Differential expression in RNA-seq: a matter of depth

Supplementary Material

Sonia Tarazona[1,2], Fernando García-Alcalde[1], Joaquín Dopazo[1], Alberto Ferrer[2], and Ana Conesa[1]

[1]Bioinformatics and Genomics Department,
Centro de Investigación Príncipe Felipe, Valencia, Spain

aconesa@cipf.es

[2]Department of Applied Statistics, Operations Research and Quality,
Universidad Politécnica de Valencia, Valencia, Spain

August 24, 2011

Table S1: Biotype classification and description. Ensembl biotypes have been grouped in 14 classes to facilitate the interpretation of results.

Group	**Meaning**	**Ensembl biotypes**
protein_coding	Contains an open reading frame (ORF).	protein_coding
processed_transcript	Doesn't contain an ORF.	processed_transcript
pseudogene	Have homology to proteins but generally suffer from a disrupted coding sequence and an active homologous gene can be found at another locus. Sometimes these entries have an intact coding sequence or an open but truncated ORF, in which case there is other evidence used (for example genomic polyA stretches at the 3' end) to classify them as a pseudogene.	pseudogene; polimorphic_pseudogene
IG	Immonuglobulin variable chain and T-cell receptor (TcR) genes.	IG_C_gene; IG_C_pseudogene; IG_D_gene; IG_J_gene; IG_J_pseudogene; IG_V_gene; IG_V_pseudogene
lincRNA	Long, intervening noncoding RNAs, that can be found in evolutionarily conserved, intergenic regions.	lincRNA
miRNA	MicroRNA	miRNA; miRNA_pseudogene
miscRNA	Miscellaneous RNA	miscRNA; misc_RNA_pseudogene
Mt	Mitochondrial	Mt_rRNA; Mt_tRNA; Mt_tRNA_pseudogene
rRNA	Ribosomic RNA	rRNA; rRNA_pseudogene
scRNA_pseudogene	Small cytoplasmic RNA pseudogene	scRNA_pseudogene
snoRNA	Small nucleolar RNA	snoRNA; snoRNA_pseudogene
snRNA	Small nucleic acid RNA	snRNA; snRNA_pseudogene
TR	T cell receptor generated	TR_C_gene; TR_J_gene; TR_V_gene; TR_V_pseudogene
tRNA_pseudogene	Transference RNA	tRNA_pseudogene

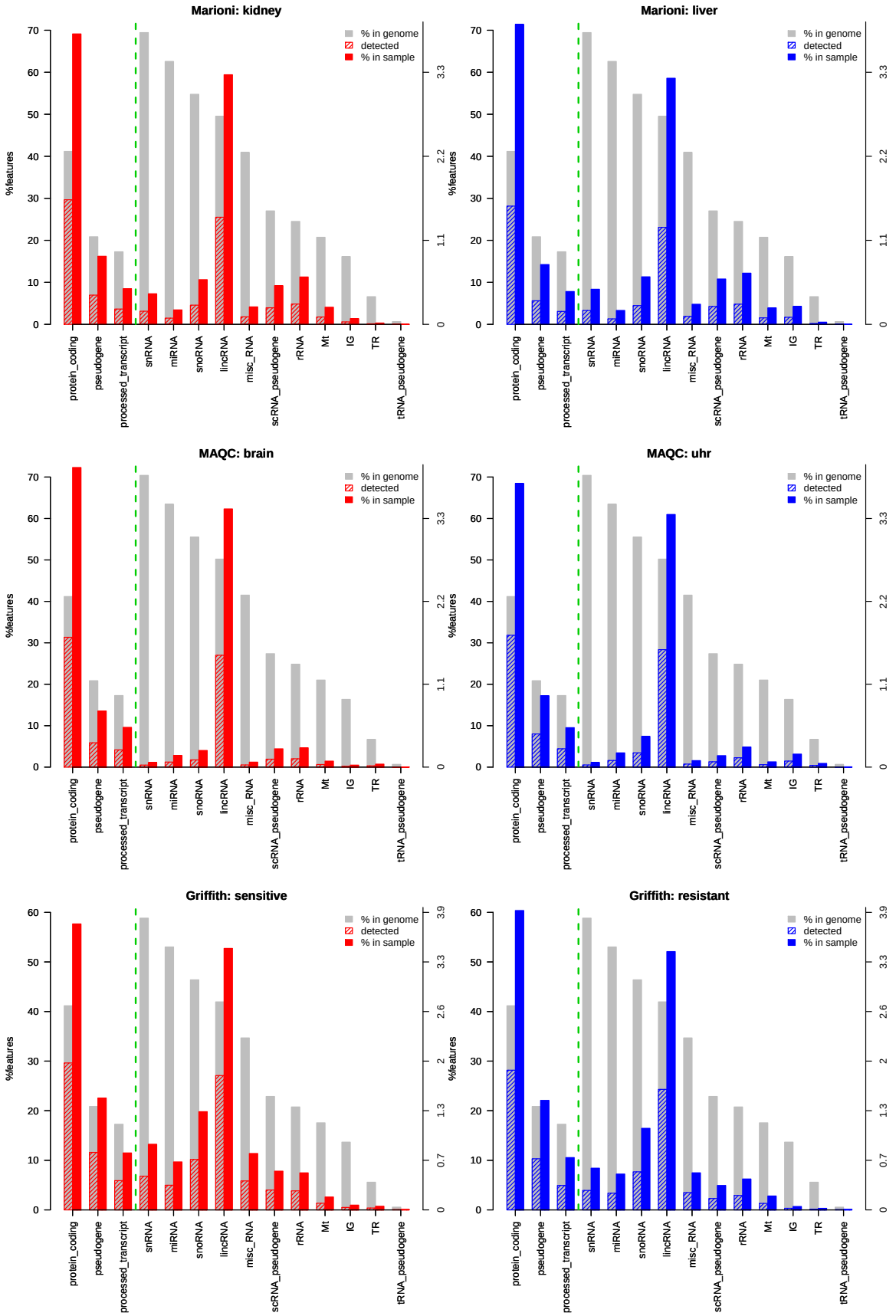

Figure S1: Detection percentages per transcript biotype for each experimental condition in Marioni's, MAQC and Griffith's datasets. Grey bar: Percentage of each biotype in the genome, representing the relative abundance of the biotype in the genome. Striped color bar: Biotype percentage detected in the sample (with more than 5 counts) with regard to the genome, representing which percentage of genes with that biotype are being detected in the sample. Solid color bar: Percentage of the biotype in the total detected features in the sample, representing the relative abundance of the biotype in the sample. Vertical line separates bars expressed in left and right y-axis scale.

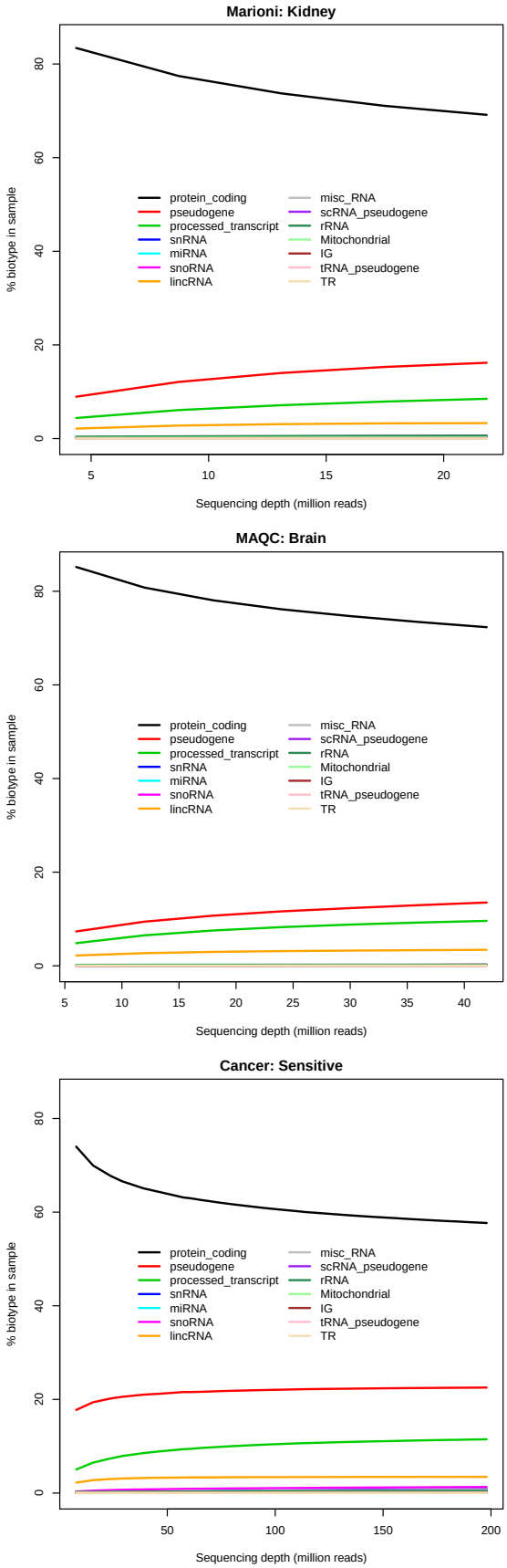

Figure S2: Percentage of each transcript biotype within total detections at increasing sequencing depth for one of the experimental conditions of each one of Marioni's, MAQC and Griffith's datasets.

