

Differential expression in RNA-seq: a matter of depth

Sonia Tarazona^{1,2}, Fernando García-Alcalde¹, Joaquín Dopazo¹, Alberto Ferrer², and Ana Conesa^{1,*}

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

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

*Corresponding author. Email: aconesa@cipf.es

August 29, 2011

Abstract

Next Generation Sequencing (NGS) technologies are revolutionizing genome research and in particular, their application to transcriptomics (RNA-seq) is increasingly being used for gene expression profiling as a replacement for microarrays. However, the properties of RNA-seq data have not been yet fully established and additional research is needed for understanding how these data respond to differential expression analysis. In this work we set out to gain insights into the characteristics of RNA-seq data analysis by studying an important parameter of this technology: the sequencing depth. We have analyzed how sequencing depth affects the detection of transcripts and their identification as differentially expressed, looking at aspects such as transcript biotype, length, expression level and fold-change. We have evaluated different algorithms available for the analysis of RNA-seq and proposed a novel approach -NOISeq- that differs from existing methods in that it is data-adaptive and non-parametric. Our results reveal that most existing methodologies suffer from a strong dependency on sequencing depth for their differential expression calls and that this results in a considerable number of false positives that increases as the number of reads grows. In contrast, our proposed method models the noise distribution from the actual data, can therefore better adapt to the size of the dataset and is more effective in controlling the rate of false discoveries. This work discusses the true potential of RNA-seq for studying regulation at low expression ranges, the noise within RNA-seq data and the issue of replication.

Introduction

The emergence of next generation sequencing (NGS) has created unprecedented possibilities for the characterization of genomes and has significantly advanced our understanding of its organization. Today, NGS technologies can be used to tackle the de novo sequencing of large genomes (Velasco et al. 2010; Argout et al. 2010; Locke et al. 2011), report individual genome differences within the same species (Durbin et al. 2010), characterize the interaction spectrum of DNA-binding proteins (Park 2009) and create genome-wide profiles of epigenetic modifications (Li et al. 2010). One of the most ground-breaking applications of short-read sequencing is the deciphering of the complexity of the transcriptome. In the last few years the use of RNA-seq technology has resulted in an incredible amount of new data that has dissected isoform and allelic expression, extended 3' UTR regions, revealed novel splice junctions, modes of antisense regulation and intragenic expression (Carninci et al. 2005; Nagalakshmi et al. 2008; Graveley et al. 2010; Trapnell et al. 2010). RNA-seq is also increasingly being used to quantify gene expression, as the number of mapped reads to a given gene or transcript is an estimation of the level of expression of that feature (Marioni et al. 2008).

Although at the dawn of RNA-seq applications it was claimed that this technology would produce unbiased, ready-to-analyze gene expression data, the reality has turned out to be very different. One of the problems that must be faced when dealing with analysis of short reads is that the quantification of expression depends on the length of the biological features under study -genes, transcripts or exons-, as longer features will generate more reads than shorter ones (Oshlack and Wakefield 2009). Common normalization methods, including division by transcript length such as RPKM (Reads Per Kb of exon model per Million mapped reads) from Mortazavi et al. 2008, mitigate but do not completely eliminate this bias (Young et al. 2010). Another drawback is the very nature of the sequencing technology, which is basically a sampling procedure from a population of transcripts, implying that differences in transcript relative distributions between samples will affect the assessment of differential expression (Bloom et al. 2009; Robinson and Oshlack 2010).

Furthermore, the ability to detect and quantify rare transcripts is obscured by the wide dynamic range of mapped reads and the concentration of a large portion of the sequencing output in a reduced number of highly expressed transcripts. However, RNA-seq technology boasts a general high level of data reproducibility across lanes and flow-cells, which reduces the need of technical replication within these experiments (Marioni et al. 2008).

Differential expression methods have also evolved with NGS technologies. Methods traditionally used for microarrays have paved the way to other approaches that take into account the discrete nature of the expression quantification and use different probability distributions to model data (Marioni et al. 2008; Robinson et al. 2010; Anders and Huber 2010; Sultan et al. 2008; Srivastava and Chen 2010; Hardcastle and Kelly 2010). Most of the methodologies proposed so far rely on parametric assumptions and use Poisson or Negative Binomial distributions to model feature counts, following the rationale of the sampling procedure in RNA sequencing. However, the subsequent confirmation of distribution assumptions is important as they might not always hold true (Bullard et al. 2010). Moreover, usually very few replicates, if any, are available, making the estimation of model parameters difficult. Additionally, parametric approaches tend to be problematic for assessing differential expression in low count features (Bullard et al. 2010).

An underlying factor that relates to several of the mentioned problems in RNA-seq analysis is the amount of reads generated in a given experiment. The more the target is sequenced, the more transcripts are identified and the higher the value of the expression level. Although most of the existing analysis methods address this issue by including a correction factor related to library size (Mortazavi et al. 2008; Bullard et al. 2010), higher sequencing rates will presumably result in a more accurate estimation of the expression level and, concomitantly, inferential methods will then enjoy increased power to identify differentially expressed features. As a consequence, our ability to find transcripts and detect differential expression is very much determined by the sequencing depth and this leads to the question of how many reads should be generated in an RNA-seq experiment to obtain robust results. Some recent reports suggest that, in a mammalian genome, ~ 700 million reads would be required to obtain accurate quantification of $> 95\%$ of expressed transcripts (Blencowe et al. 2009), but as yet, there has not been a systematic analysis on how sequencing coverage affects differential expression calls (Oshlack et al. 2010). Knowledge on the relationship between sequencing depth, feature detection and differential expression is needed for experimental design purposes and for understanding the characteristics of the analysis results. In this paper, we set out to gain insight into the effect that sequencing depth has on the statistical analysis of RNA-seq data. We evaluate how this parameter relates to the identification of expressed genes, sequencing noise, transcript length and differential expression. We propose a novel methodology for the assessment of differentially expressed features, NOISeq, that empirically models the noise in count data, is reasonably robust against the choice of sequencing depth and can work in the absence of replication. Our proposal has been tested on three human RNA-seq datasets with different sequencing depths and also on simulated data. We compare NOISeq to published methods for RNA-seq such as Fisher's Exact Test, edgeR (Robinson et al. 2010), baySeq (Hardcastle and Kelly 2010) and DESeq (Anders and Huber 2010).

Results

Saturation, gene length and reads distribution

In RNA-seq technology, saturation would be reached when an increment in the number of reads does not result in additional true expressed transcripts being detected, or in more features called as differentially expressed when two or more conditions are compared. Detection of transcripts can be studied directly on mapped data, while differential expression calls will depend on the statistical methodology of choice. We first evaluated the number of detected genes -defined as genes with more than 5 mapped reads- and the new detections rate NDR -number of newly detected genes in one million additional reads- as a function of the sequencing depth for each of the three datasets used in this study. Note that in this paper the gene is taken as the expression unit, but results can be extended to other features such as transcripts or exons, provided that an appropriate quantification of their expression was obtained.

