Differential gene expression analysis

Zuying Chai 柴祖映

Peking University

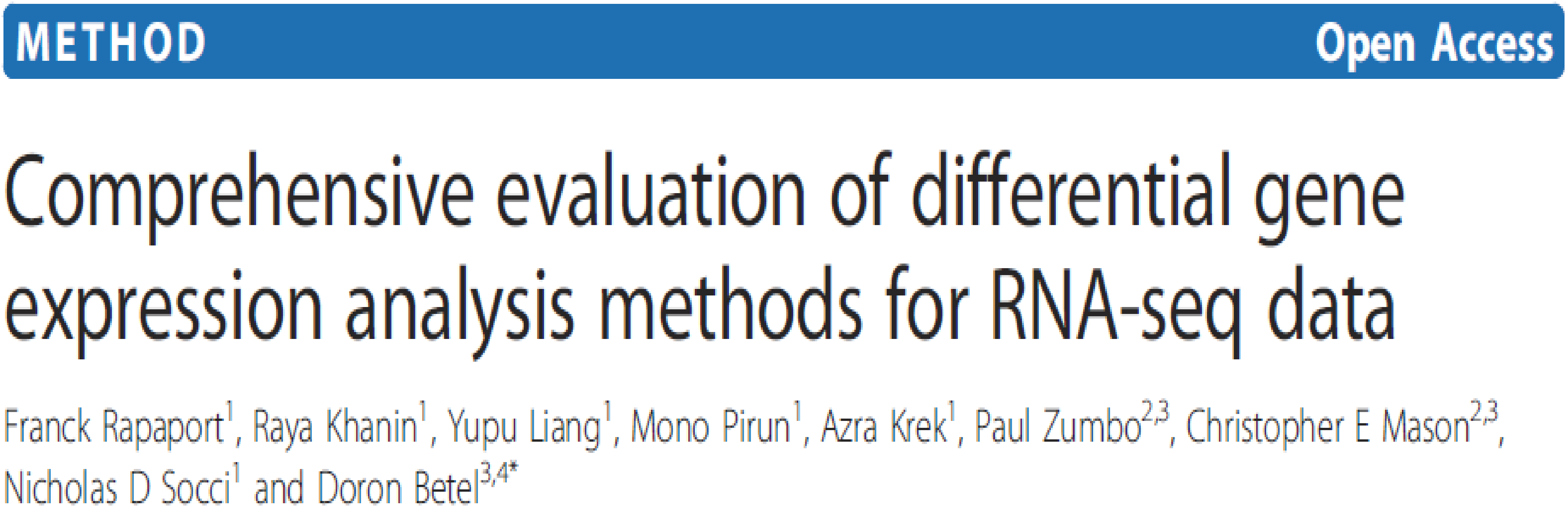
Background

- High-throughput sequencing technology is rapidly becoming the standard method for measuring RNA expression levels (aka RNA-seq).

- One of the main goals of these experiments is to identify the differentially expressed genes in two or more conditions.

Differential gene expression analysis

- 3 steps:
- 1. Normalization of counts
- 2. parameter estimation of the statistical model
- 3. Test for differential gene expression


METHOD
Open Access

Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data

Franck Rapaport[1], Raya Khanin[1], Yupu Liang[1], Mono Pirun[1], Azra Krek[1], Paul Zumbo[2,3], Christopher E Mason[2,3], Nicholas D Socci[1] and Doron Betel[3,4*]

Goal : Comparison of different analysis methods for RNA-seq data from different perspectives.

Such as, Cuffdiff, edgeR, DESeq, PoissonSeq, baySeq, and limma.

Datasets for Research

They used two benchmark datasets:

1 The first is the Sequencing Quality Control (SEQC) dataset, which includes replicated samples of the human whole body reference RNA and human brain reference RNA along with RNA spike-in controls.

2 The second dataset is RNA-seq data from biological replicates of three cell lines that were characterized as part of the ENCODE project.

The measures of their analysis

- The analysis in this paper focused on a number of measures that are most relevant for detection of differential gene expression from RNA-seq data
- i) normalization of count data;
- ii) sensitivity and specificity of DE detection;
- iii) performance on the subset of genes that are expressed in one condition but have no detectable expression in the other condition;
- iv) the effects of reduced sequencing depth and number of replicates on the detection of differential expression.