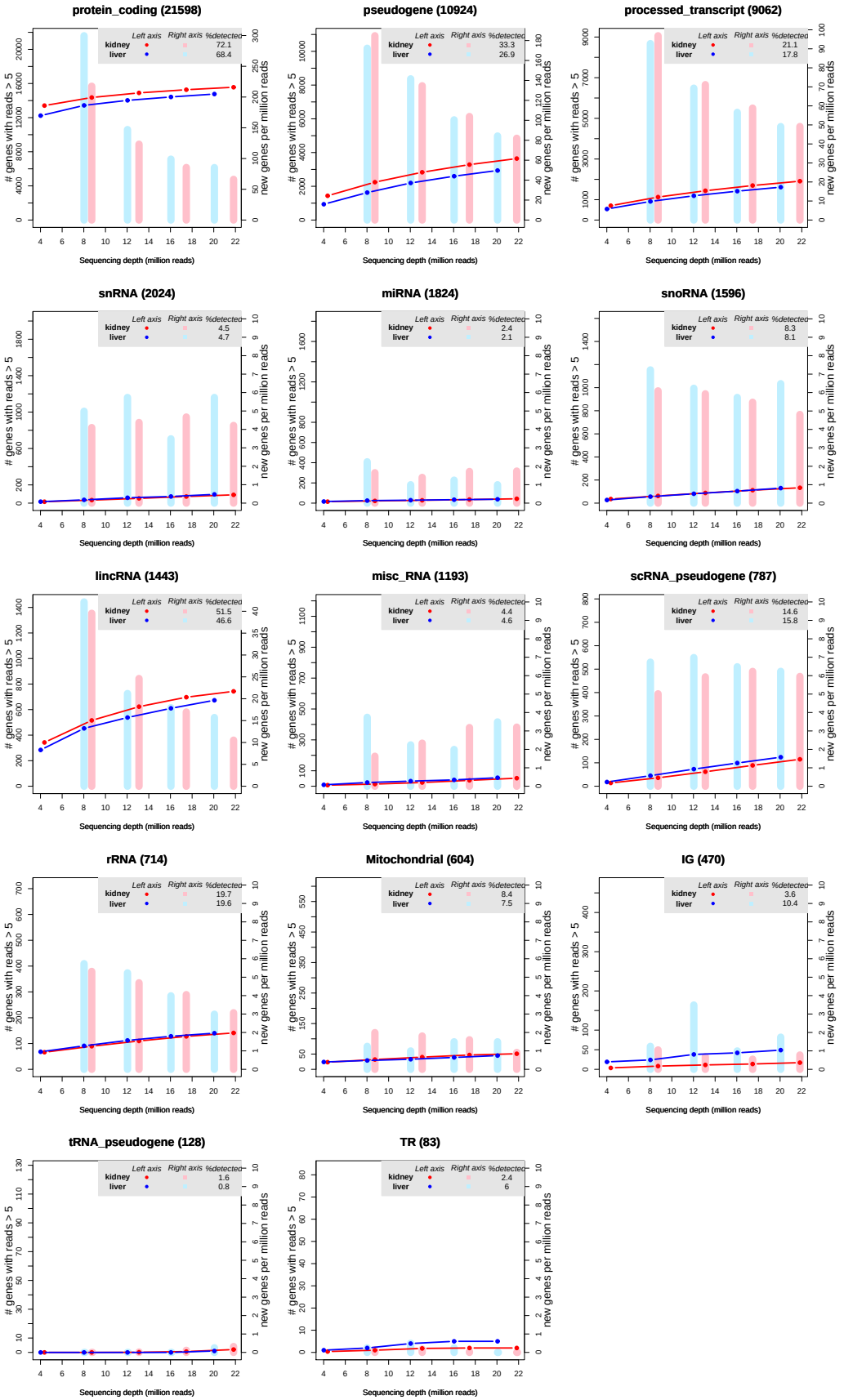

Figure S3a: Marioni's dataset. Saturation curves per biotype display the number of genes detected by more than 5 uniquely mapped reads as a function of the sequencing depth for each experimental condition (left y-axis). Vertical bars represent the number of newly detected genes per million additional reads (NDR, right y-axis) for each experimental condition.

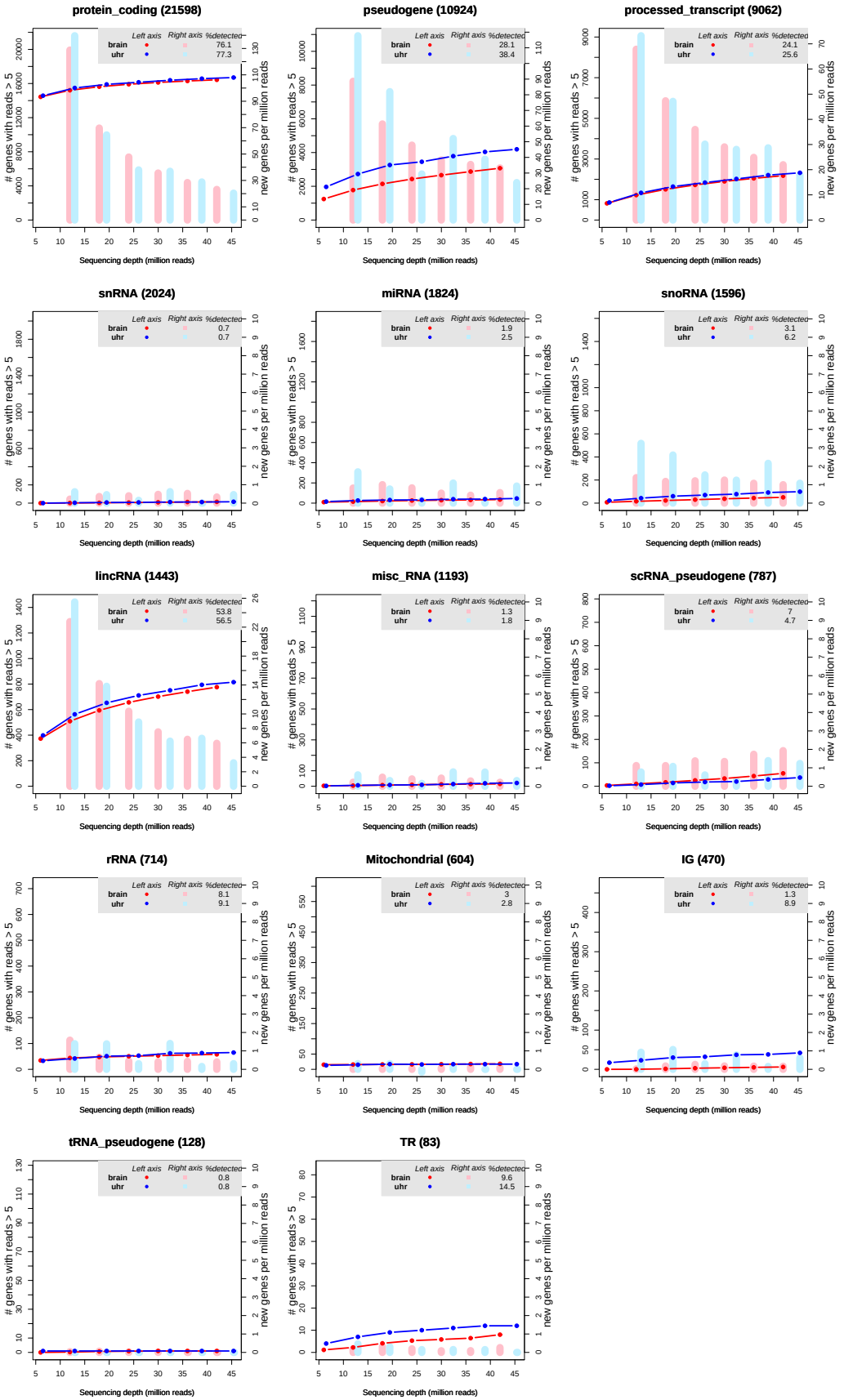

Figure S3b: MAQC dataset. Saturation curves per biotype display the number of genes detected by more than 5 uniquely mapped reads as a function of the sequencing depth for each experimental condition (left y-axis). Vertical bars represent the number of newly detected genes per million additional reads (NDR, right y-axis) for each experimental condition.

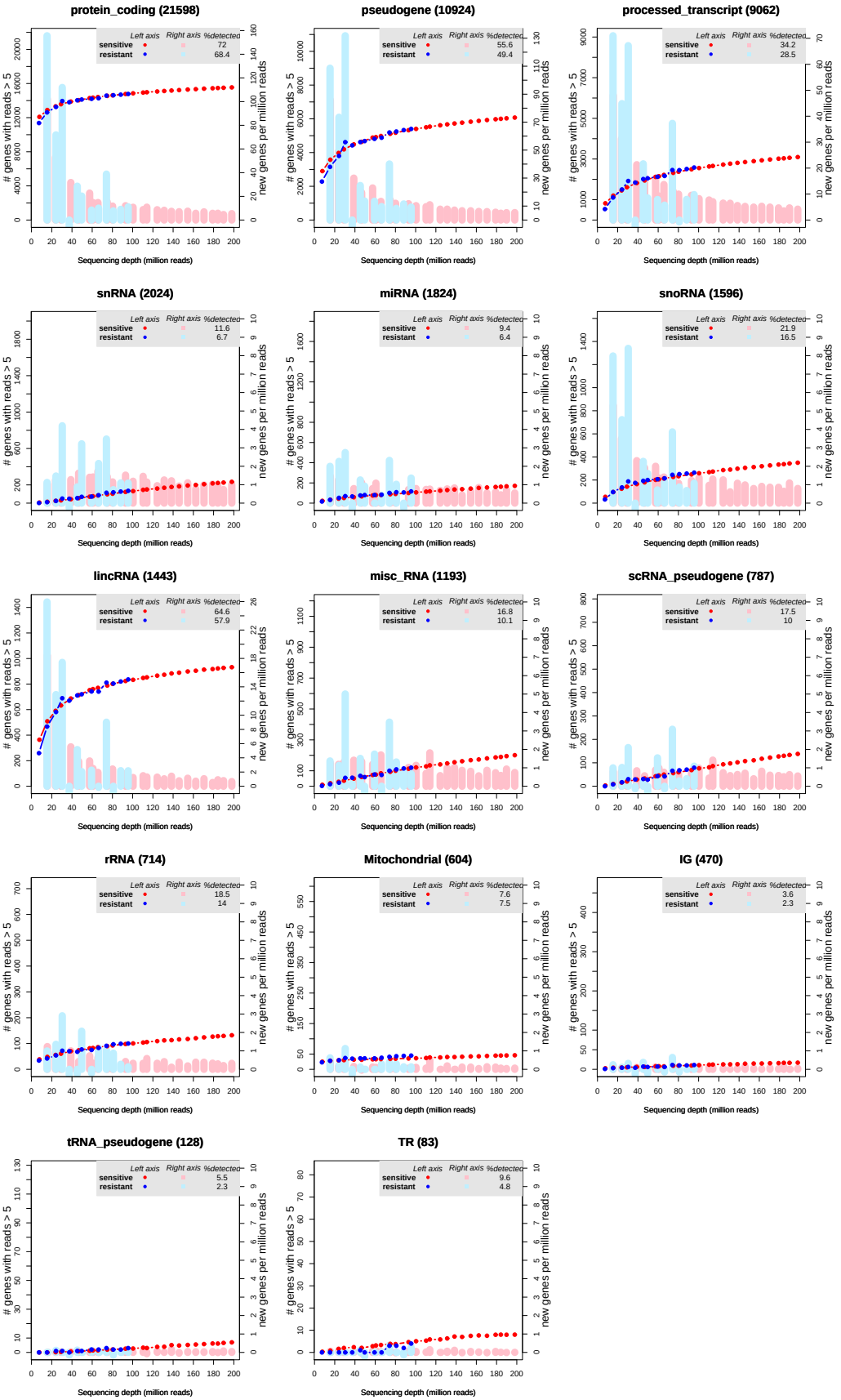

Figure S3c: Griffith's dataset. Saturation curves per biotype display the number of genes detected by more than 5 uniquely mapped reads as a function of the sequencing depth for each experimental condition (left y-axis). Vertical bars represent the number of newly detected genes per million additional reads (NDR, right y-axis) for each experimental condition.

Figure S4: Small RNAs within intronic regions

When considering the per-biotype break-down analysis, it should be noticed that miRNAs and other small RNAs are often embedded in the introns of protein-coding genes and that reads assigned to these small RNAs species could also be intronic reads, particularly if the gene is highly expressed. Although the count procedure applied in this work only used reads mapping to exons (and not to intronic regions), the annotation scheme of *gff* files might still result in some intronic miRNAs reads being included. Removing these intronic small RNAs from the mapping data resulted in a reduction of the number of detected miRNAs genes in about 500 (Marioni's data), but this did not substantially altered the sequencing depth analysis and the conclusions of this work. Figure S4 shows saturation curves for Marioni dataset after removing small RNA intronic reads. It can be observed that saturation dynamics are practically identical to those recorded for the full dataset (see Fig. S3a).