Mapped reads accumulative plots (Fig. 1) suggest that for all three experiments saturation is not entirely reached, since the number of scored genes keeps on increasing with the number of reads considered. However, as each dataset has a different total readout, NDRs at the deepest coverages are substantially different. While Marioni's data (22 Million reads) end at a NDR of 232 genes, in the MAQC experiment (45 Million reads) this value is 70 and in Griffith's dataset (200 Million reads) drops to 19. It is interesting to note that for a given number of reads NDR values are broadly similar across datasets (for example, in the Griffith data, the NDR at 20 and 45 million reads is 210 and 75, respectively), suggesting that these saturation figures could be indicative of the saturation dynamics of the Solexa technology, at least in human datasets.

[Figure 1 about here.]

We next asked whether this growing detection of genes resulted from the identification of rare transcripts or from the inclusion of (un)specific noise in the data. We evaluated saturation plots for different transcript biotypes, including protein-coding, processed transcript, pseudogenes, miRNAs, tRNAs, rRNAs, snRNAs, snoRNAs and scRNAs (Supp. Table S1). All experimental datasets used in this study followed the standard Illumina protocol for mRNA library preparation (Illumina 2009), that includes poly-A mRNA isolation, RNA fragmentation and size-selection from gel. Therefore, transcripts should be polyadenylated and larger than the size selection cutoff -typically around 200 bps- to be captured by the sequencing procedure. Polyadenylation signals are present in protein coding genes but have also been identified in long range non-coding transcripts (Carninci et al. 2005) and some snoRNAs (Grzechnik and Kufel 2008; Lemay et al. 2010). Expression of pseudogenes is controversial, but reports indicate that these might be transcribed, giving rise to non functional messengers in a tissue specific manner (Zheng et al. 2007). Furthermore, poly-A stretches might be present in retrotransposed pseudogenes that originate from genome insertion events of transcribed messengers (Zheng et al. 2007). Poly-A tails are also added to pri-miRNAs, nascent miRNA transcripts that undergo processing to reach the mature miRNA state (Kim et al. 2009). Although pri-miRNAs can be long molecules, they are of transient nature and miRNAs are typically not captured by mRNA-seq library preparation protocols. Alternatively, miRNAs embedded in introns of coding genes could still be sequenced from partially processed transcripts. Other RNAs such as tRNAs, snRNA, snoRNA and rRNA may undergo cytoplasmic polyadenylation as targeting for degradation (Anderson 2005; Slomovic et al. 2006). Additionally, rRNA depletion usually precedes mRNA preparation and rRNA presence is considered as contamination in mRNA-seq experiments. In general, these small RNA species can be considered as not targeted by the mRNA-seq procedure.

[Figure 2 about here.]

As expected, for all datasets, the protein-coding biotype represented the large majority of the detected transcripts (60-70%). Other species such as pseudogene, processed-transcript and lincRNA were also readily found (Fig. 2A and Supp. Fig. S1), whereas small RNAs were only marginally detected. The distribution of biotypes observed among detected features evolved with increasing sequencing depths, with the relative abundance of protein-coding transcripts steadily decreasing, whereas non-coding genes gained a proportional presence (Fig. 2B and Supp. Fig. S2). Moreover, transcript-type specific saturation curves indicated that the coding transcriptome was more successful in reaching relative saturation than other relevant transcript species, that progressed with more steep detection curves, and that in ultra-highthroughput sequencing datasets, such as the Griffith's experiment, a non-negligible percentage of off-target RNA species might also be identified (Fig. 2C and Supp. Fig. S3). Removing small RNA intronic reads from mapping data did not alter observed saturation dynamics (Supp. Fig. S4).

Finally, we also observed a sequencing-depth dependency for the length of detected transcripts. This effect was more pronounced for lincRNAs, processed-transcripts and pseudogenes than for protein-coding RNAs (Fig. 2D and Supp. Fig. S5), which may be a consequence of the lower count value of non-coding RNAs that would create a strong dependence between transcript length and detection. However, in all four biotypes, the median length of the identified genes was always larger than the targeted genome median for that biotype, indicating a general bias of the technology towards longer transcripts.

Taken together, this analysis suggests that a relatively stable detection of protein-coding genes is reached at moderate sequencing depths and that ultra high-throughput sequencing mainly benefits the detection of non-coding, low-expression RNAs of putative regulatory function, but might also result in the sequencing of off-target transcript species which in turn has an influence in the relative proportion of transcript types. Therefore we concluded that for differential expression analysis, a balanced sequencing depth between conditions is advisable. We also suggest using the “per-biotype transcript detection” and “length” accumulative curves to estimate the saturation and contamination levels of any particular mRNA-seq dataset. Finally, we must highlight that only human datasets were used in these analyses and therefore the presented figures are conditioned by the magnitude of the human transcriptome.

Differential expression

Once we obtained a comprehensive picture of how NGS library size affects the identification of expressed genes, we next asked how the available number of reads influences the capacity of this technology to detect gene expression changes. In this section, we introduce the NOISeq algorithm and evaluate the behavior of this and other differential expression methods in relation to sequencing depth.

NOISeq is a novel non-parametric approach for the identification of differentially expressed genes (d.e.g.) from count data that aims to be robust against the number of available reads. Essentially, NOISeq creates a

null or noise distribution of count changes by contrasting fold-change differences (M) and absolute expression differences (D) for all the genes in samples within the same condition. This reference distribution is then used to assess whether the M - D values computed between two conditions for a given gene are likely to be part of the noise or represent a true differential expression (Fig. 3A). In practice NOISeq creates the noise distribution by joining (M, D) values from all possible pair-wise comparisons between replicates of either condition (see Methods for further details).

Two variants of the method were implemented: NOISeq-real uses replicates, when available, to compute the noise distribution and, NOISeq-sim simulates them in absence of replication. It should be noted that the NOISeq-sim simulation procedure assimilates to technical replication and does not reproduce biological variability, which is necessary for population inferential analysis. However, current mRNA-seq experiments are still sparse in replication, thus the ability of statistical methods to work with technical replicates or in their absence altogether, is relevant. Simulation in NOISeq-sim is basically controlled by two parameters: the number of simulated samples or replicates (nss) and the size of each replicate, given as a percentage of the total number of reads (pnr). We determined that NOISeq-sim worked best when at least five replicates were simulated and replicate size was 20% of the total amount of reads in the corresponding condition. With these parameters NOISeq-sim resulted in similar differential expression calls as NOISeq-real, with a slight higher detection rate for the simulation version of the algorithm (Supp. Material).

Performance assessment of mRNA-seq differential expression methods

We compared NOISeq to a selection of RNAseq differential expression methods obtained after evaluation with simulated data (Supp. Material), namely edgeR (Robinson et al. 2010), baySeq (Hardcastle and Kelly 2010), DESeq (Anders and Huber 2010) and FET (Fisher's Exact Test). These are all parametric approaches (except for FET) in contrast to NOISeq for which no assumptions are made on the distribution of the M and D statistics. All methodologies were applied to the three benchmarking datasets. Moreover, both MAQC and Griffith's experiments included RT-PCR measurements for a number of genes. In these two cases, we identified positive (RT-PCR differentially expressed) and negative (RT-PCR non differentially expressed) genes following the same previously reported procedure (Bullard et al. 2010; Griffith et al. 2010) (see Methods) and used them to obtain performance plots. We also included the analysis of gene length corrected data with methods that permitted this input. Note that FET was applied on counts normalized by library size.

