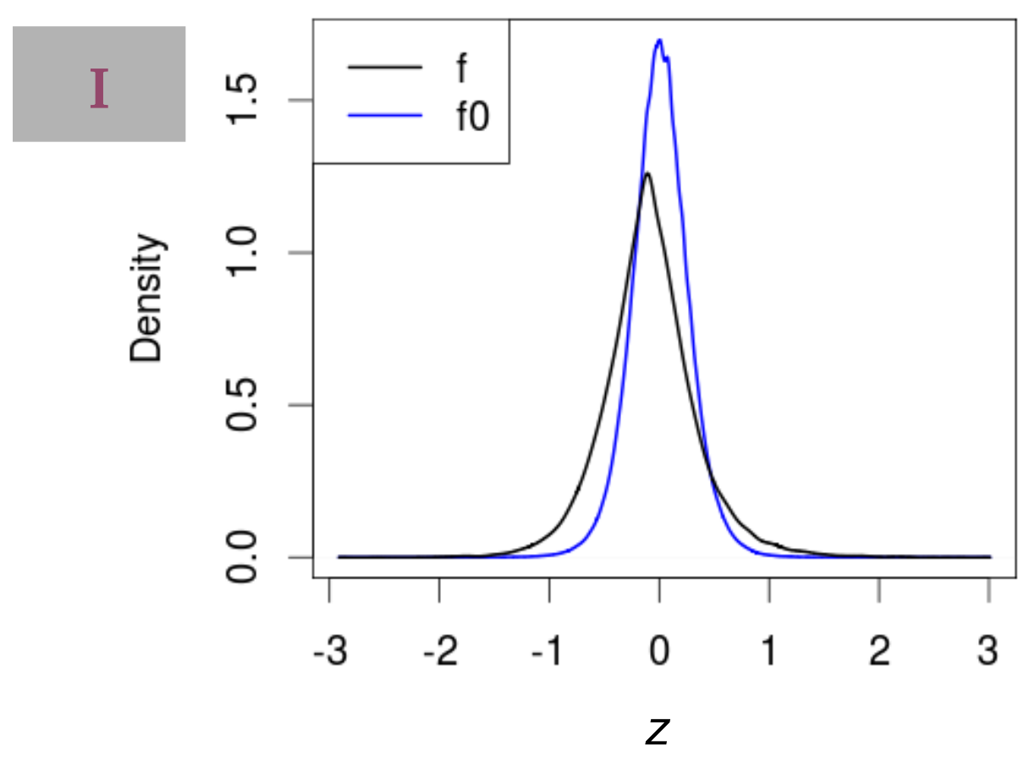
- Signal scores (*Z*) measure the change in expression and take into account not only the difference (*D*) but also the fold-change (*M*) between conditions. *M* and *D* are corrected for the individual biological variability of each feature (*M** and *D**).

$$Z = \frac{M^* + D^*}{2}$$

- Non-parametric method: Noise scores are generated by resampling. Signal (*f*) and noise (f_0) distributions are estimated using a Kernel Density Estimator. Signal distribution can be written as a mixture:

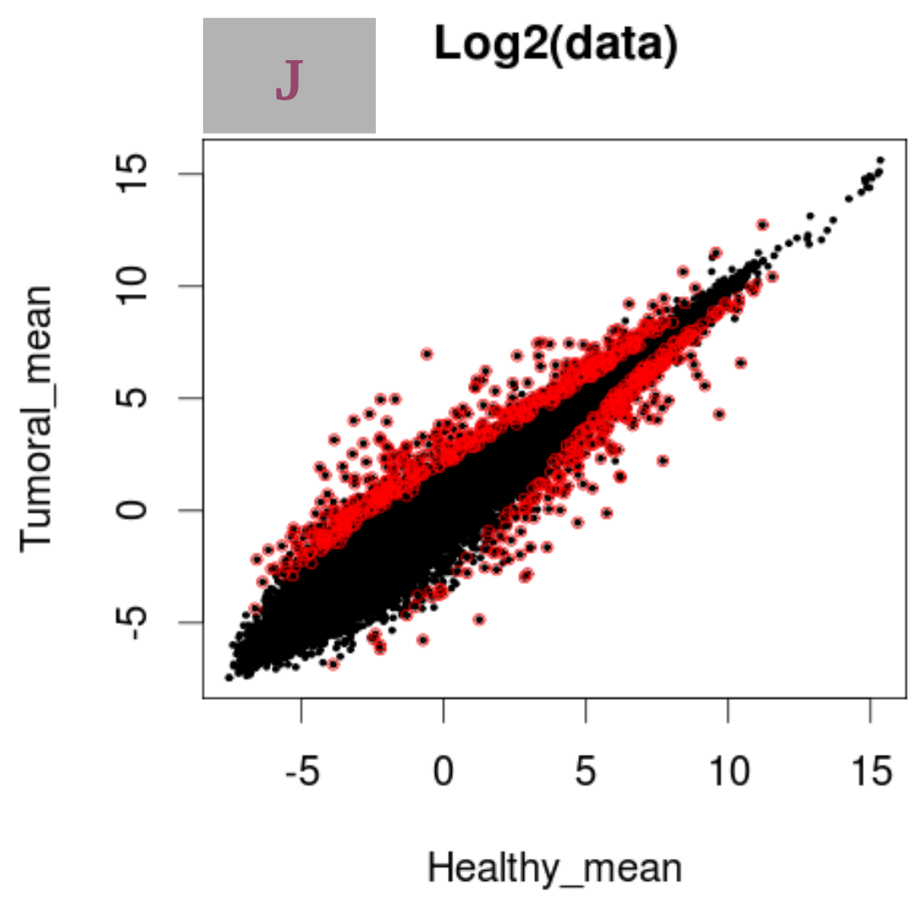
$f(z) = p_0 f_0(z) + p_1 f_1(z)$ (Figure (**I**))