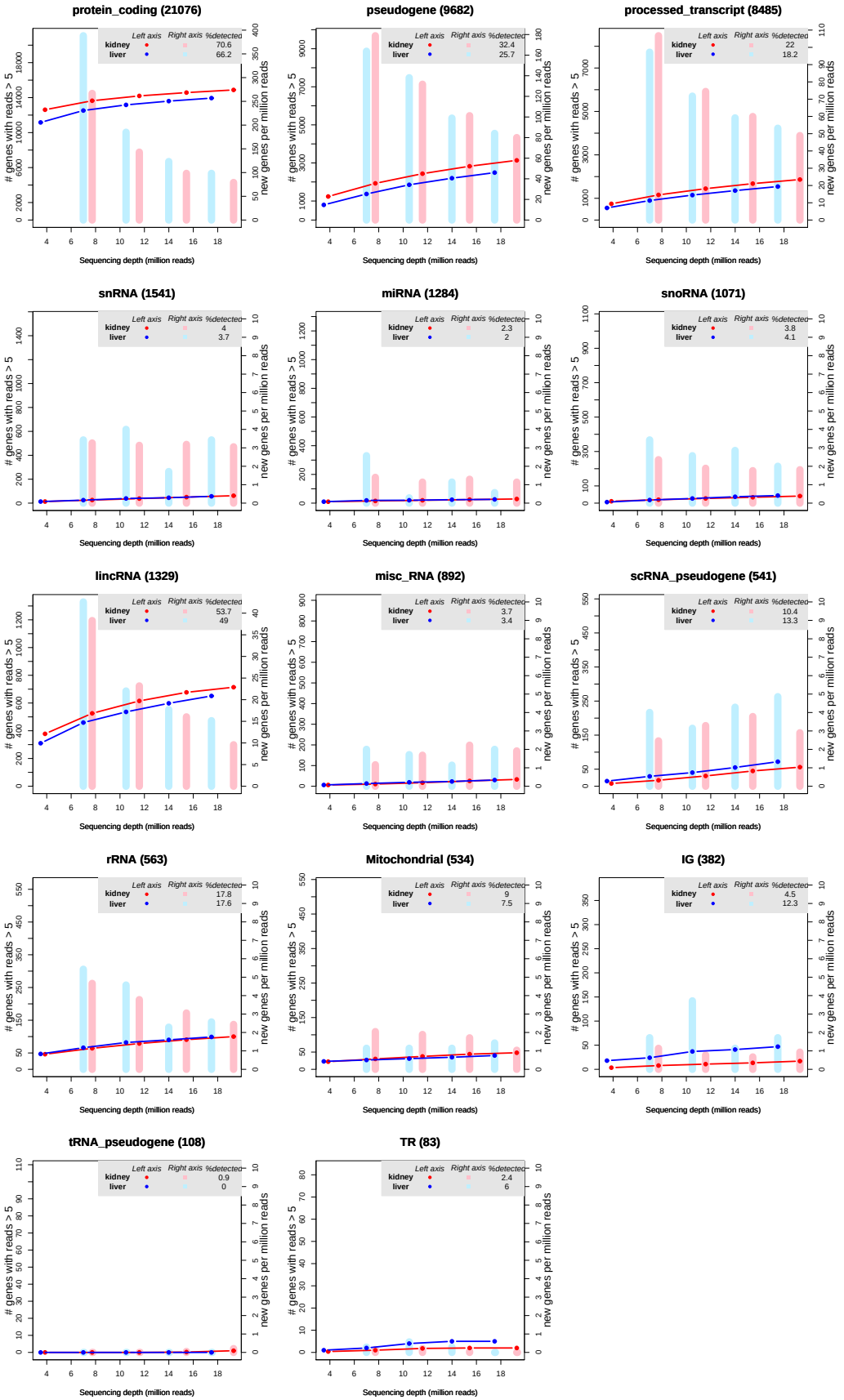

Figure S4: Marioni's dataset (excluding intronic regions). Saturation curves per biotype display the number of genes detected by more than 5 uniquely mapped reads as a function of the sequencing depth for each experimental condition (left y-axis). Vertical bars represent the number of newly detected genes per million additional reads (NDR, right y-axis) for each experimental condition.

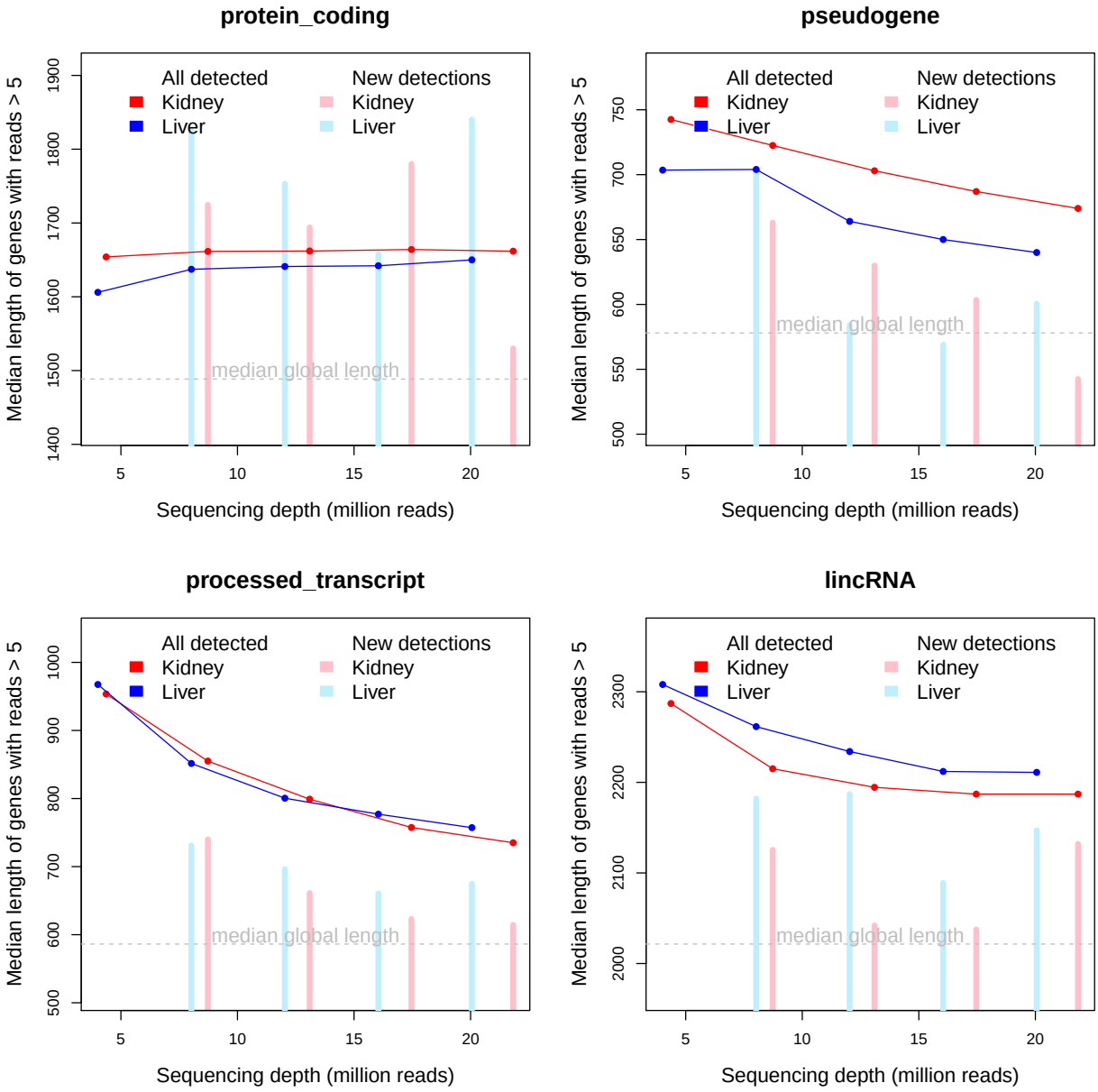

Figure S5a: Marioni's dataset. Median length of genes with more than 5 counts is represented according to different sequencing depths for each sample (red and blue curves). Pink and light blue bars represent the median length of the new detected genes when increasing sequencing depth.

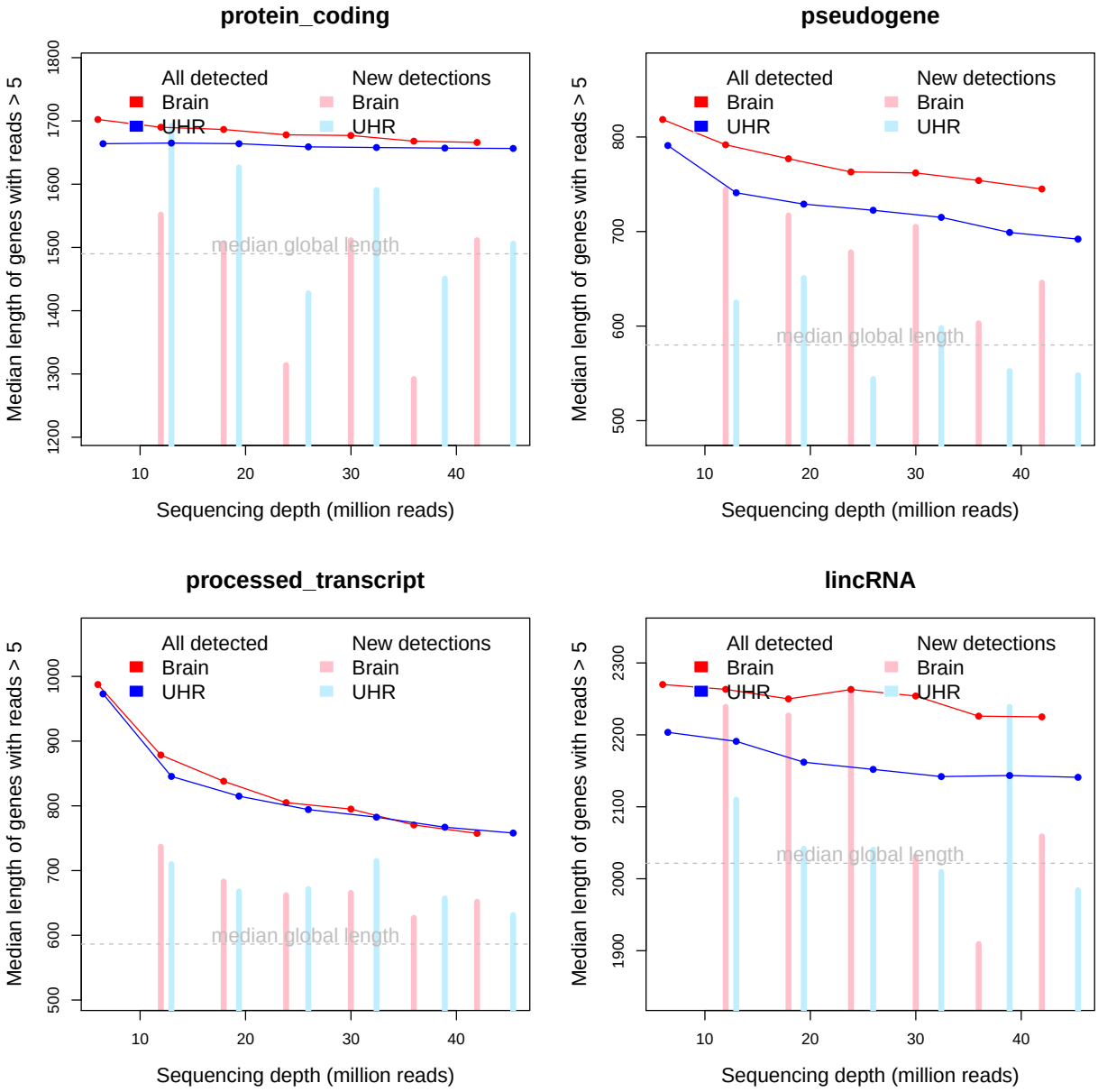

Figure S5b: MAQC dataset. Median length of genes with more than 5 counts is represented according to different sequencing depths for each sample (red and blue curves). Pink and light blue bars represent the median length of the new detected genes when increasing sequencing depth.