On the MAQC dataset, two performance indicators, Precision-Recall Curves (PRC) and False Discovery Rate (FDR) indicated a better behavior of NOISeq compared to other methodologies (Fig. 3B). Specifically, false discoveries were higher for edgeR, DESeq and baySeq. FET had a low FDR regardless of the significance threshold but also showed a poorer precision-recall figure. Interestingly, PRC and FDR were very similar on data with and without length correction. Griffith's RT-PCR data were more limited but led to the same conclusions (Supp. Table S3).

In summary, our performance analysis highlighted differences between RNA-seq differential expression methods when using a fixed library size and pointed to NOISeq as a high performing methodology. We next investigated how these methods behaved with different numbers of mapped reads.

[Figure 3 about here.]

Differential expression and sequencing depth

Comparing statistical approaches were applied to each experimental dataset taking an increasing number of lanes until the nominal sequencing depth of the experiment was reached. In the case of Griffith's data, only half of the lanes were used from the sensitive cell line to equilibrate sequencing depth in both samples to around 100 million reads. As different methodologies use different parameters to select significant features it was not always clear which cutoff values would produce comparable analysis scenarios. In this study we chose $q=0.8$ for NOISeq, a probability of 0.999 for baySeq and an adjusted p-value threshold of 0.001 for the remaining methods. Less restrictive values for compared methodologies resulted in far too large a number of selected genes. We performed our study using library size normalized count data, as all evaluated methods allowed this possibility. Next, we introduced feature length normalization into the analysis for those methodologies that permitted this option.

Sequencing depth dependence in number and type of differential expression calls

[Figure 4 about here.]

We first investigated the number of differential expression calls as a function of the sequencing depth (SD) (Fig. 4 and Supp. Tables S4). A very pronounced dependency between gene selection and read

number was observed for edgeR, DESeq and baySeq. FET did not show this dependency but did identify a reduced number of significant genes. NOISeq had an intermediate behavior with a moderate number of d.e.g. in Marion and MAQC datasets, and increased only slightly with SD. Results for Griffith's data were slightly different. While FET and NOISeq identified a small number of significant genes (between 150 and 200), close to the figure reported in the original paper, other methods resulted in larger selection sets. Moreover, both FET and NOISeq-real lost significant calls as more lanes were considered, reflecting the high variability of this dataset. We then looked at differential expression curves by transcript biotype and noticed that for parametric approaches a significant and increasing number of off-target transcripts were selected as more reads were considered (Supp. Fig. S11), whereas NOISeq again behaved moderately here. In fact, NOISeq significant calls were the most enriched in protein-coding genes, where other methods included higher proportions of non-polyadenylated transcripts (Supp. Fig. S12).

Sequencing depth influence on length, expression and fold-change of significant genes

[Figure 5 about here.]

To better understand how SD affects other properties of differential expression we plotted the transcript length, fold-change (M) and mean expression level of significant genes as a function of the available number of reads (Fig. 5 and Supp. Fig. S13). The pattern of differences between methods was similar to that observed in previous analyses. The edgeR, DESeq and baySeq methods showed SD dependency whereas NOISeq and FET did not. FET had large and constant values for these three parameters.

In the parametric approaches, the mean transcript length of statistically significant genes decreased as the number of lanes grew. This length shortening effect was only very moderately present in NOISeq which, at the highest sequencing depths, generally selected larger genes than the other methods. This difference is in agreement with the observed higher selection of small non-coding RNAs by the parametric approaches. Furthermore, the mean fold-change of the genes detected by compared methodologies was greatly influenced by the total read number. The larger the sequencing output the smaller the count differences between samples declared as significant, and this was specially notable in the large Griffith's dataset (100 million reads), where mean M values for differentially expressed genes dropped below 1. NOISeq, on the contrary, selected genes with larger count differences and had a robust behavior with changing sequencing depth. Finally, we also observed a strong dependency on the level of expression. Current RNA-seq statistical methods tend to identify genes with a lower relative abundance as the number of available reads grows. Again here NOISeq, and especially NOISeq-real, offered a more constant and intermediate result, selecting genes with lower expression at smaller sequencing depths and genes with larger count numbers at higher depths than parametric RNA-seq methods.

Most statistical analysis methods for RNA-seq suffer from high false discovery rates

All previous results indicated that d.e.g. identified by parametric approaches strongly increase in number as more sequencing is generated and that this results in calling significant genes with smaller fold changes. Although this could be explained by an apparent higher accuracy of gene expression estimates in large sampling sizes, the prominent discrepancy with a data-driven methodology such as NOISeq and the results of our initial performance analysis led us to suspect a general failure of those methods in controlling FDR as the sequencing output increase. To verify this, we analyzed the available MAQC RT-PCR data as a function of the SD, looking both at the false (FPR) and true (TPR) positive rates. As suspected, current RNA-seq analysis methods progressively incorporated false calls as more sequencing data were used, reaching above 60% of false positives in edgeR (Fig. 6). In contrast, NOISeq maintained a stable and low FPR throughout the increasing number of lanes. Only FET had better FPR performance, however at a significant cost of the number of true detections. The TPR obtained from the other compared methods was slightly higher than that of NOISeq, which is logically the consequence of the large number of the d.e.g. called by these methodologies. Furthermore, we verified that false positives were basically genes with shorter length, decreasing expression level and smaller fold-change differences at each SD value (Supp. Fig. S16a). Notably, genes selected in common by NOISeq and other approaches did contain a functional signature, i.e. were significantly enriched in many biological functions while those only detected by parametric methods had no specific functional charge (Supp. Material).

[Figure 6 about here.]

Effect of normalization by feature length on sequencing depth biases

Lastly, we evaluated whether normalization of count data by a feature length correction method, such as RPKM, affected the observed patterns of sequencing depth dependence. We introduced length normalization

into NOISeq-sim, NOISeq-real, FET and baySeq and repeated our analysis (edgeR and DESeq packages do not allow for this correction). Figures were essentially the same as in non length-normalized data regarding number (Supp. Fig. S14), mean fold-change and mean expression value of d.e.g. (Supp. Fig. S15 b and c). However, the dependence between library size and transcript length was significantly changed and all methodologies showed now a constant behavior and a shorter mean length value than non-normalized counterparts (Supp. Fig. S15a). Finally, False and True Positive curves for MAQC data (Fig. 7A, 7B and Supp. Fig. S16b) again resembled previous results: baySeq increasingly detected false positives with increasing sequencing depth and FET and NOISeq maintained a low level of true positive detection.

[Figure 7 about here.]