Normalized counts by log expression correlation

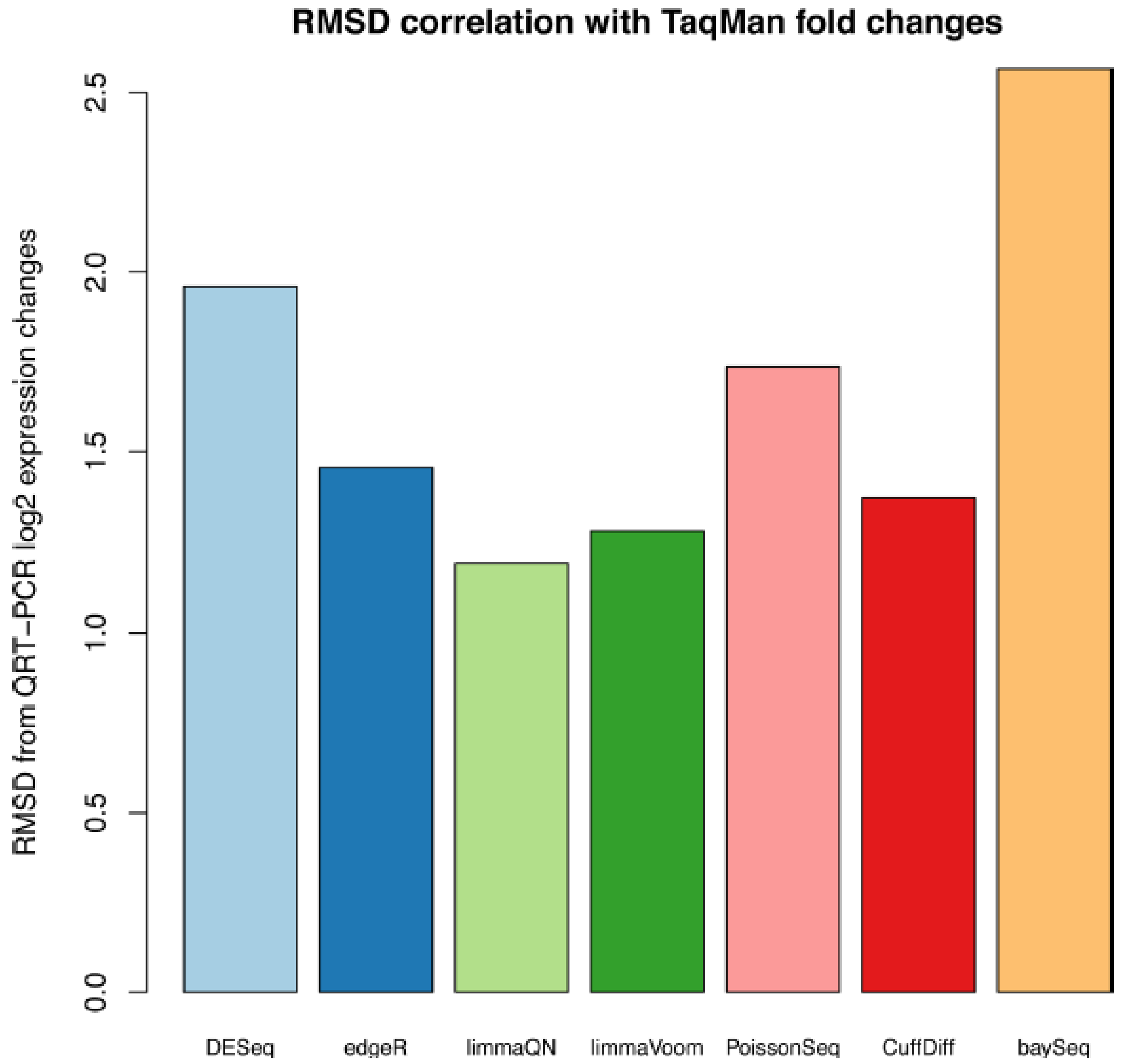


Figure 1 RMSD correlation between qRT-PCR and RNA-seq $\log_2$ expression changes computed by each method. Overall, there is good concordance between $\log_2$ values derived from the DE methods and the experimental values derived from qRT-PCR measures. Upper quartile normalization implemented in baySeq package is least correlated with qRT-PCR values. DE, differential expression; RMSD, root-mean-square deviation.