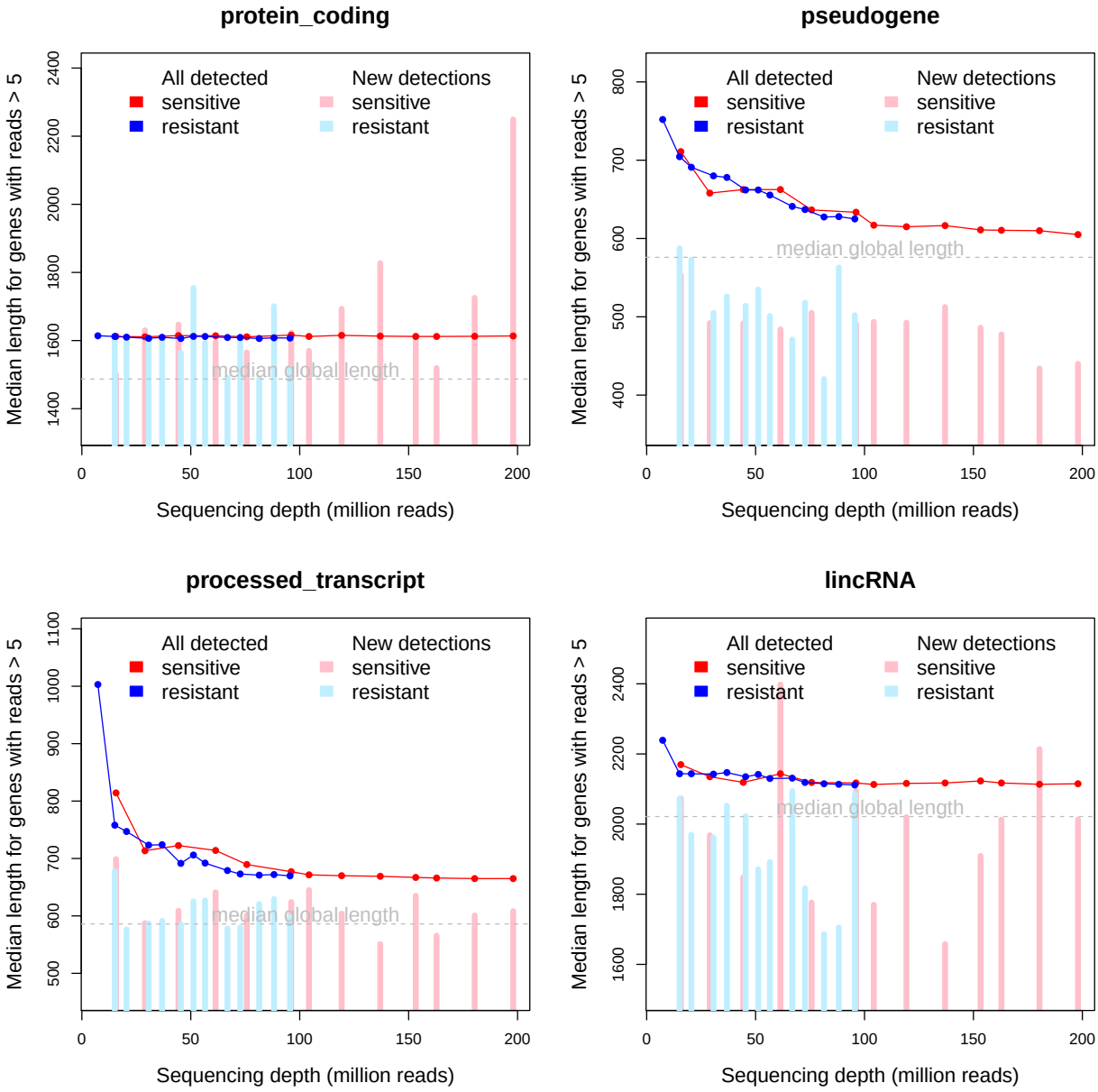

Figure S5c: Griffith's dataset. Median length of genes with more than 5 counts is represented according to different sequencing depths for each sample (red and blue curves). Pink and light blue bars represent the median length of the new detected genes when increasing sequencing depth.

NOISeq-sim parameters

The simulation procedure to generate replicates in NOISeq-sim relies on two adjustable input parameters: the number of replicates to be generated (or number of simulated samples, *nss*) and the size of each simulated replicate which is given as a percentage of the total number of reads (parameter *pnr*). Optimal values for these parameters are those that would produce the same results as when using true technical replicates with NOISeq-real and also present good rates for false positives and negatives. The best-performing values of *nss* and *pnr* were determined using a synthetic dataset where the number and identity of the differentially expressed genes is known. This synthetic dataset was obtained from the baySeq package (Hardcastle et al. 2010) and consisted of counts for 1000 features in 5 replicates for each condition, being the first 100 features differentially expressed. Using these data, we evaluated Precision-Recall Curves (PRC) and False Discovery Rates (FDR) at different values of *pnr* and *nss*.

We first evaluated different sample sizes (*pnr*), while maintaining the number of replicates *nss* equal to the number of them in the simulated data, i.e., 5. PRC were highly similar for different *pnr* values, while FDR was best at a *pnr* of 0.2 (Fig. S6). Naturally, the lowest FDR was obtained when using real replicates. Next, we varied the value of *nss* maintaining *pnr* at 0.2. Fig. S7 shows a high stability of results for the number of replicates for both PRC and FDR. Therefore, we set as default values for NOISeq-sim 5 simulated replicates ($nss = 5$) and $pnr = 0.2$, thus maintaining the total number of simulated replicates reads at the same level as in the starting data.

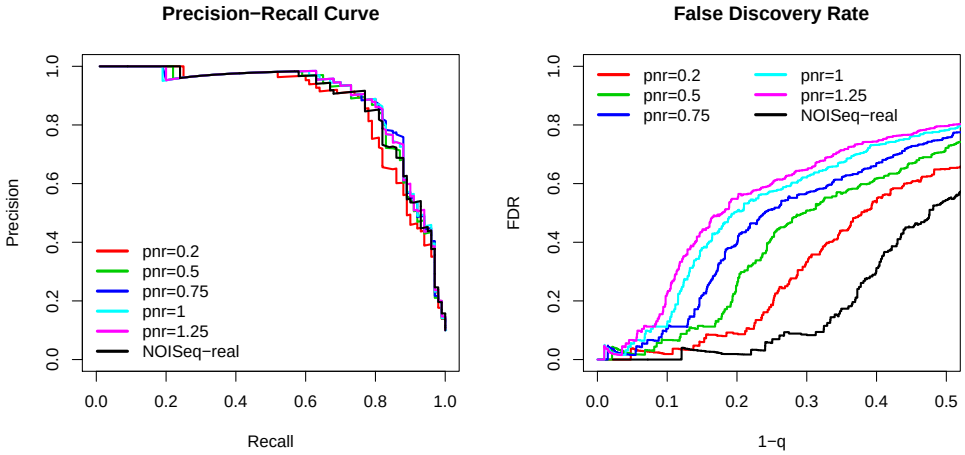

Figure S6: Precision-Recall curves and False Discovery Rates for NOISeq-sim with different values of *pnr* parameter compared to NOISeq-real on synthetic data.

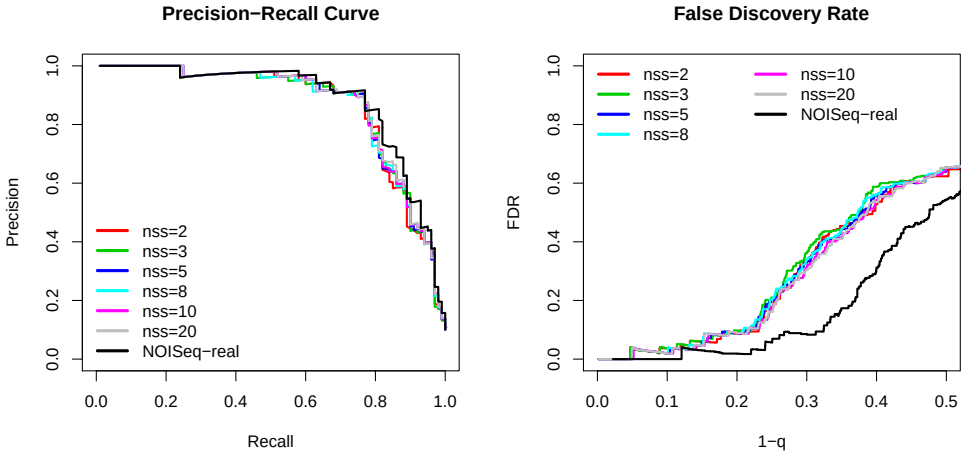

Figure S7: Precision-Recall curves and False Discovery Rates for NOISeq-sim with different values of *nss* parameter compared to NOISeq-real on synthetic data.

Fig. S8 shows the number of differentially expressed genes (d.e.g.) identified by each method in the three benchmarking datasets. A similar gene selection was obtained for both Marioni and MAQC data using the two algorithm variants, but NOISeq-sim repeatedly resulted in the detection of additional significant features. Differences between the two methods were bigger in the Griffith's dataset, presumably due to the characteristics of these data. In Griffith's work, two human colorectal cancer lines, differing in their resistance phenotype to fluorouracil, were compared and few gene expression differences were expected. The original paper reported 259 differentially expressed genes. We observed a large variability among lanes within the same sample for these data and also noted that lanes from both experimental conditions became mixed (Fig. S9), unlike other datasets. This high variance might explain the poorer simulation performance in mimicking the actual data, and hence the differences in gene selection between NOISeq-sim and NOISeq-real. Nonetheless, a replicate relative size of 20% was also optimal for this dataset. Therefore, for the sake of comparison, NOISeq-sim parameters were set at nss=5 simulated replicates per condition with a size of pnr=20% of the total amount of reads in that condition.

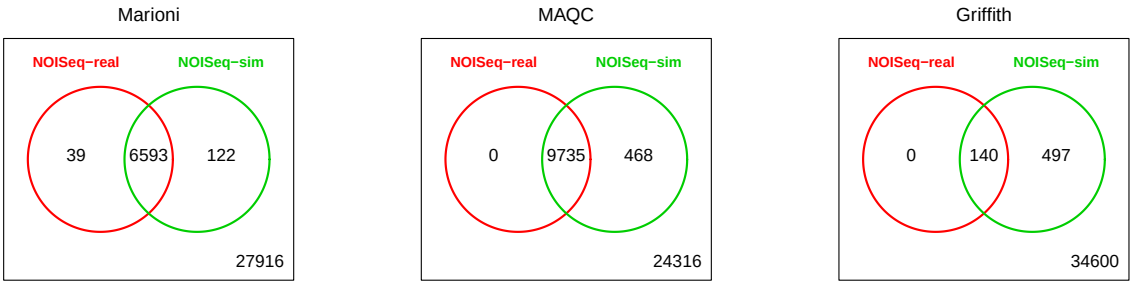

Figure S8: Differentially expressed genes detected by NOISeq-real and NOISeq-sim. Five replicates per condition were generated in NOISeq-sim, each with 20% of the total number of reads in the corresponding condition. Counts data were not normalized by gene length.

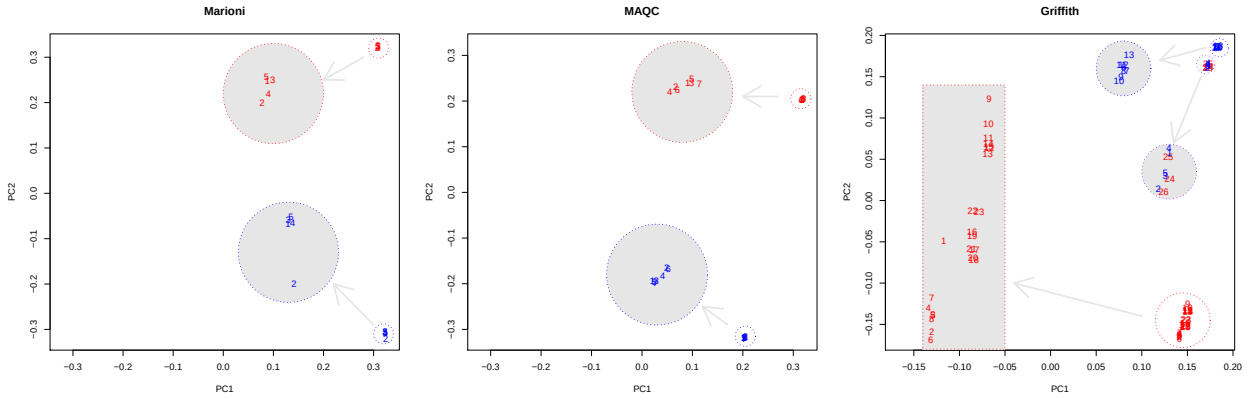

Figure S9: Score Plots from the Principal Component Analysis applied to the Marioni's, MAQC and Griffith's datasets. Points inside the small circles are the real score values for the available samples, and show a high reproducibility of replicates for Marioni's and MAQC datasets, and a high variability between replicates within the same condition in Griffith's dataset. The clusters of samples in small circles have been zoomed (grey figures) to allow the identification of each sample.

Performance assessment of RNA-seq differential expression methods

Current methodologies proposed for the differential expression analysis of mRNA-seq data differ in the parametric assumptions they rely on and in the statistical strategies they employ (Oshlack *et al.*, 2010). Table S2 summarizes all the methods compared in this study.

To concentrate our sequencing depth analysis on a subset of representative and differentiating approaches we first evaluated their general performance and compared them to NOISeq using synthetic data included in the baySeq R package (Hardcastle and Kelly, 2010). This dataset contains counts for 1000 features evaluated in two experimental conditions with 5 replicates each. The first hundred features are differentially expressed. To compute differential expression, a library size correction was applied whenever the methods included this option and Precision-Recall curves (PRC) and False Discovery Rate plots (FDR) were generated in each case (see Methods). Both PRC and FDR plots separated RNA-seq statistical methodologies in two groups (Fig. S10). The methods in edgeR package (CD and TWD), DESeq, baySeq-NB and both versions of NOISeq showed good accuracy and consistent control of False Discoveries, whereas MARS, LRT, FET and baySeq-PO showed a poorer performance. Therefore, we selected edgeR-CD, baySeq-NB, DESeq and FET (the last one included for its extended use) together with NOISeq-sim and NOISeq-real for further analysis.

Table S2: Differential expression methods for RNA-seq data considered in this study.

Method	**Description**	**R package**	**Reference**
MARS	MA-plot based method, estimating noise by Random Sampling.	DEGseq	Wang et al., 2010
LRT	Likelihood Ratio Test based on a Poisson model.	DEGseq	Marioni et al., 2008
FET	Fisher's Exact Test.	stats	Fisher, 1970
edgeR	Exact test based on a Negative Binomial model with two variants: • CD estimates a Common Dispersion for all tags. • TWD estimates a Tag-Wise dispersion.	edgeR	Robinson et al., 2010
DESeq	Exact test based on a Negative Binomial model.	DESeq	Anders and Huber, 2010
baySeq	Empirical bayesian method computing models posterior probabilities from: • Poisson data distribution (PO). • Negative Binomial data distribution (NB).	baySeq	Hardcastle and Kelly, 2010

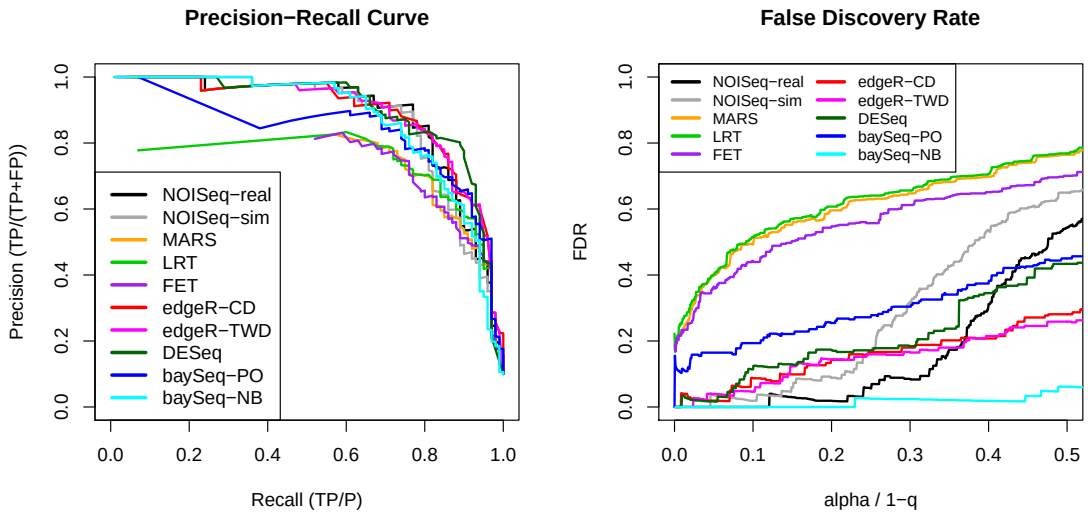

Figure S10: Precision-Recall and False Discovery Rate curves for the differential expression methods compared, applied to the synthetic dataset included in baySeq package.

Table S3: True and False Positive Rates for differential expression methods applied to Griffith's dataset and using RT-PCR as gold-standard. Genes considered as true positives are those declared differentially expressed on RT-PCR data (82 genes). Genes considered as true negatives are those not found differentially expressed by RT-PCR (12 genes).

Method	**Length correction**	**# TP**	**TPR**	**# FP**	**FPR**
NOISeq-real	No	47	57.3%	0	0.0%
NOISeq-real	Yes (RPKM)	35	42.7%	0	0.0%
NOISeq-sim	No	63	76.8%	2	16.7%
NOISeq-sim	Yes (RPKM)	61	74.4%	0	0.0%
FET	No	14	17.1%	0	0.0%
FET	RPKM	9	11.0%	0	0.0%
edgeR	No	73	89.0%	5	41.7%
DESeq	No	70	85.4%	4	33.3%
baySeq	No	58	70.7%	3	25.0%
baySeq	Yes	61	74.4%	3	25.0%

Table S4a: Number of differentially expressed genes declared by each method in Marioni's dataset when using an increasing number of lanes in the analysis. Methods were applied on data without length normalization and on length-normalized data when allowed by the method.

Methods	**1 lane**	**2 lanes**	**3 lanes**	**4 lanes**	**5 lanes**
No length correction					
NOISeq-real	–	6446	6564	6588	6632
NOISeq-sim	4683	5596	6095	6444	6715
FET	2199	2171	2158	2157	2167
edgeR	6308	8745	10221	11352	12176
DESeq	–	6728	8119	9154	10009
baySeq	–	2504	5688	7472	8750
Length correction					
NOISeq-real	–	6686	6885	6961	7012
NOISeq-sim	4532	5641	6251	6678	7063
FET	1681	1642	1632	1625	1617
baySeq	–	2733	5364	7868	8787

Table S4b: Number of differentially expressed genes declared by each method in MAQC dataset when using an increasing number of lanes in the analysis. Methods were applied on data without length normalization and on length-normalized data when allowed by the method.

Methods	**1 lane**	**2 lanes**	**3 lanes**	**4 lanes**	**5 lanes**	**6 lanes**	**7 lanes**
No length correction							
NOISeq-real	–	9141	9480	9546	9615	9661	9735
NOISeq-sim	7446	8455	8887	9369	9566	9957	10203
FET	5096	5090	5072	5066	5053	5059	5046
edgeR	11050	13487	14926	15954	16699	17443	17872
DESeq	–	11874	13773	14782	15700	16473	16923
baySeq	–	6068	10536	12438	13546	14385	15110
Length correction							
NOISeq-real	–	9349	9694	9873	9925	10021	10072
NOISeq-sim	7281	8591	9182	9729	10052	10501	10699
FET	3871	3880	3843	3825	3823	3826	3824
baySeq	–	7153	10928	12858	14260	14751	15495

As it can be observed in Tables S4a, S4b and S4c, the number of differentially expressed genes is quite higher for edgeR, DESeq and baySeq than for NOISeq. We perfomed a number of functional enrichment analyses on MAQC data using FatiGO tool from Babelomics suite (Medina *et al.*, 2010) to compare the biological information contained in the sets of differentially expressed genes (DEG) in common among methods and in the sets of genes detected by the other methods but not by NOISeq. We excluded Fisher's Exact Test (FET) from this comparison because FET declared less DEG than the other methods, showing a low sensibility, so it makes no sense to study the potential false positives in this case. Table S5 summarizes the number of DEG in each set of the comparison (classified according to the condition in which they are up-regulated) and also the number of significantly enriched GO terms resulting from the functional enrichment analysis of each set against the rest of the genome. The results show that the genes additionally detected by the other methods and not by NOISeq do not contain specific functional information, since none or very few enriched GO terms are found for them. However, when considering genes detected in common by NOISeq and the other methods, many significantly enriched terms are obtained. This indicates that genes selected by NOISeq bear the major functional charge to the biological differences between samples, whereas those identified by other methods lack this characteristic and may therefore be less biologically significant.

Table S4c: Number of differentially expressed genes declared by each method in Griffith's dataset when using an increasing number of lanes in the analysis. Methods were applied on data without length normalization and on length-normalized data when allowed by the method.

Methods	**1 lane**	**2 lanes**	**3 lanes**	**5 lanes**	**7 lanes**	**9 lanes**	**11 lanes**	**13 lanes**
No length correction								
NOISeq-real	–	338	333	333	153	148	152	140
NOISeq-sim	352	421	478	473	622	664	730	637
FET	216	194	186	181	168	171	169	156
edgeR	1674	1195	1880	3207	4099	5415	6576	6056
DESeq	–	1004	1741	2904	3236	4092	5069	5005
baySeq	–	342	598	1071	1861	2824	3621	3247
Length correction								
NOISeq-real	–	386	341	321	152	140	139	125
NOISeq-sim	392	456	487	522	626	714	764	733
FET	239	197	189	188	197	191	182	166
baySeq	–	137	472	1165	1705	2666	3555	3320

Table S5: Comparison between the genes declared as differentially expressed by NOISeq and by the other methods, including the number of significantly enriched GO terms for each set of genes.

	FET	edgeR	DESeq	baySeq
In common between NOISeq and the other method	4799	**9735**	**9693**	**9712**
Up in BRAIN	–	3468	3431	3457
Up in UHR	–	6267	6262	6255
# GO terms (Up in BRAIN)	–	192	178	190
# GO terms (Up in UHR)	–	486	481	485
Detected by the other method and not by NOISeq	247	**137**	**7230**	**5398**
Up in BRAIN	–	2731	3707	1826
Up in UHR	–	5406	3517	3572
BRAIN = UHR	–	0	6	0
# GO terms (Up in BRAIN)	–	0	0	2
# GO terms (Up in UHR)	–	0	4	1
Detected by NOISeq and not by the other method	4936	0	42	23

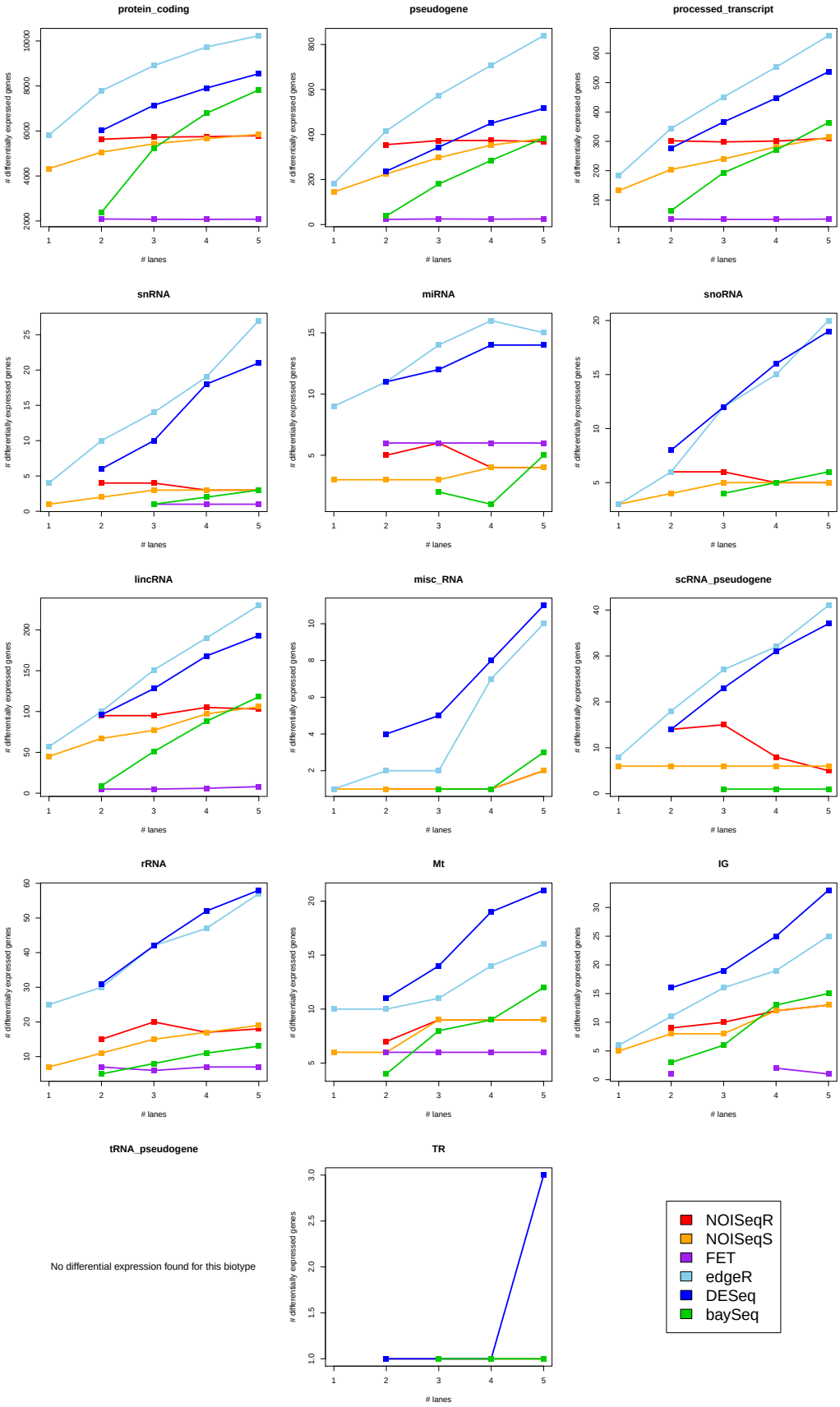

Figure S11a: Number of differentially expressed genes according to the number of lanes used, per biotype, in Marioni's dataset. Count data was not length-normalized.

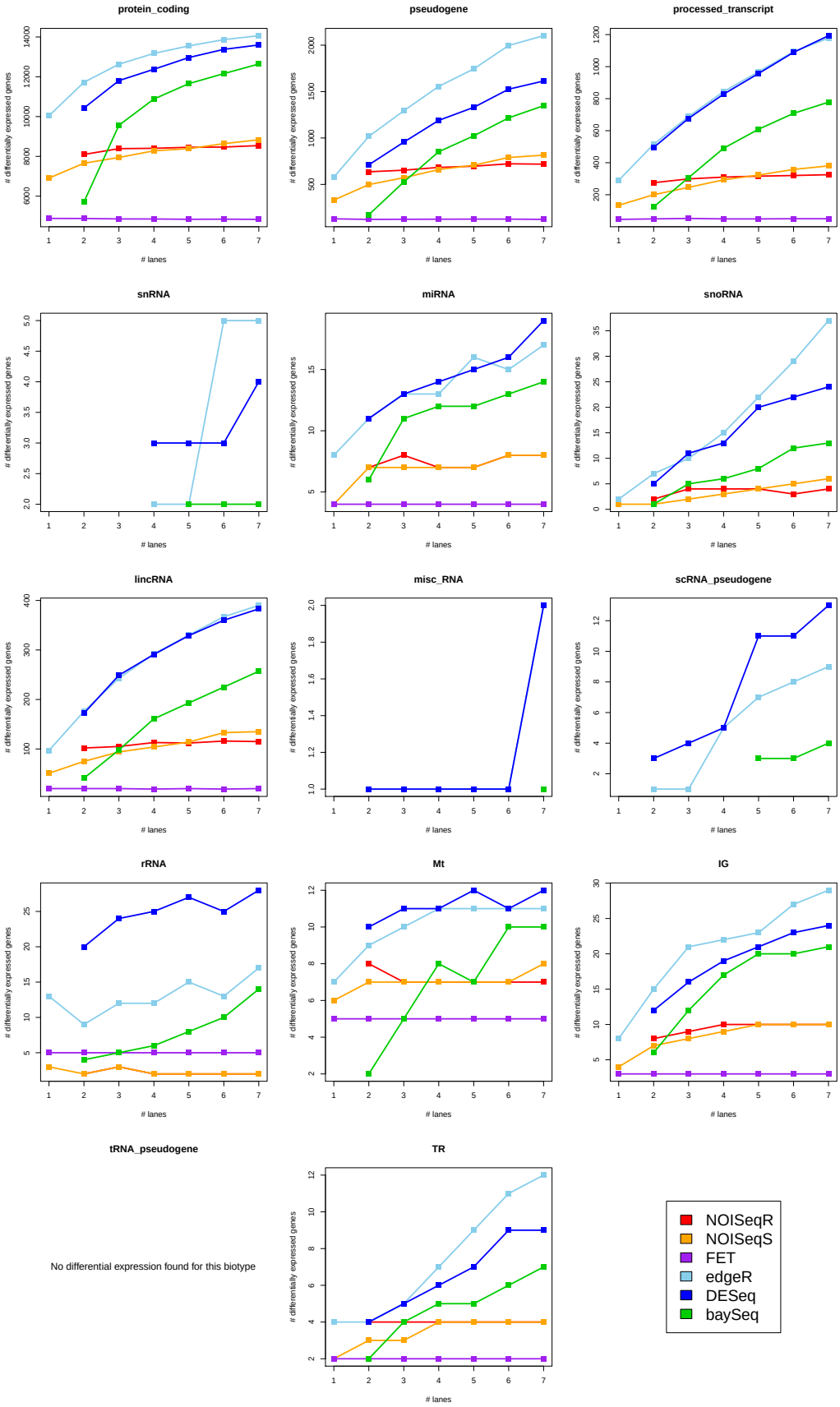

Figure S11b: Number of differentially expressed genes according to the number of lanes used, per biotype, in MAQC dataset. Count data was not length-normalized.

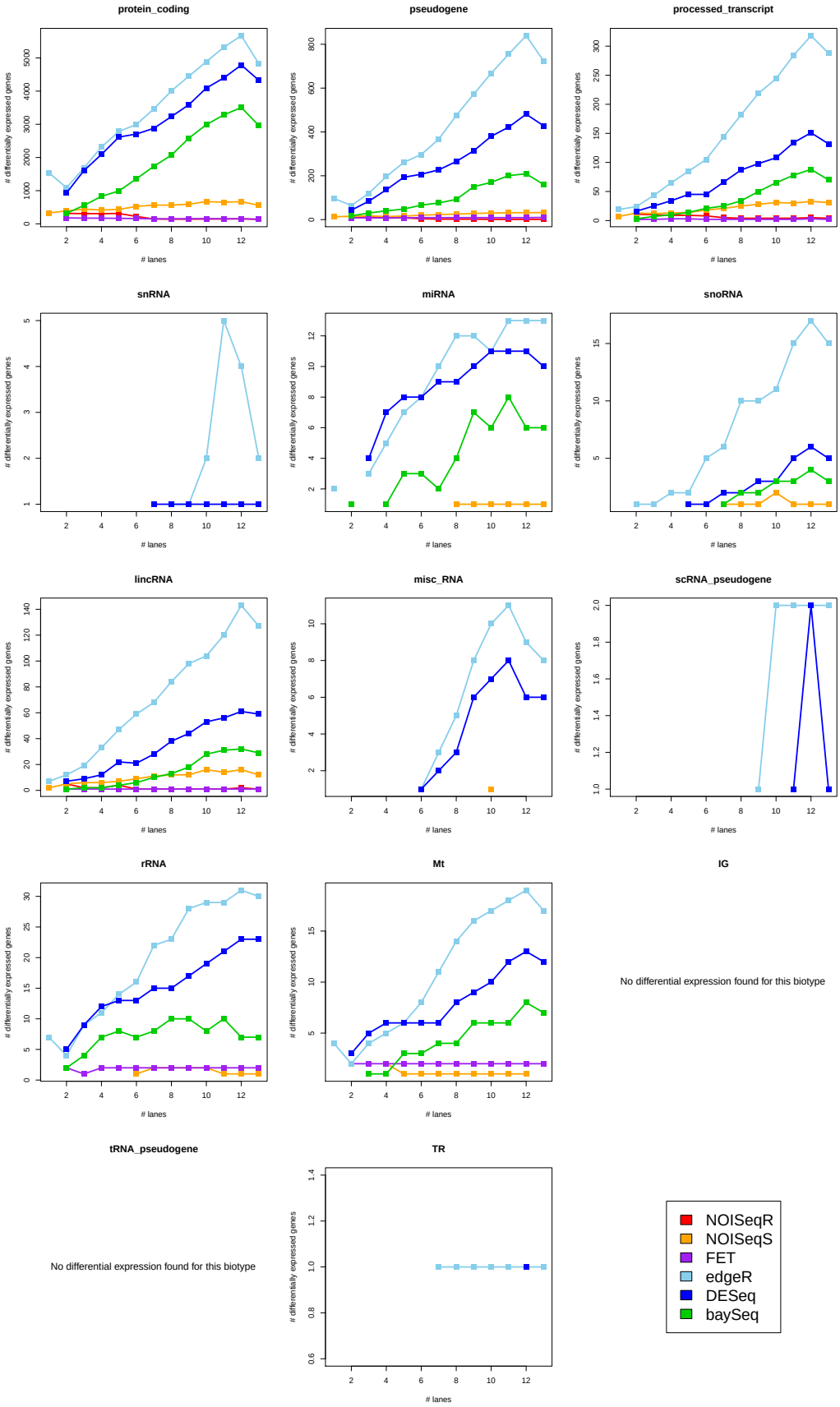

Figure S11c: Number of differentially expressed genes according to the number of lanes used, per biotype, in Griffith's dataset. Count data were not length-normalized.

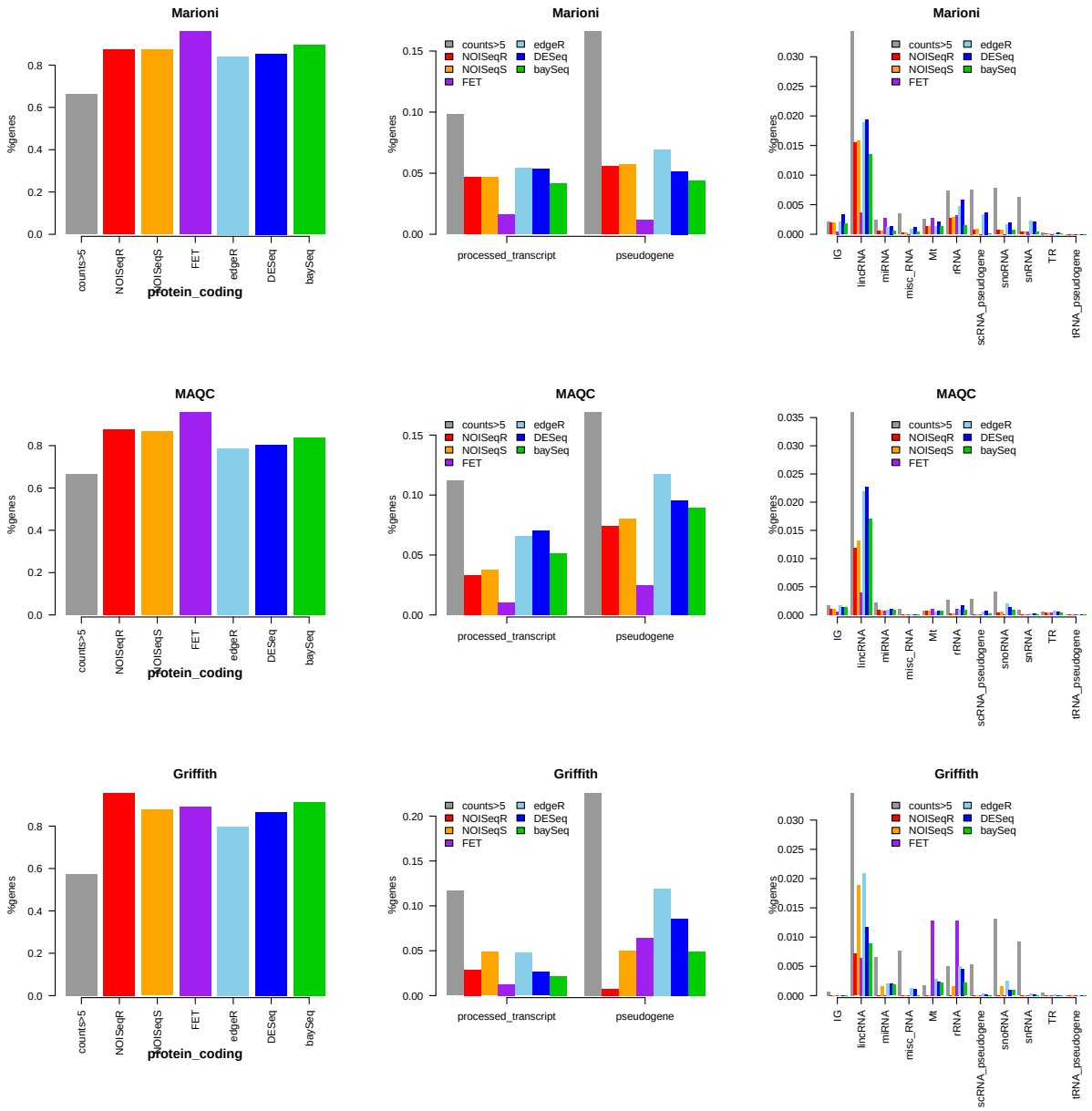

Figure S12: Relative percentage of biotypes within differentially expressed genes compared to detected genes in each dataset and for different methods. No gene length correction was applied to the data.

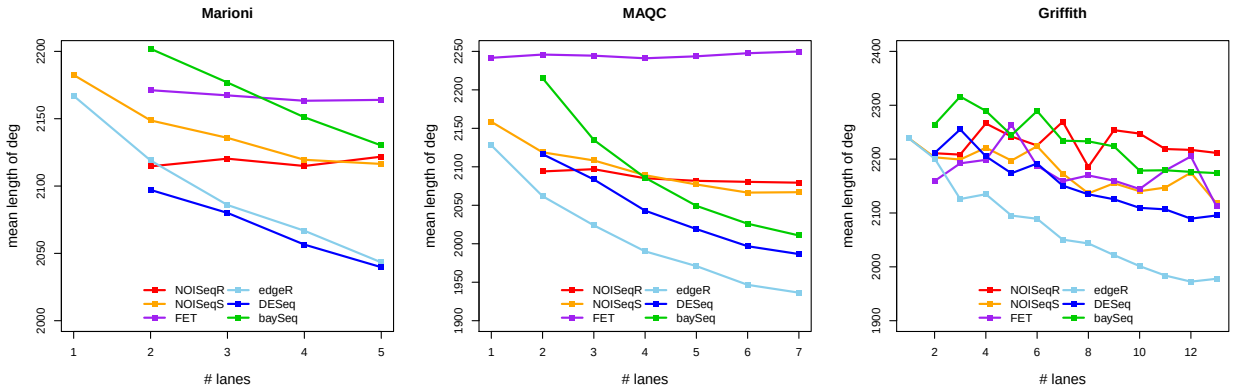

Figure 13a: Mean length of differentially expressed genes according to the number of lanes used and the method, for Marioni's, MAQC and Griffith's datasets. Count data were not normalized by the gene length.

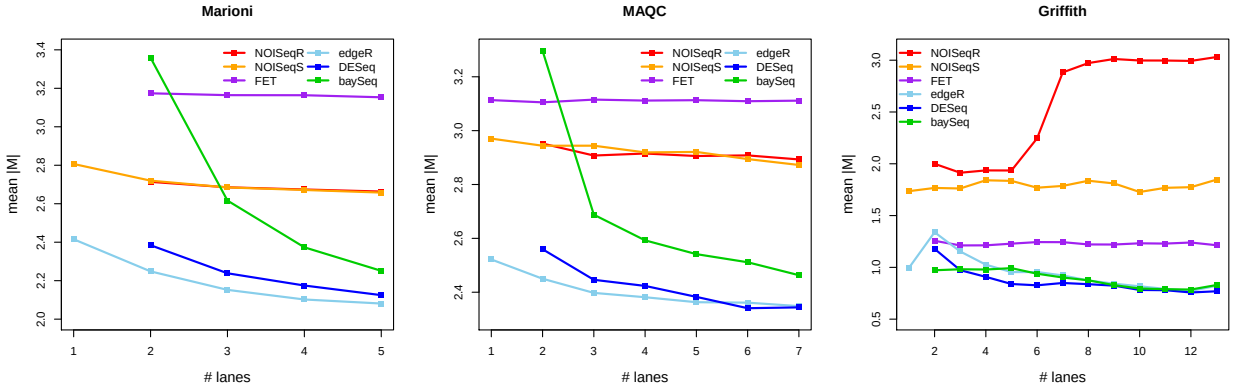

Figure 13b: Mean fold-change (M) of differentially expressed genes according to the number of lanes used and the method, for Marioni's, MAQC and Griffith's datasets. Count data were not normalized by the gene length.

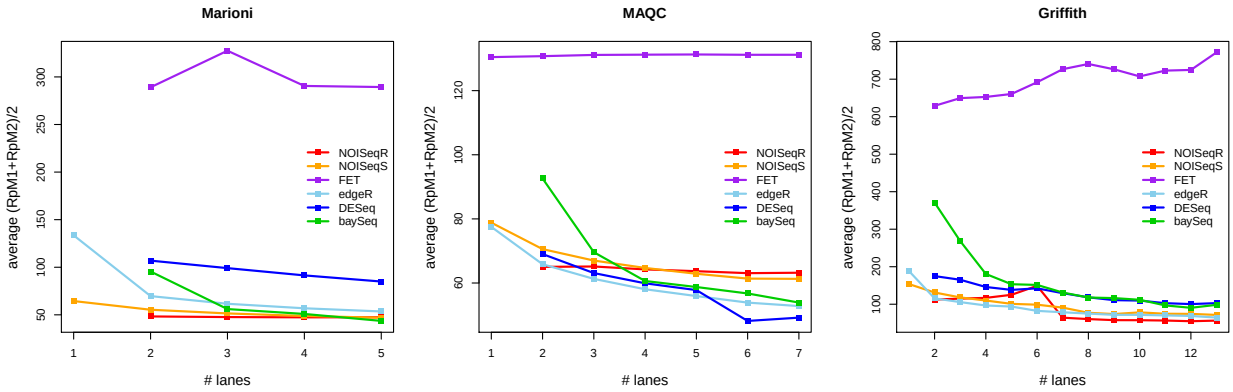

Figure 13c: Mean expression level of differentially expressed genes according to the number of lanes used and the method, for Marioni's, MAQC and Griffith's datasets. RpM_i is the number of reads in condition i per million reads, i.e. $RpM_i = \frac{10^6 \times \text{ gene counts in condition } i}{\text{total counts in condition } i}$. Count data were not normalized by the gene length.

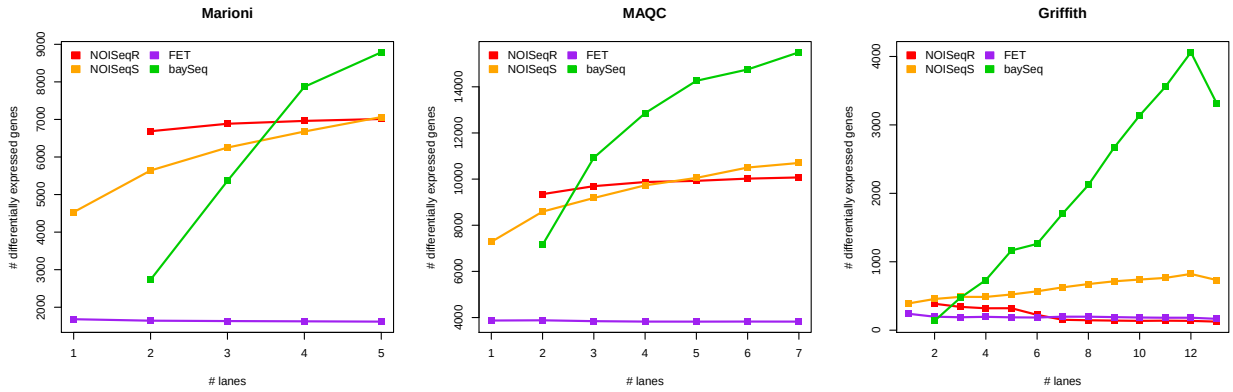

Figure S14: Number of differentially expressed genes according to the number of lanes used, per method, in Marioni's, MAQC and Griffith's datasets. Count data were length-normalized.

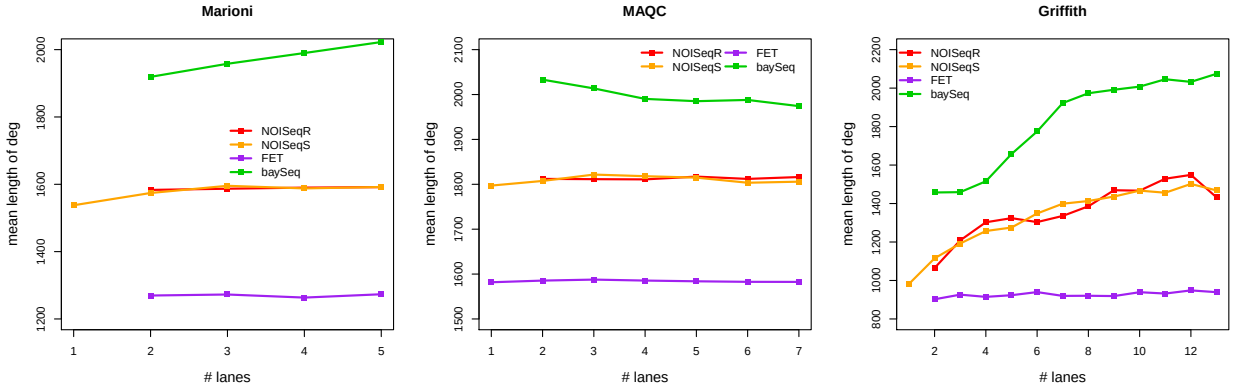

Figure 15a: Mean length of differentially expressed genes according to the number of lanes used and the method, for Marioni's, MAQC and Griffith's datasets. Count data were normalized by the gene length.

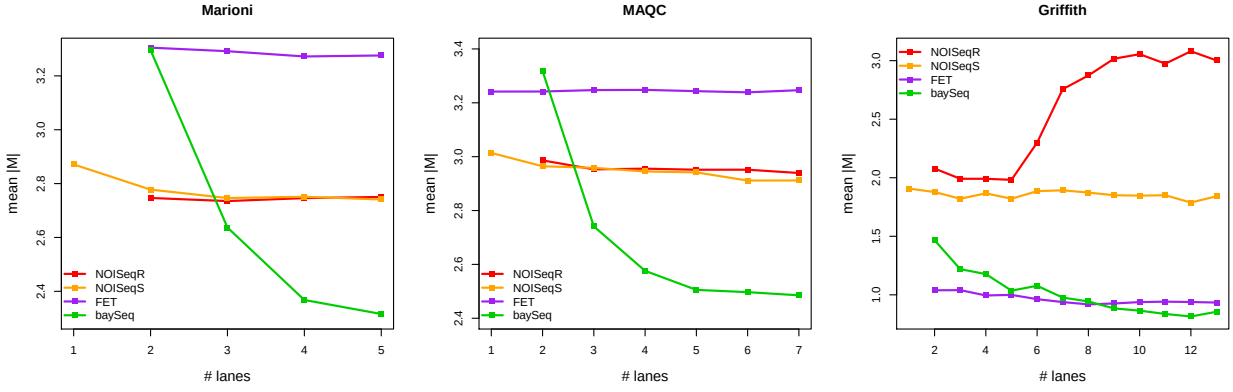

Figure 15b: Mean fold-change (M) of differentially expressed genes according to the number of lanes used and the method, for Marioni's, MAQC and Griffith's datasets. Count data were normalized by the gene length.

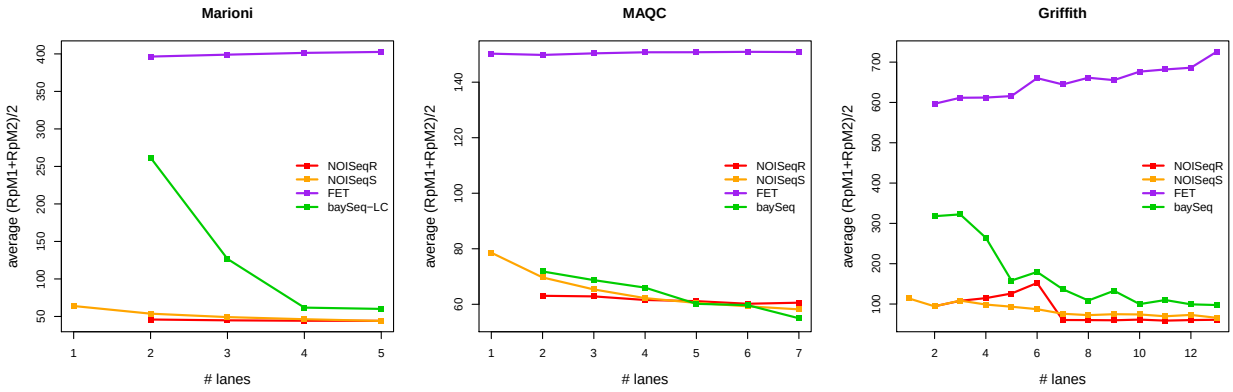

Figure 15c: Mean expression level of differentially expressed genes according to the number of lanes used and the method, for Marioni's, MAQC and Griffith's datasets. RpM_i is the number of reads in condition i per million reads, i.e. $RpM_i = \frac{10^6 \times \text{ gene counts in condition } i}{\text{total counts in condition } i}$. Count data were normalized by the gene length.

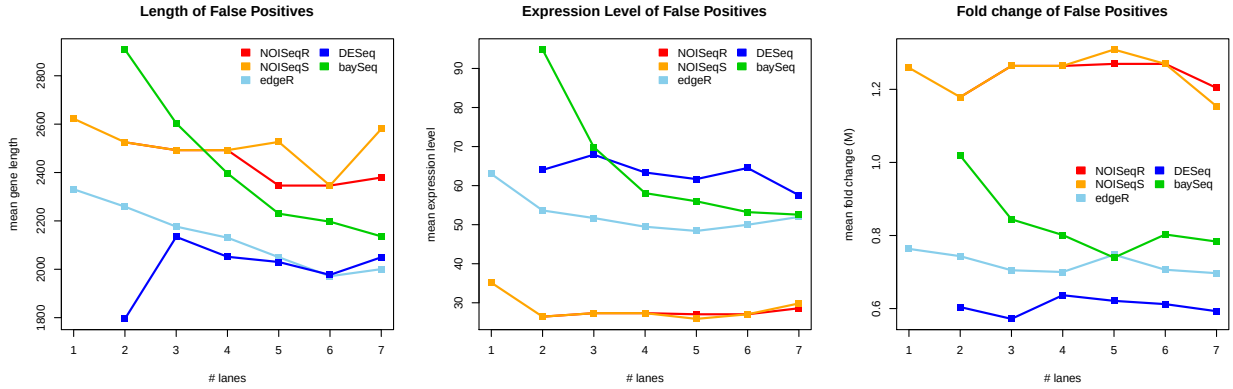

Figure S16a: Properties (length, expression level and M) for false positive genes in RT-PCR MAQC experiment for differential expression methods applied to data with no length normalization.

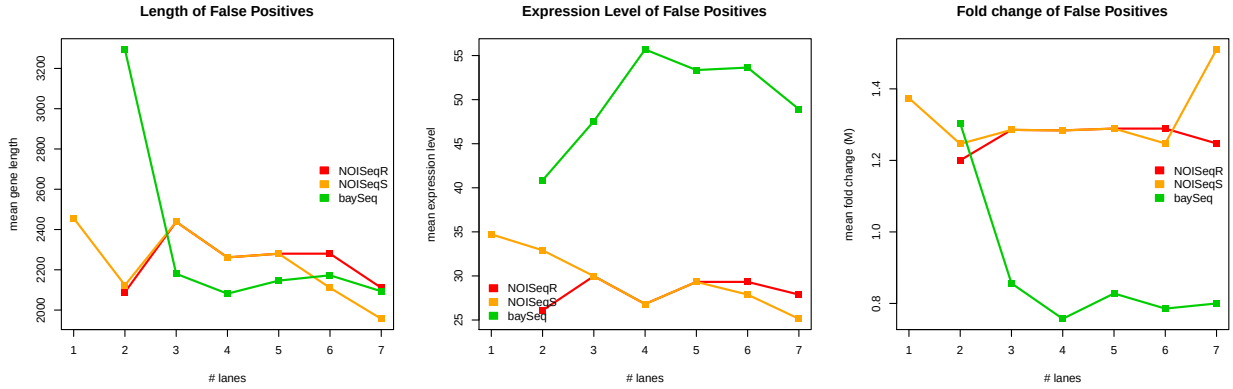

Figure S16b: Properties (length, expression level and M) for false positive genes in RT-PCR MAQC experiment for differential expression methods applied to length-normalized data.

Figure S17: Considering multihits

This work only used uniquely mapping reads to compute gene expression estimates, which is still a common practice in RNAseq analysis. However, discarding multiple hits might have as a consequence the underestimation of large protein families and could imply that some genes would not be detectable at the read lengths used by Marioni and in the MAQC data.

To evaluate how the numbers of differentially expressed genes would change upon a more accurate usage of multihit reads we partially redid Marioni's data analysis to consider these reads. We remapped Marioni's dataset giving to multireads a weight of 1/number of hits, computed gene counts and obtained differential expression calls with different methods.

The following plot (Fig. S17) compares the results for the uniquely mapped reads and for the reassignment of multihits. It can be seen that, in general, the number of differentially expressed genes was very similar at the maximum sequencing depth for uniquely mapped reads and when multihits were considered and that, for lower sequencing depths, the differences were in general bigger. However, the pattern of differences between methods and how these hold along sequencing depth was basically the same considering or not multihits. NOISeq remained the method that was both robust across the number of reads and detected a reasonable number of differentially expressed genes. This result reinforces the suitability of NOISeq to adapt to actual library sizes across different mapping strategies.

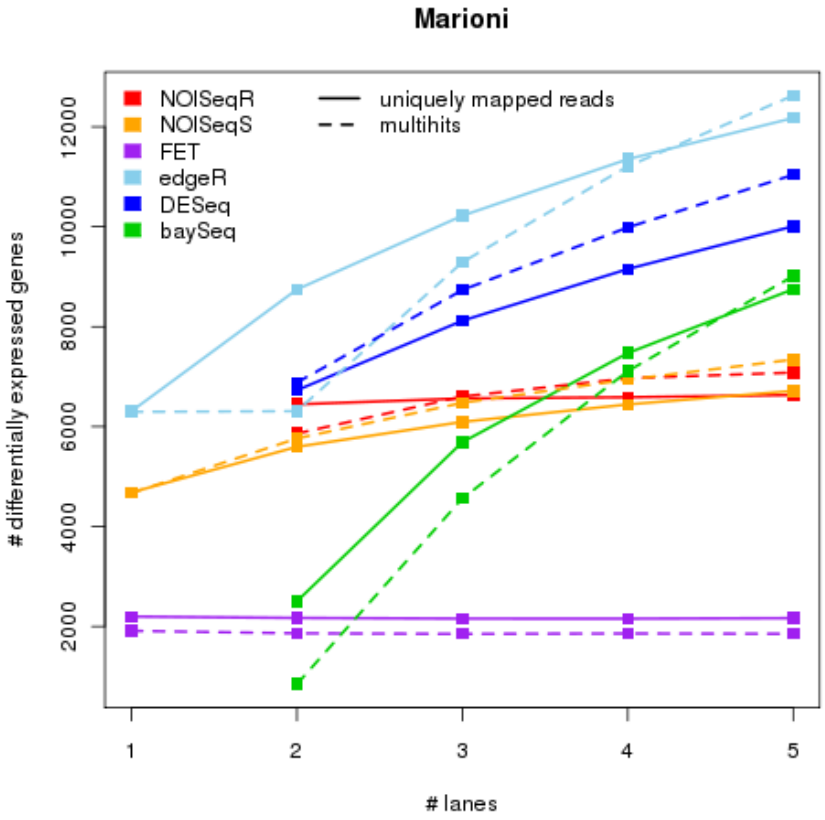

Figure S17: Marioni's dataset. Comparison of the number of differentially expressed genes according to sequencing depth and method when considering or not mapping multihits.