Discussion

Estimation of gene expression levels by sequencing is conceptually simple and has been seen as a very straightforward task. Sequencing reads the population of RNA molecules in a given sample and renders a direct quantification of the abundance of each transcript, mapping ambiguities and sequencing errors issues apart. Although this is fundamentally true, as shown in studies on correspondence of RNA-seq data with microarrays and RT-PCR (Bullard et al. 2010; Griffith et al. 2010; Marioni et al. 2008), we believe that there is still some work to be done to fully understand the characteristics of RNA-seq data and their processing by statistical methods. One of the biases that rapidly became evident was the effect of transcript length in the quantification and identification of differential expression. The nature of the short read procedure makes it inevitable that longer transcripts will be preferentially detected over shorter ones, and this has been shown to have implications in the biological interpretation of the data (Oshlack and Wakefield 2009; Young et al. 2010). Another important element is the magnitude of the depth of the sequencing experiment, the subject of this study. Due to the large dynamic range of gene expression, ultra-high throughput sequencing seems advisable to detect transcripts with low expression values. However, we have seen that, as more sequencing output is considered, the diversity and quantity of detected off-target RNA species, such as several types of small RNAs, also increases (Fig. 2B). The extent to which each of these biotypes and transcripts are purification artifacts or have a biological significance warrants a separate study but it does show an important property of RNA-seq data: the effect that sequencing depth has on the distribution of reads among transcripts and the quantification of expression, essentially a percentage in the case of this technology. Robinson and Oshlack (2010) have already highlighted the implications that different transcript distributions might have in RNA-seq normalization and differential expression. Our observations suggest that it is advisable to take equal sequencing depths between samples in order to support accurate statistical analysis.

We have evaluated several RNA-seq differential expression methods regarding their behavior throughout sequencing depths: edgeR, DESeq, baySeq, the traditional Fishers Exact Test (FET) and a novel method proposed here: NOISeq. edgeR, DESeq and baySeq use the Negative Binomial (NB) distribution. The first two apply an exact test, while baySeq is a Bayesian method. NOISeq creates an empirical distribution of count changes adapted to the available data, from which the probability of differential expression for each feature can be derived. In this non-parametric approach, differential expression does not rely on individual transcript measurements, but in the joint distribution of M - D values for all the features within the dataset. We studied the effect of sequencing depth on the number of differentially expressed genes, their length, fold-change value and expression level. The pattern produced by NOISeq and FET was more constant across the different variables analyzed, whereas the other three methods showed a pronounced dependence. The parametric approaches strongly increased the number of significant calls as more sequencing output was included, resulting in a considerable number of false positives (Fig. 6). The newly detected genes were shorter, of lower relative expression, and had smaller fold-change differences than those obtained with less data, and contained many off-target RNA species (Fig. 5 and Supp. Fig. S12). False positive genes identified in the analysis of the MAQC data had similar characteristics, suggesting that large library size datasets analyzed by these parametric approaches incorporate many falsely called significant genes at the low expression range and/or with small fold-change differences. The constant pattern of FET was intrinsically due to a low detection power that identified only highly expressed transcripts. However, NOISeq showed more robustness against these sequencing depth biases while maintaining a high true positive detection rate. We believe that, given the number of reads sequenced and the specific characteristics of the data analyzed, this approach creates a more realistic estimation of the probability that a given count difference will occur by chance, and also results in the stable control of false positives. The compared parametric approaches do not have this flexibility and tend to render small fold-changes as significant when sequencing numbers grow.

One striking difference in the way the two types of methods work relates to how differential expression calls increased. edgeR, DESeq and baySeq added new significant genes to the pool of already detected features with each new lane summed into the library size. In contrast, NOISeq selected new but also discarded

some genes (data not shown), depending on how the variability introduced by the additional sequence input reshaped the noise distribution. We believe that this property makes our approach robust to large count values and helps to control false discovery rates. This aspect was especially notable when working with Griffith's dataset. Variability between lanes was surprisingly large if compared to the other two datasets—which resulted in fewer significant genes declared at the highest sequencing depths (Fig. 7) and an erratic behavior when considering other parameters analyzed. High technical reproducibility has been claimed for the RNA-seq technology (Marioni et al. 2008; Mortazavi et al. 2008) but our observations suggest that this should be checked for each dataset. Unfortunately, the cancer study only provided us with a reduced number of negatives upon which to evaluate sequencing depth-related trends, however RT-PCR data in this study also indicated a higher FDR for NB-based methods than for NOISeq (Supp. Table S3), again indicating a large artefactual gene selection by those methods in this dataset. Moreover, biological replicates (which remain uncommon in RNA-seq analysis) are expected to have higher variability rates. The nature of the NOISeq methodology, in particular NOISeq-real, makes it a suitable approach for accounting for the variability of biological replication. On the other hand, it is important to remember that inferential approaches such as those implemented in edgeR, DESeq and baySeq rely on the analysis of biological replication to achieve their true competency, and therefore performance results of these methods using technical replicates might not be completely applicable to biological replicates.

With regards to the two variants of NOISeq, overall NOISeq-sim and NOISeq-real performed similarly throughout the whole study, although a slightly higher detection rate and sequencing depth dependency was observed with NOISeq-sim. The two variants were more different at Griffith's data. These results indicate that the simulation procedure of NOISeq-sim works well to replace technical replicates but may tend to overestimate d.e.g. in data with high variability among replicates.

We also analyzed how normalization by transcript length modified our conclusions. In general, figures were equivalent when the different statistical methods were applied to length-normalized data (Supp. Fig. S14 and S15), except for the sequencing-depth (SD) influence on the length of significant genes, which was not observed. Other SD biases, such as relative expression, fold-change differences and FDR were maintained, indicating that the tendency towards the detection of shorter genes when using larger libraries is simply the consequence of lower relative expression rather than length itself, since normalization of expression value by length eliminated, or reduced, this bias. Other normalization procedures, such as Upper Quartile (Bullard et al. 2010) or TMM (Robinson and Oshlack 2010), have been proposed and it remains to be studied how sequencing depth influences results in these cases.

This study raises the question of the true potential of RNA-seq to investigate the regulation of rare transcripts. Our results indicate that although deep sequencing effectively enhances our view on the diversity of the transcriptome, the identification of true differential expression at a low count range might not be so easy to achieve. More reads imply the detection of more genes, but also result in noisier data, which makes the assessment of differential expression increasingly difficult. This is suggested by the observation that NOISeq, which models noise on the actual number of reads, does not indefinitely increase selection of low count-number transcripts as sequencing depth grows, and by the fact that increasing library sizes confines the false positive calls to low expressed genes (Fig. 6). Undoubtedly, improvements in RNA-seq library preparation protocols, sequencing accuracy and mapping precision will help to reduce noise and improve differential expression analysis. However, the distribution of count differences within one RNA-seq sample will still be influenced by the nature of short-read technology and the characteristics of the analyzed transcriptome. For example, we repeated our analysis considering allocation of multihit reads and although slight variations in d.e.g. numbers occurred, the pattern of sequencing depth dependency showed in this study remained (Supp. Fig. S17). We believe that the NOISeq method is an effective strategy to capture the variability of count data and provide the statistical framework for differential expression assessment.

In conclusion, this work sheds new light on the properties of RNA-seq and points to important issues that should be evaluated when developing new approaches for the statistical analysis of these data.