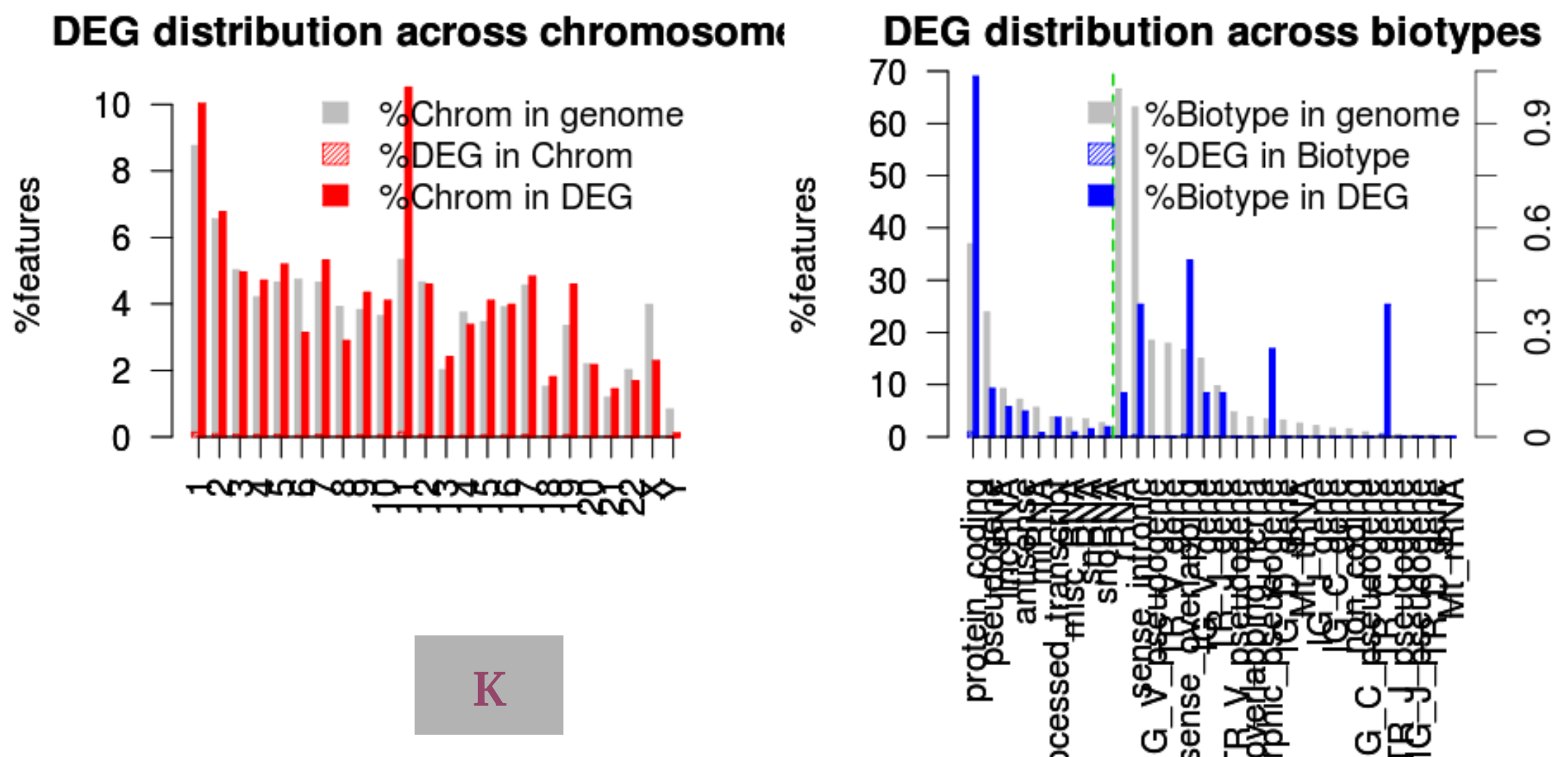
- The probability of differential expression (p_1) for a given feature is estimated from the mixture and is considered to be equivalent to 1-FDR.



Differential expression results

The NOISeq package includes some plots to explore differential expression results. For the prostate cancer dataset, the differentially expressed genes (DEG) were obtained using a probability cutoff of 0,95 (in red in (**J**)). Figure (**K**) shows the proportion of DEG per chromosome and per biotype.

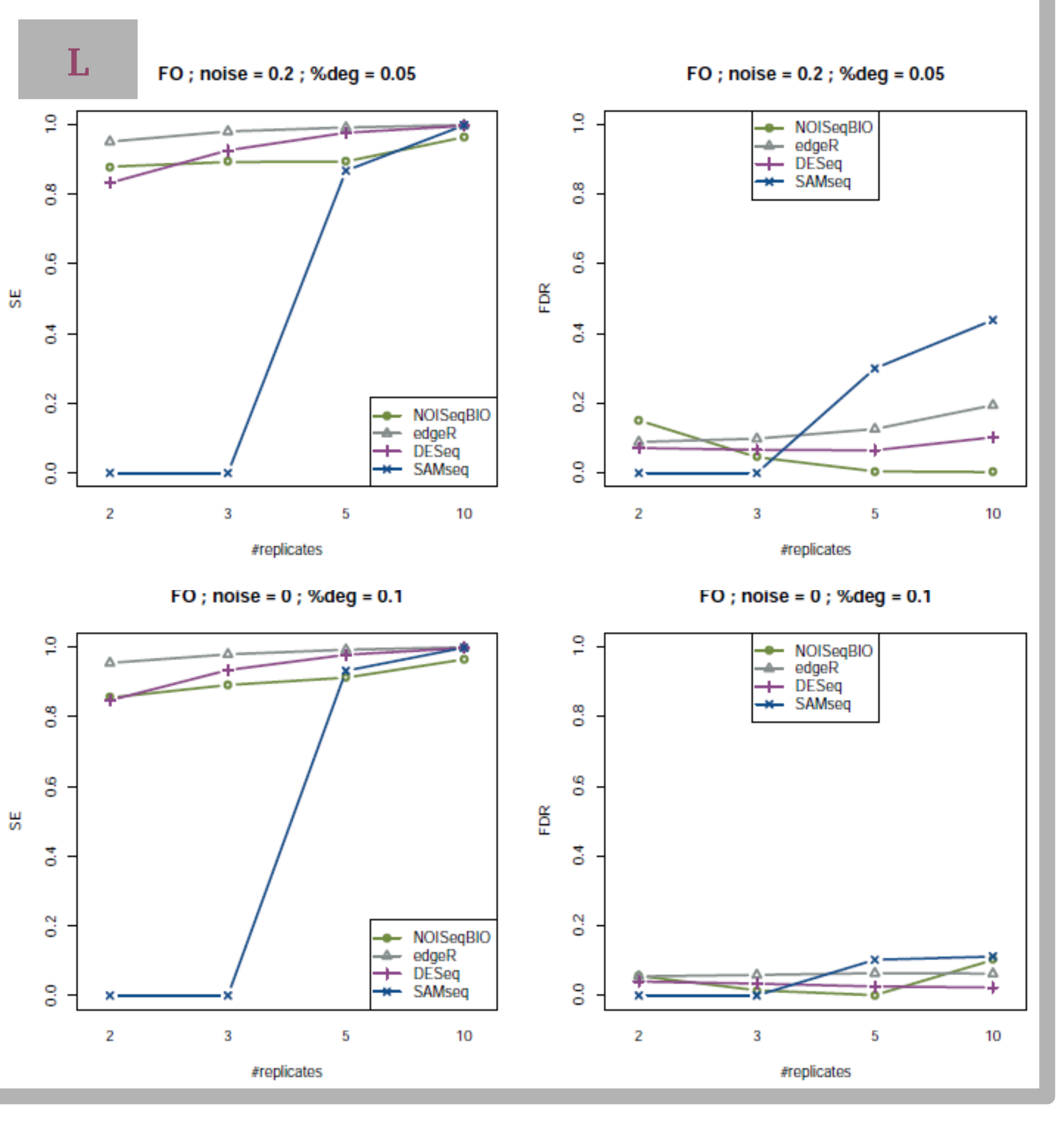




Comparing NOISeqBIO to other differential expression methods

NOISeqBIO was compared to the widely used edgeR (Robinson *et al.*, 2010) and DESeq (Anders & Huber, 2010), and to another non-parametric method, SAMseq (Li & Tibshirani, 2011). The performance of the methods was obtained on datasets simulated under a variety of experimental scenarios considering different noise levels, proportion of DEG or number of replicates.

In Figure (**L**), it can be seen that NOISeqBIO is efficient at controlling the False Discovery Rate (FDR), while maintaining an acceptable sensitivity (SE).


L
FO ; noise = 0.2 ; %deg = 0.05
FO ; noise = 0 ; %deg = 0.1
SE
FDR
NOISeqBIO
edgeR
DESeq
SAMseq
#replicates