Differential expression analysis

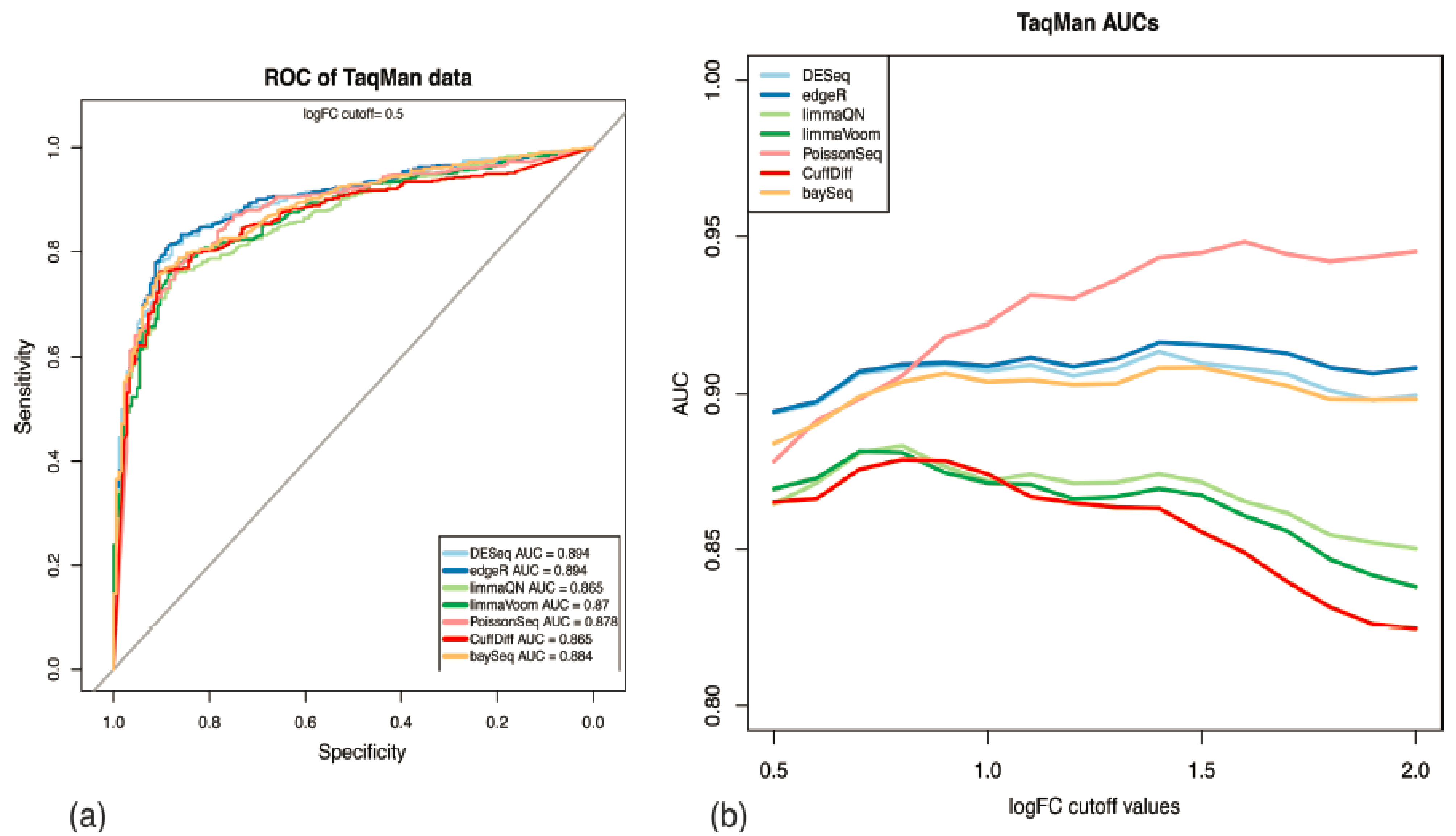
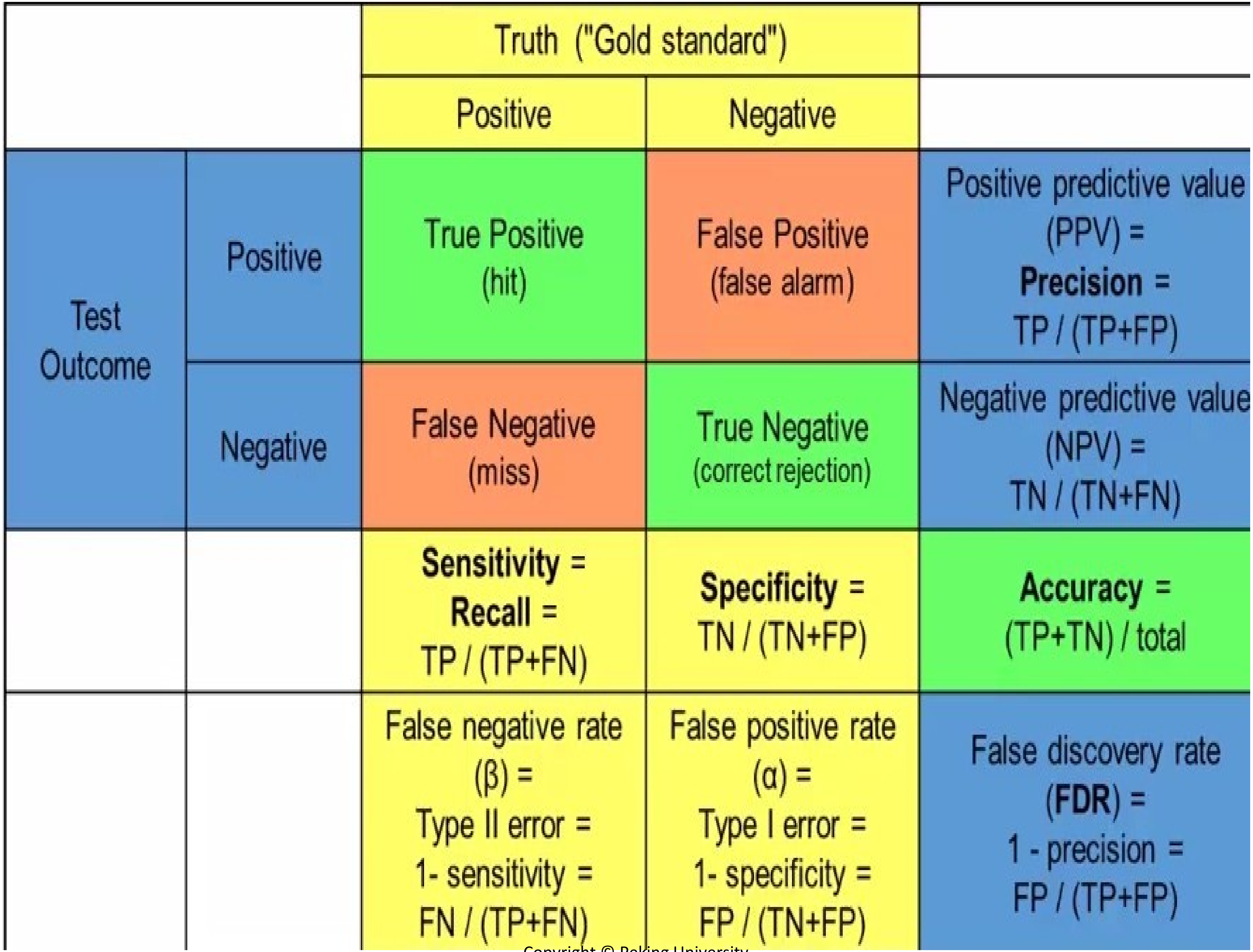
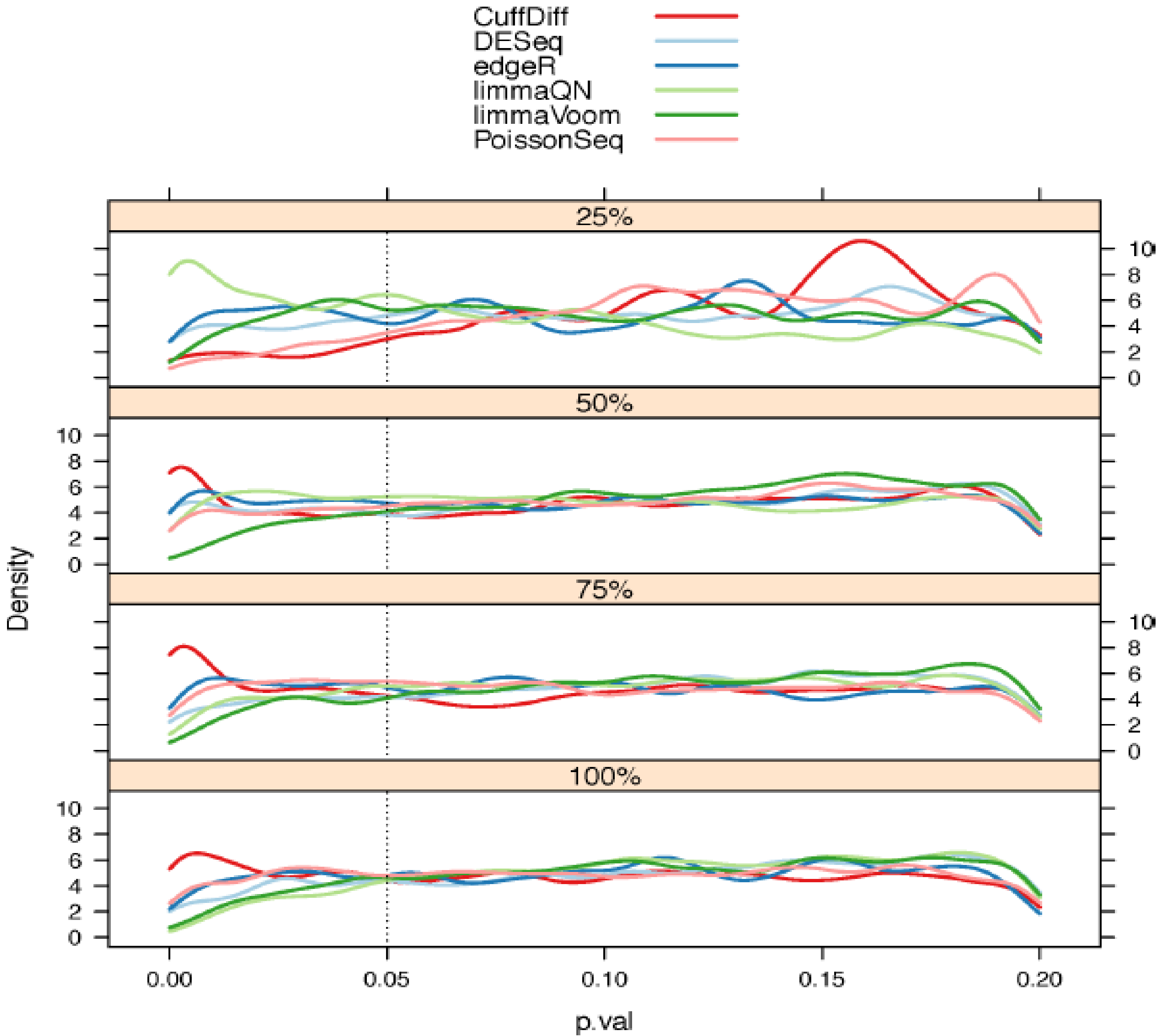