Methods

Datasets

Three publicly available human RNA-seq datasets with different sequencing depths were used in this study. Marioni's pioneering work (Marioni et al. 2008) compares gene expression in kidney and liver tissues and has a sequencing depth of around 20 million reads (distributed in 5 lanes) for each sample. The MAQC dataset (Bullard et al. 2010; Shi et al. 2006) was generated for benchmarking purposes on RNA-seq. It consists of two samples: Ambion's human brain reference RNA (Brain) and Stratagene's human universal reference RNA (UHR). Each sample comprises seven lanes, providing 42 and 45 million reads respectively. This project additionally has RT-PCR data for validation of RNA-seq analysis results. The third dataset was published

by Griffith et al. 2010 and contains 96 and 198 million paired-end reads, respectively, of the transcriptome of two human colorectal cancer cell lines only differing in the fluorouracil (5-FU) resistance phenotype. Also here RT-PCR data were available for a number of genes.

In all the three experiments Solexa technology was used. Raw *fastq* files were downloaded from the SRA archive (Leinonen et al. 2011) and mapped against the *Homo sapiens* high coverage assembly *Hg19* from Ensembl (Flicek et al. 2011) using Tophat (Trapnell et al. 2009), allowing up to 2 mismatches and discarding reads mapping at multiple locations. Counts for each gene were computed by means of the HTSeq Python package (Anders 2010) using the annotation of the Ensembl genes (version 60) and only exonic reads. This was also used to obtained biotype for each gene, as well as a corresponding length value computed as the median length of its annotated transcripts.

Differential expression method: NOISeq

NOISeq method computes differential expression between two conditions given the expression level of the considered features. In this study, the gene was used as the expression unit, although the methodology can be equally applied to transcripts or exons, provided the quantification of their expression is supplied. The gene expression level is the number of reads or in the library mapping to a gene, i.e. the read counts.

Let c_{gj}^i be the number of read counts for each gene i in the j -th sample (or replicate or lane) from the experimental condition or group g ($g = 1$ or 2), where j varies from 1 to the number of samples in group g . Then, the library size or sequencing depth s_{gj} can be computed as the sum of counts c_{gj}^i over all the genes for the j -th replicate in experimental condition g . In order to avoid library size bias, the NOISeq method corrects the counts by a factor closely related to the sequencing depth. The default option is taking the number of counts per million reads, so the corrected expression values would be $x_{gj}^i = c_{gj}^i \times 10^6 / s_{gj}$. Other implemented normalization techniques are the Upper Quartile (UQUA) from Bullard et al. 2010, the Trimmed Mean of M values (TMM) from Robinson and Oshlack 2010 or RPKM from Mortazavi et al. 2008 (when the length of the features is provided). Regardless of the normalization procedure used, NOISeq permits applying a feature length correction which consists of dividing the expression level by a factor equal to any power of the feature length. NOISeq also accepts processed expression values instead of gene counts to allow other normalization procedures.

Hence, NOISeq takes these corrected values or pseudo-counts x_{gj}^i to obtain the statistics needed to derive differential expression. Let x_g^i be the expression value that summarizes all the replicates in the experimental condition g . In the case that there are no replicates at all, x_g^i is the corrected expression value. When technical replicates are available, $x_g^i = \sum_j x_{gj}^i$. If biological replicates are used, x_g^i is computed as the mean or median of the x_{gj}^i for all the replicates.

The differential expression statistics in NOISeq are the log-ratio (M) and the absolute value of difference (D). These statistics collect the information on fold-change and also the absolute pseudo-counts difference, thereby compensating the unstable behavior of M at low expression values. They can be defined for a certain gene i as $M^i = \log_2 \left(\frac{x_1^i}{x_2^i} \right)$ and $D^i = |x_1^i - x_2^i|$.

To avoid the indetermination in calculating M when expression level is 0, zero counts were replaced by $k=0.5$ before normalization. The k parameter can also be set by the user or, if normalized counts are provided, calculated as the middle point between 0 and the minimum expression value for detected genes. In addition, genes with 0 counts in all the replicates and conditions are excluded from the analysis, considering that they are obviously not expressed.

Once M and D values have been obtained for each gene, a threshold for these values must be established in order to classify genes as differentially or non-differentially expressed. A gene is considered to be differentially expressed if the corresponding M and D values are very likely to be higher than noise values. The M and D probability distribution in noise data is computed by contrasting gene counts within the same experimental condition. To obtain this distribution, each replicates pair are considered and values are pooled together. Absolute values of M are used, since the sign of changes is arbitrary and only the magnitude of the change is biologically meaningful.

Let M^* and D^* be the random variables describing noise distribution. Let G^i be a random variable which takes the value 1 if gene i is differentially expressed between two experimental conditions and 0 when it is not. We are interested in determining the probability of G^i taking a value of 1. A gene i has been considered to be differentially expressed when the corresponding values for $|M|$ and D ($|m^i|$ and d^i) are likely to be higher than in noise ($|M^*|$ and D^* values). Then, the probability of a gene being differentially expressed given the expression levels in both conditions can be written as follows:

$$P(G^i = 1 | x_1^i, x_2^i) = P(G^i = 1 | M^i = m^i, D^i = d^i) = P(|M^*| < |m^i|, D^* < d^i) \quad (1)$$

Thus, the probability of not being differentially expressed between the two conditions can be easily derived

as: $P(G^i = 0|M^i = m^i, D^i = d^i) = 1 - P(|M^*| < |m^i|, D^* < d^i)$. The odds $P(G^i = 1|M^i = m^i, D^i = d^i) / P(G^i = 0|M^i = m^i, D^i = d^i)$ may be used to decide whether a gene is differentially expressed between the two conditions or not. For instance, an odds value of 4:1 is equivalent to $P(G^i = 1|M^i = m^i, D^i = d^i) = 0.8$ and it means that the gene is 4 times more likely to be differentially expressed than non-differentially expressed. This is the probability threshold we used throughout the paper.

As it has been stated above, the NOISeq algorithm compares replicates within the same condition to estimate noise distribution. Two versions of NOISeq method have been developed: NOISeq-real computes noise from replicates when these are available, and NOISeq-sim simulates technical replicates from the data.

NOISeq-real.

The algorithm estimates the probability distribution for M^* and D^* in an empirical way, computing M and D values for every pair of replicates within the same experimental condition and for every gene. Then, all these values are pooled together to generate the noise distribution. Two replicates in one of the experimental conditions is sufficient to run the algorithm. If J_g is the number of samples in experimental condition g , the number of comparisons within this condition would be $\binom{J_g}{2}$. If $\binom{J_g}{2}$ is higher than 30, in order to reduce computation time, 30 pairwise comparisons are randomly chosen out of these $\binom{J_g}{2}$ when estimating noise distribution. It should be noted that biological replicates are necessary to make any inference about the population. whole population.

NOISeq-sim.

When there are no replicates for any of the experimental conditions, the algorithm can simulate them. The simulation relies on the assumption that read counts follow a multinomial distribution, where probabilities for each class (gene) in the multinomial distribution are the probability of a read to map to that gene. These mapping probabilities are approximated using counts in the only sample of the corresponding experimental condition. Counts equal to zero are replaced with $k=0.5$, to give all genes some chance to appear. Given the sequencing depth of the unique available sample, sequencing depth for the simulated samples is generated randomly from a uniform distribution in the interval $[(pnr-v)*s_g, (pnr+v)*s_g]$. The parameter pnr determines the number of reads of each simulated replicate and is a percentage of the sequencing depth s_g of the available sample g , and v is a parameter representing the variability of sequencing depth across samples. Both parameters can be chosen by users. NOISeq-sim also allows users to choose the number of replicates to be simulated (nss). We recommend $nss \geq 5$, $pnr=0.2$ and $v=0.02$.

NOISeq has been implemented in the statistical language R and is available at <http://bioinfo.cipf.es/noiseq>.

Validation of differential expression calls

RT-PCR data available from MAQC and Griffith's experiments were used to evaluate performance of statistical methods. *Positive* and *negative* RT-PCR differentially expressed genes were obtained directly from the original works and matched to Ensembl IDs. After discarding replicates and eliminating unmatched genes, a total of 330 and 82 *positive* genes and 83 and 12 *negative* genes for MAQC and Griffith's dataset, respectively, were taken to compute True and False Positive Rates (TPR and FPR). Additionally, Precision-Recall Curves (PRC) and False Discovery Rate (FDR) plots were generated both for simulated and RT-PCR datasets. "Recall" is the true positive rate (TPR) and "Precision" is defined as $TP/(TP+FP)$, so it is equal to $1-FDR$. PRC are good performance estimators when the number of negatives greatly exceeds the number of positives, as is the case of expression datasets (Davis and Goadrich 2006).

Acknowledgments

This research was supported by grants BIO2008-05266-E and BIO2008-04638-E from the Spanish Ministry of Science and Innovation (MICINN), in the frame of ERA-Net Pathogenomics; grant BIO2009-10799 from the MICINN; BIO2008-04212 from the MICINN and PROMETEO/2010/001 from the GVA-FEDER. We also thank the support of the National Institute of Bioinformatics (www.inab.org) and the CIBER de Enfermedades Raras, both initiatives of the ISCIII, MICINN. This work is also partly supported by a grant (RD06/0020/1019) from Red Tematica de Investigacion Cooperativa en Cancer (RTICC), ISCIII, MICINN.

Conflict of interest statement.

None declared.

List of Figures

1	Saturation curves	11
2	Detection and sequencing depth	12
3	NOISeq method: description and performance	13
4	Differential expression methods and sequencing depth	14
5	Gene length, M and expression level	15
6	TP and FP vs sequencing depth	16
7	Differential expression on length-normalized data	17

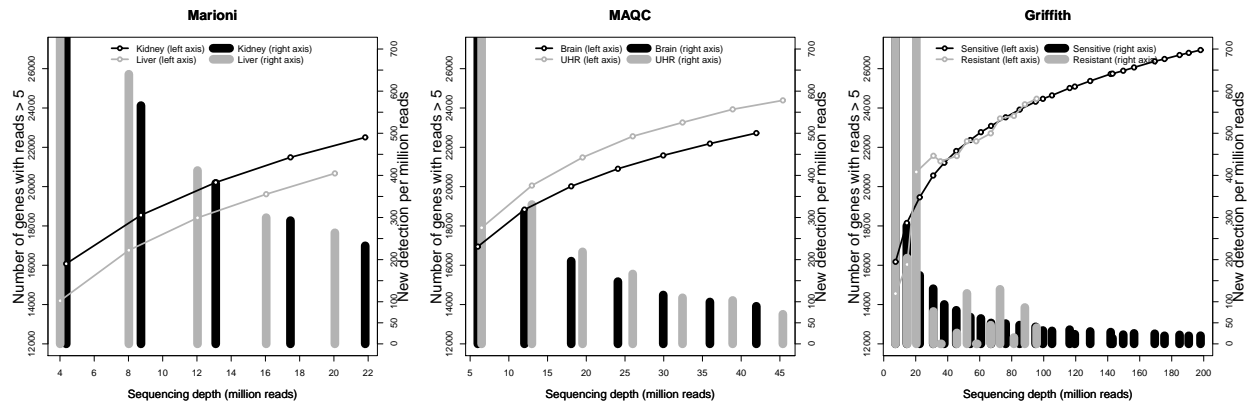


Figure 1

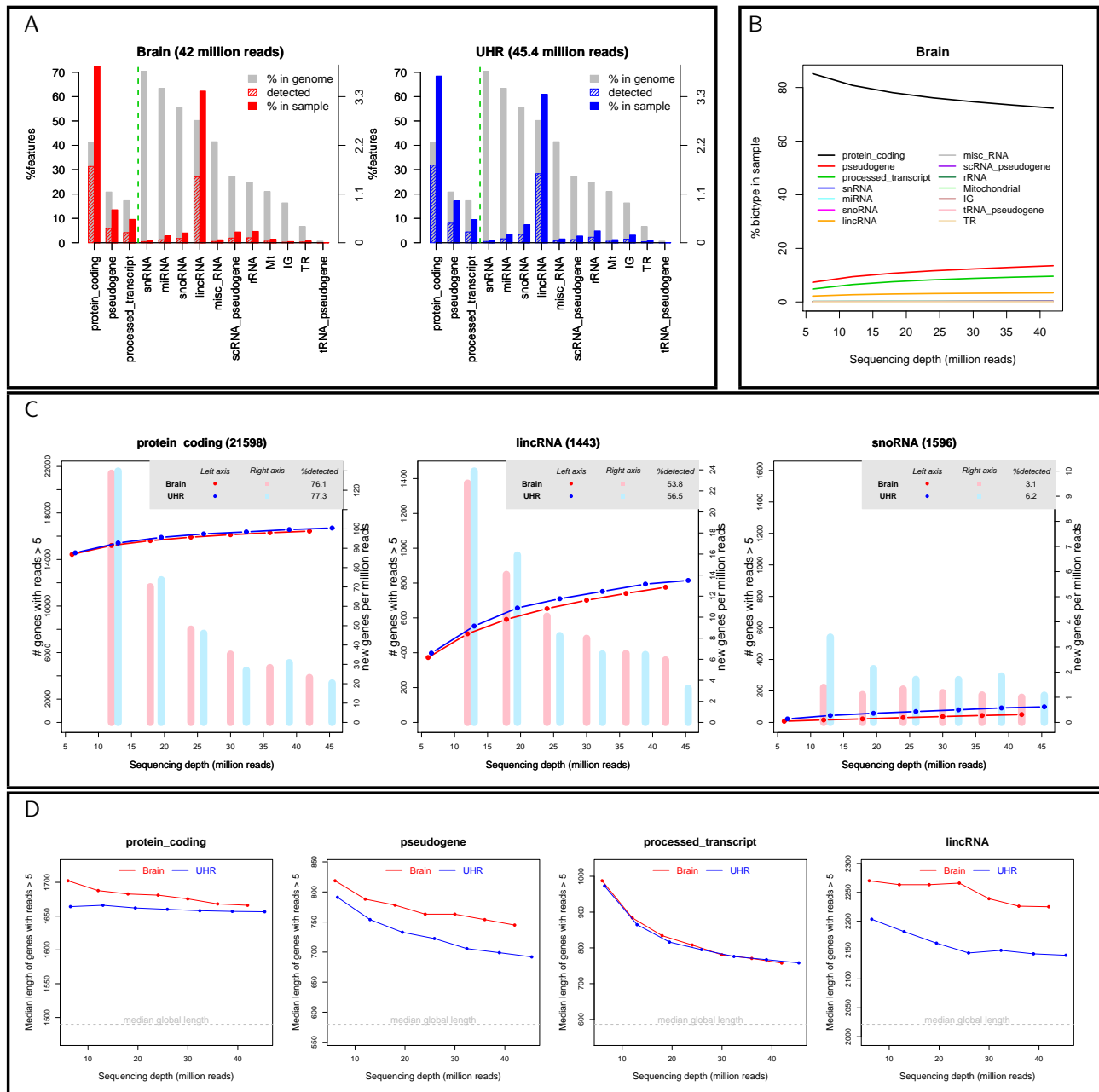


Figure 2

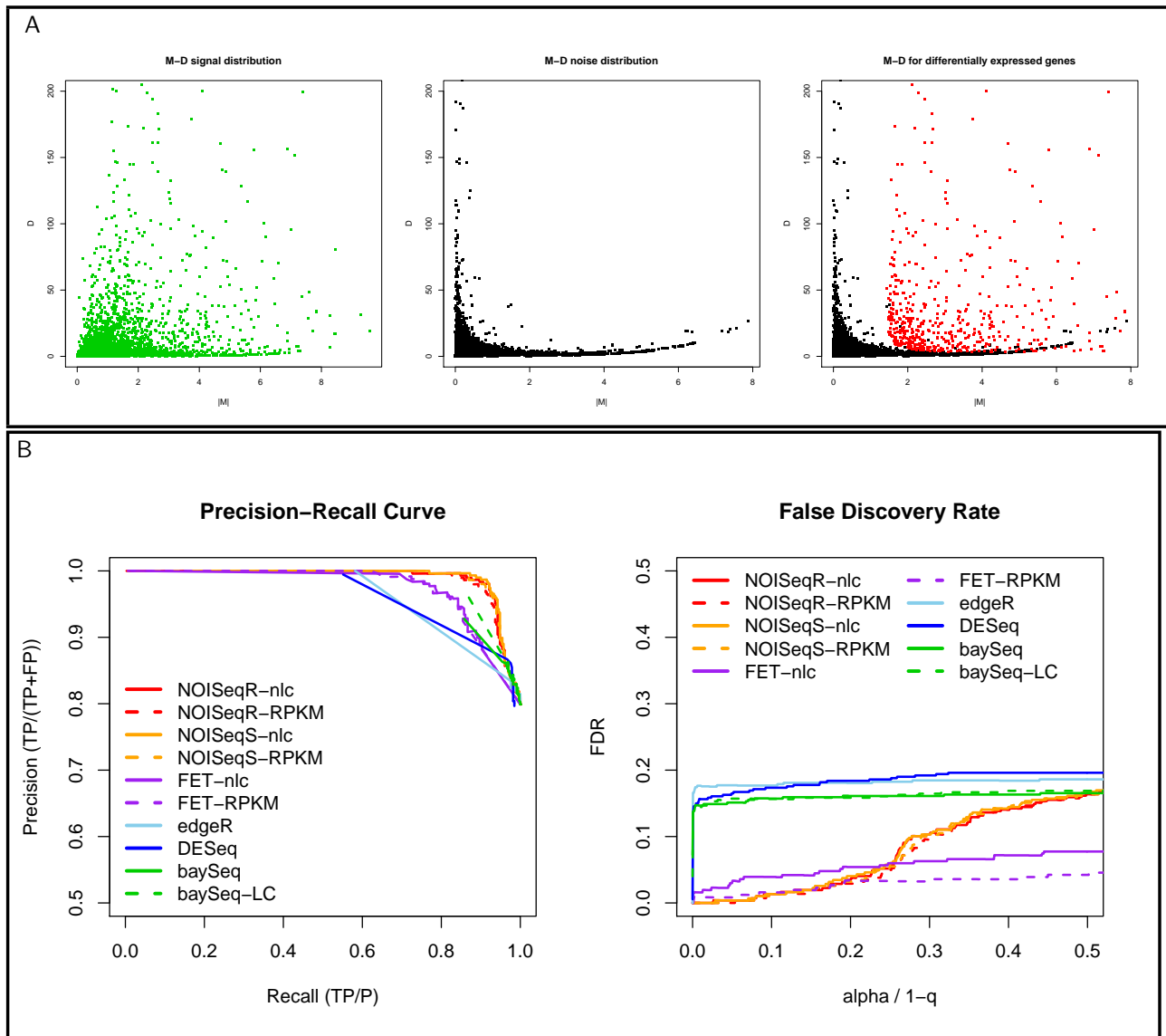


Figure 3

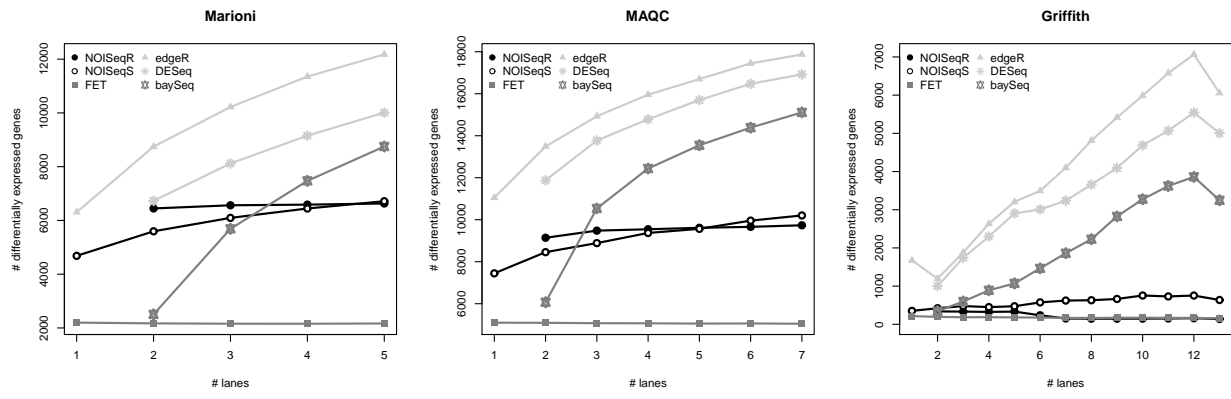


Figure 4

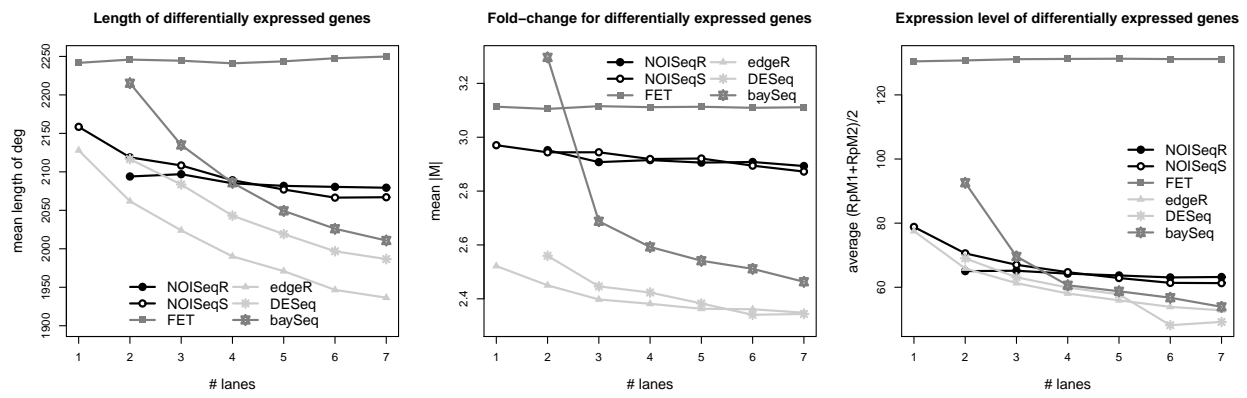


Figure 5

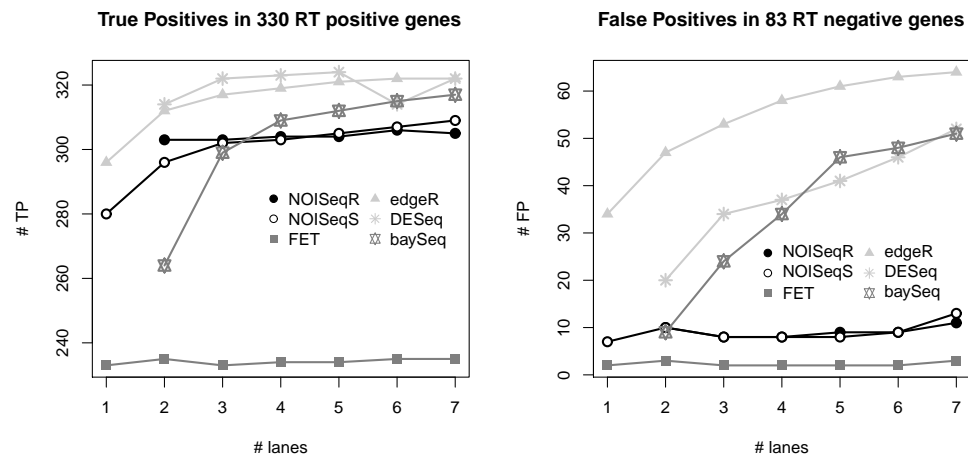


Figure 6

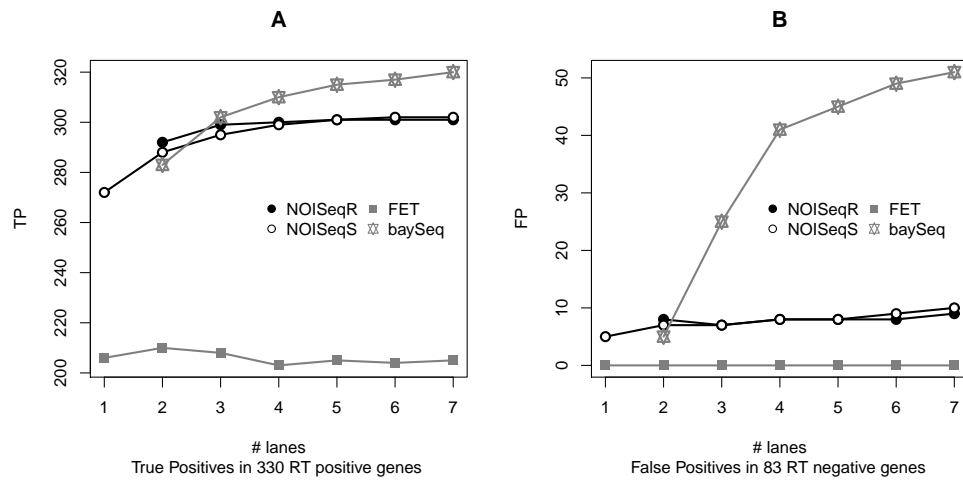


Figure 7

Figure legends

Figure 1. Saturation curves display the number of genes detected by more than 5 uniquely mapped reads as a function of the sequencing depth for each experimental condition in the three datasets (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 2. Feature detection and sequencing depth for MAQC data. (A) Detection percentages per transcript biotype. Gray bar indicates genome percentage, striped color bar is the percentage detected by the sample with regard to the genome, solid color bar is the percentage the biotype represents in the total detected features in the sample. Vertical line separates bars expressed in left and right y-axis scales. (B) Percentage of each transcript biotype within total detections at increasing sequencing depth (Brain sample). (C) Saturation curves and NDR bars for protein-coding, lincRNA and snoRNA. (D) Median transcript length as a function of sequencing depth for protein-coding, pseudogene, processed transcript and lincRNA biotypes. The median global length of each biotype is computed considering genes with median transcript length greater than 150 nts.

Figure 3. NOISeq method: description and performance. (A) Schematic representation of the NOISeq methodology. M-D distribution in noise (black), signal (green) and differentially expressed genes (red). Both axis scales have been trimmed to improve visualization. (B) Precision-Recall curves and False Discovery Rates for the differential expression methods compared on MAQC dataset using RT-PCR results as a gold-standard.

Figure 4. Differentially expressed genes according to sequencing depth for each dataset and method. No gene length correction was applied to the data.

Figure 5. Relationship between gene length, fold-change M and expression level of differentially expressed genes and the number of lanes used, for each method in MAQC dataset. No length correction was applied to the data. RpM_i is the number of reads in condition i per million reads, i.e. $RpM_i = \frac{10^6 \times \text{gene counts in condition } i}{\text{total counts in condition } i}$.

Figure 6. Relationship between the number of True Positives (TP) and False Positives (FP) and sequencing depth. TP and FP were obtained applying different statistical methods on MAQC dataset and comparing results to RT-PCR positive and negative genes.

Figure 7. Differential expression in MAQC dataset according to sequencing depth for methods with gene length correction using RT-PCR data as a gold-standard. (A) True Positives. (B) False Positives.

References

- Anders, S. 2010. Htseq: Analysing high-throughput sequencing data with python. <http://www-huber.embl.de/users/anders/HTSeq/>.
- Anders, S and Huber, W. 2010. Differential expression analysis for sequence count data. *Genome Biology* **11**(10):R106.
- Anderson, J. 2005. RNA turnover: unexpected consequences of being tailed. *Current biology* **15**(16):R635–R638.
- Argout, X, Salse, J, Aury, J, Guitinan, M, Droc, G, Gouzy, J, Allegre, M, Chaparro, C, Legavre, T, Maximova, S, et al. 2010. The genome of *Theobroma cacao*. *Nature Genetics* **43**(2):101–108.
- Blencowe, BJ, Ahmad, S, and Lee, LJ. 2009. Current-generation high-throughput sequencing: deepening insights into mammalian transcriptomes. *Genes & Development* **23**(12):1379–1386.
- Bloom, J, Khan, Z, Kruglyak, L, Singh, M, and Caudy, A. 2009. Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. *BMC Genomics* **10**(1):221+.
- Bullard, JH, Purdom, E, Hansen, KD, and Dudoit, S. 2010. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11**(1):94+.
- Carninci, P, Kasukawa, T, Katayama, S, Gough, J, Frith, M, Maeda, N, