

Whole transcriptome sequencing data analysis workshop

P-value correction methods and plots:
ROC, FDR, FPs and precision-recall

Simone Tiberi, University of Zurich

1-8/02/2019

All models are wrong

George Box:

“Essentially, all models are wrong, but some are useful.”

“Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful.”

David Cox:

“The very word model implies simplification and idealization. The idea that complex physical, biological or sociological systems can be exactly described by a few formulae is patently absurd. The construction of idealized representations that capture important stable aspects of such systems is, however, a vital part of general scientific analysis and statistical models.”

Andrew Gelman:

“The saying, “all models are wrong”, is helpful because it is not completely obvious...This is a simple point...But, the trouble is, many people don't realize that all models are wrong.”

1. p-value correction

2. Plots

The p-value

- First of all, for a test we need a system of hypothesis, H_0 vs H_1 , and a test statistic.
- A p-value is the probability of generating, under H_0 , a value of the statistic as extreme (or more) as the one observed in our sample.
- A p-value of 0.05 means that, under H_0 , there is a 5% probability of generating a value as extreme (or more) than the observed one: so, under H_0 , every 100 tests, on average, 5 will be false positives (FPs) at the 5% significance threshold.
- If we perform 20,000 tests (one per gene) and H_0 is true for all of them, on average 1,000 will be FPs.
- Given a p-value, we can separate our features (genes, transcripts, exons, etc...) in “significant” (p-value $< \alpha$) and “non-significant” (p-value $> \alpha$), where typically $\alpha = 0.05$ or 0.01.
- Features can also be separated according to other statistics, e.g., the fold-change (FC) between conditions.

Definitions

Test result	Real state:		
	Positive	Negative	
Significant	TP	FP	S
Non-significant	FN	TN	NS
	P	N	

- True positive (TPs) are positive cases that are (correctly) detected by a test (i.e., significant p-value).
- True negatives (TNs) are negative cases that are (correctly) not detected by a test (i.e., non-significant p-value).
- False positive (FPs) are negative cases that are (mistakenly) detected by a test (i.e., significant p-value).
- False negatives (FNs) are positive cases that are (mistakenly) not detected by a test (i.e., non-significant p-value).

Sensitivity and Specificity

Test result	Real state:		
	Positive	Negative	
Significant	TP	FP	S
Non-significant	FN	TN	NS
	P	N	

- The true positive rate (TPR), also called Sensitivity, power or recall, measures the proportion of actual positives that are correctly identified as such: $TPR = TP/P = TP/(TP + FN)$.
- The true negative rate (TNR), also called Specificity, measures the proportion of actual negatives that are correctly identified as such: $TNR = TN/N = TN/(TN + FP)$.
- The type I error, false positive rate (FPR) or 1 - Specificity, is the proportion of negatives wrongly classified as significant.
- The type II error, false negative rate (FNR) or 1 - Sensitivity, is the proportion of positives wrongly classified as non-significant.

FWER & FDR

- Alternative ways of selecting the significant genes (w.r.t. the p-value):
 - ▶ family wise error rate (**FWER**);
 - ▶ false discovery rate (**FDR**).
- The family wise error rate (FWER) is the probability of making at least 1 type I error: $\text{FWER} = \Pr(\text{FP} > 0)$.
- FWER is very strict when testing thousands of genes.
- Define the proportion of FPs among significant results as:
 $Q = \text{FP} / S = \text{FP} / \text{TP} + \text{FP}$.
- The FDR is the expected proportion of FPs in the set of significant results: $\text{FDR} = E(Q)$
- Unlike the measures introduced so far, the FDR represents a property of the set of significant features.

Controlling the FDR

- Instead of selecting the significant genes according to the p-value, we can correct the p-value such that:
 - ▶ the ordering of genes is preserved;
 - ▶ the FDR is controlled at the α threshold.
- The Benjamini-Hochberg (BH) correction is the most popular method for controlling the FDR.
- We select the genes whose corrected p-value (often called q.value) is $< \alpha$: therefore for $\alpha = 0.05$, we expect that 5% of the selected genes will be FPs and 95% will be TPs.
- This is typically performed in both differential expression and differential splicing methods.

Independent filtering

- We often filter out genes *a priori*, i.e. before testing.
- Typically we filter genes with very low expression, below a specified threshold.