Figure 2 Differential expression analysis using qRT-PCR validated gene set. (a) ROC analysis was performed using a qRT-PCR $\log_2$ expression change threshold of 0.5. The results show a slight advantage for DESeq and edgeR in detection accuracy. **(b)** At increasing $\log_2$ expression ratios (incremented by 0.1), representing a more stringent cutoff for differential expression, the performances of the Cuffdiff and limma methods gradually reduce whereas PoissonSeq performance increases. AUC, area under the curve.

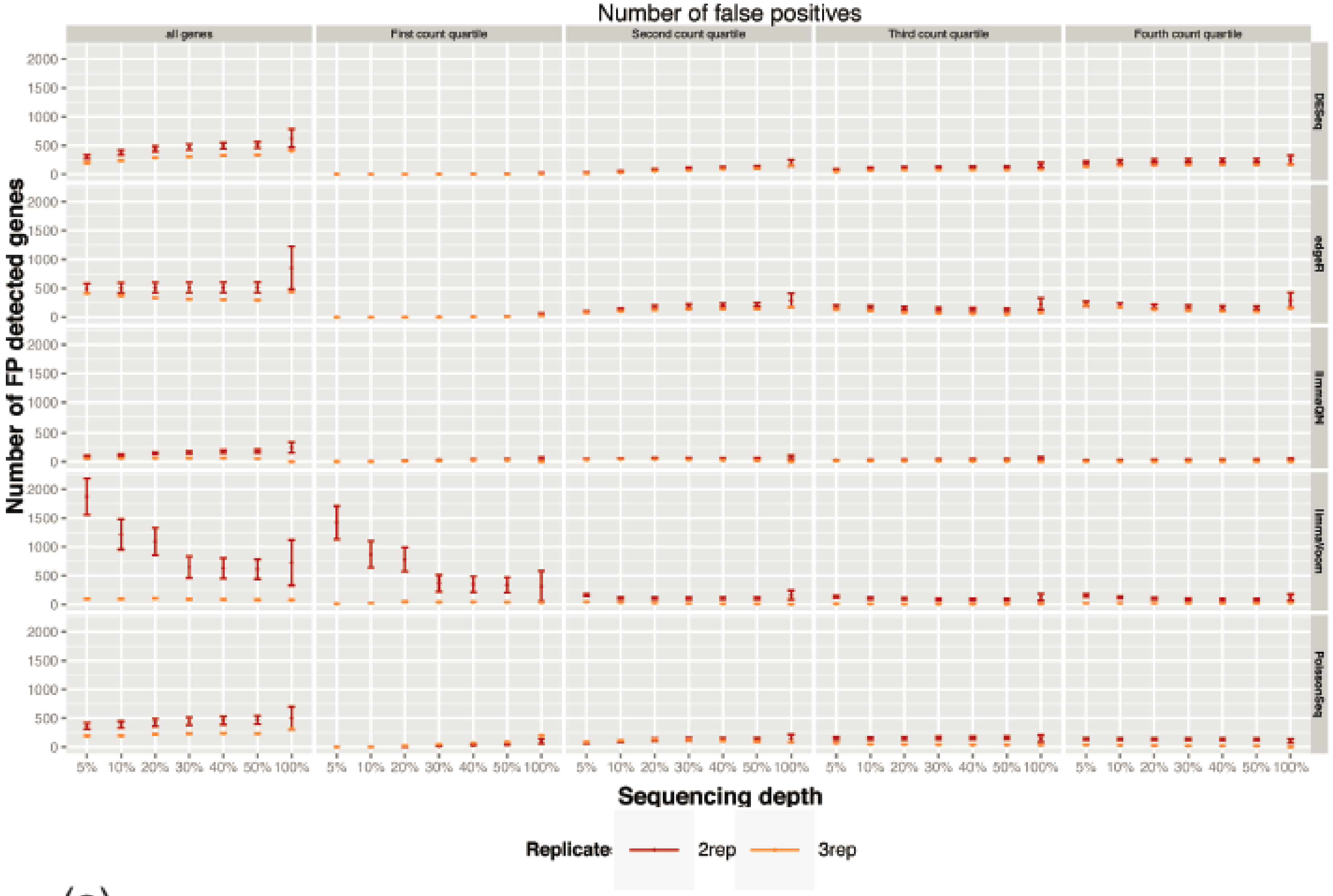


		Truth ("Gold standard")		
		Positive	Negative	
Test Outcome	Positive	True Positive (hit)	False Positive (false alarm)	Positive predictive value (PPV) = **Precision** = TP / (TP+FP)
	Negative	False Negative (miss)	True Negative (correct rejection)	Negative predictive value (NPV) = TN / (TN+FN)
		Sensitivity = **Recall** = TP / (TP+FN)	**Specificity** = TN / (TN+FP)	**Accuracy** = (TP+TN) / total
		False negative rate (β) = Type II error = 1- sensitivity = FN / (TP+FN)	False positive rate (α) = Type I error = 1- specificity = FP / (TN+FP)	False discovery rate (**FDR**) = 1 - precision = FP / (TP+FP)

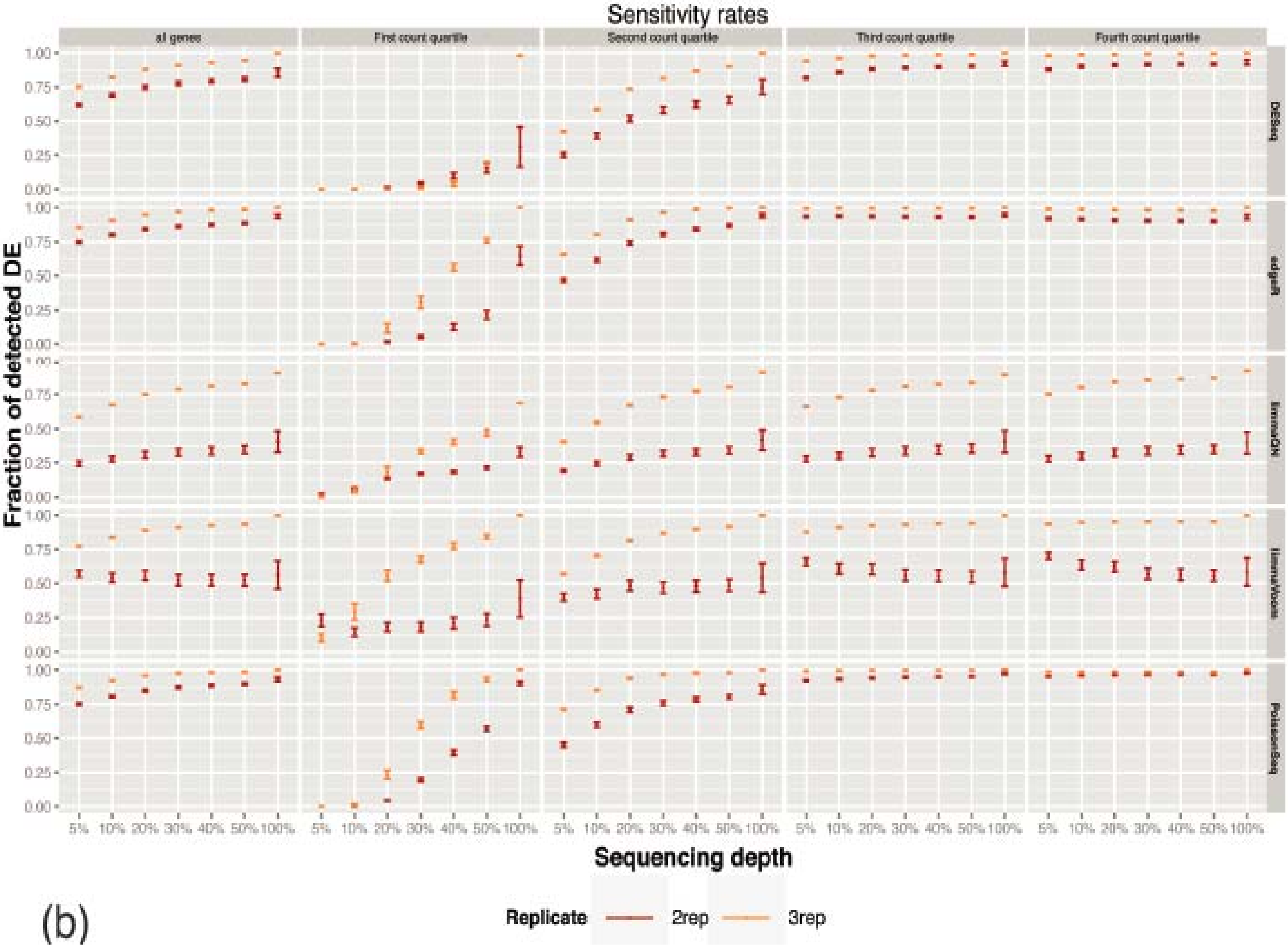
Null model evaluation of type I error



Impact of sequencing depth and number of replicate samples on DE analysis



(a)


Sensitivity rates
all genes
First count quartile
Second count quartile
Third count quartile
Fourth count quartile
DESeq
edgeR
limmaQN
limmaVoom
PoissonSeq
Fraction of detected DE
0.00
0.25
0.50
0.75
1.00
Sequencing depth
5%
10%
20%
30%
40%
50%
100%
Replicate
2rep
3rep
(b)

Conclusion

1 In most benchmarks Cuffdiff performed less favorably

- ✓ with a higher number of false positives
- ✓ without any increase in sensitivity.

2 Our results conclusively demonstrate that the addition of replicate samples provides substantially greater detection power of DE than increased sequence depth.

- Hence, including more replicate samples in RNA-seq experiments is always to be preferred over increasing the number of sequenced reads.

Thanks for your attention!