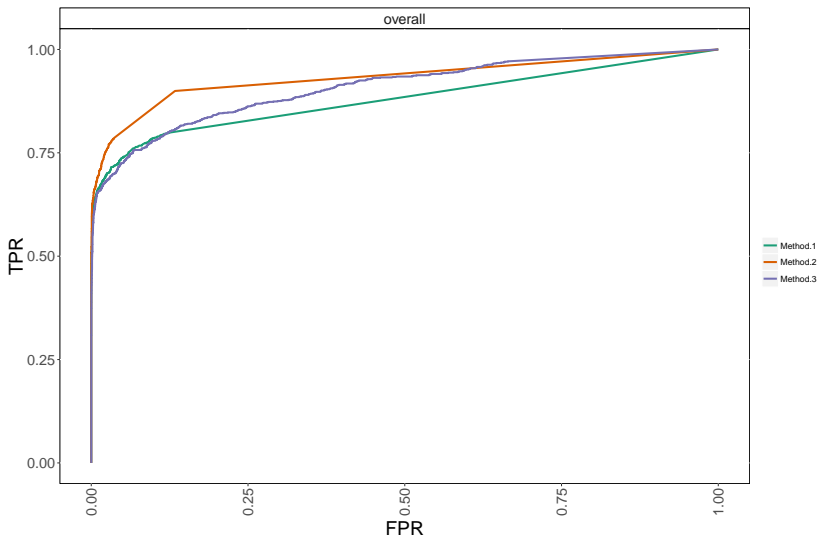
1. p-value correction

2. Plots

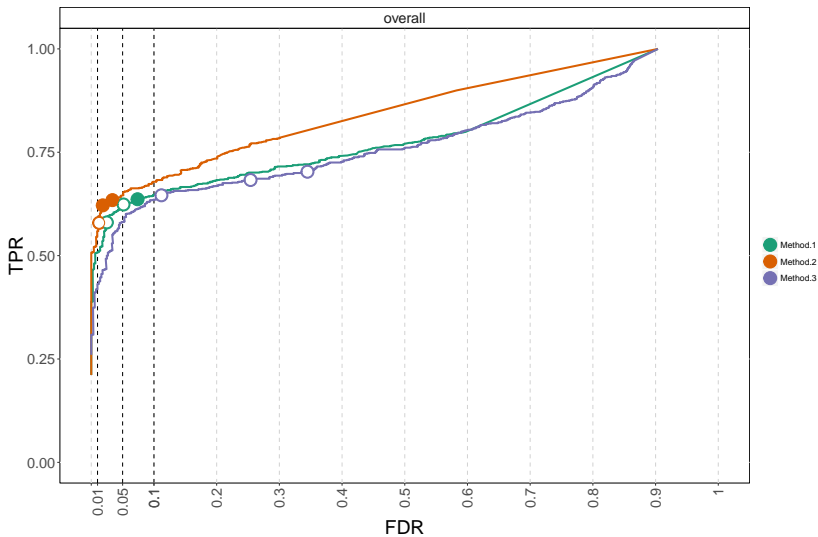
Assessing the performance of a method

- When assessing the performance of methods (DGE, eQTL, DTU, etc...) and comparing them, we mostly investigate several plots.
- This is always done in simulations where we know the **truth**.
- The main plots are:
 - ▶ **ROC**: true positive rate (TPR) vs false positive rate (FPR);
The diagonal line, from (0,0) to (1,1), represents the blind model, which chooses at random.
 - ▶ **FDR**: TPR vs FDR (more interesting than the ROC in this field): important to check if the FDR is controlled at a chosen threshold;
 - ▶ **FPs** vs top significant genes: it tells us how well a model does in the very top genes, which are those we are mostly interested in;
 - ▶ Precision vs recall plot, where:
 - ▶ $\text{precision} = \text{TP} / (\text{TP} + \text{FP})$, i.e. TP over all significant results,
 - ▶ $\text{recall} = \text{TP} / (\text{TP} + \text{FN})$, i.e. TP over (actual) positive results.
- **iCOBRA** is a very useful R package for plotting these curves.

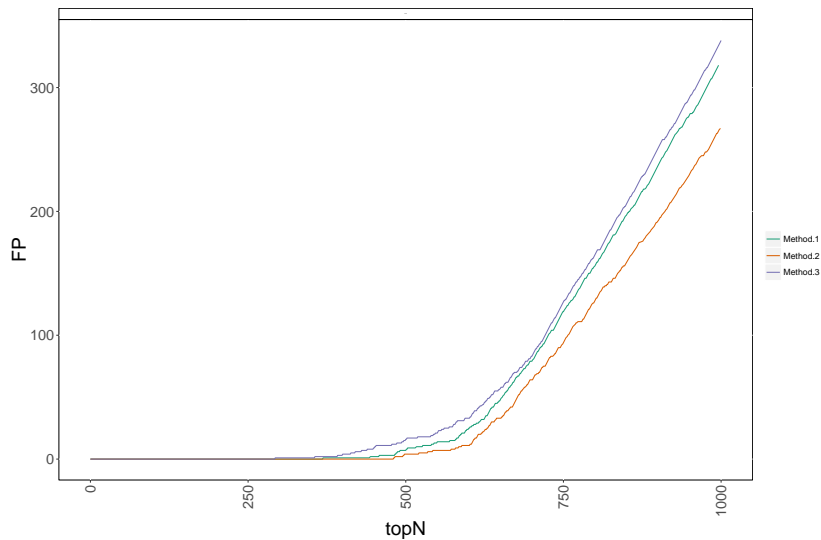
ROC curve



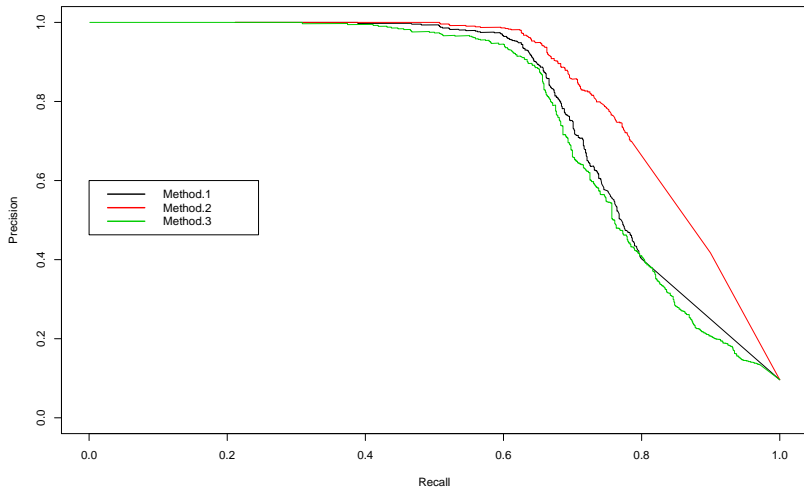
FDR plot



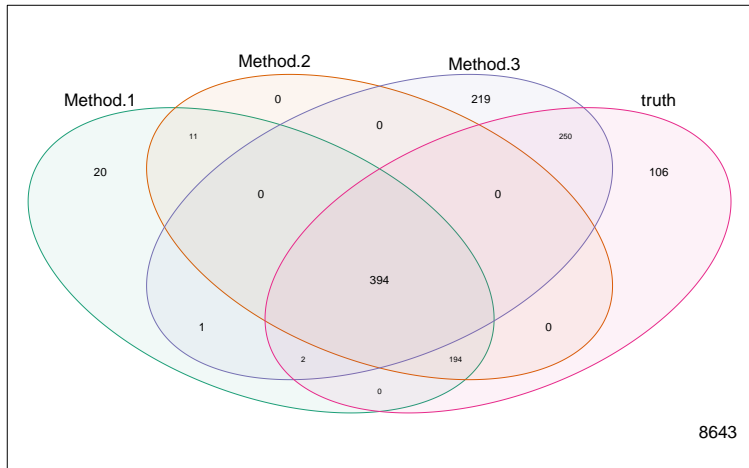
FPs vs top selected genes



Precision recall plot



Venn diagram



What comes after RNA-seq studies?

- The main aim of RNA-seq studies is to identify genes (but also transcripts or pathways) which show interesting characteristics (DGE, differential splicing, etc...): these genes will typically be studied more in depth in further biological analyses.
- This is why we are mostly interested in selecting correctly the top genes: we often want to select a small amount of (very) significant genes, with very few FPs, that might undergo further studies.

Questions?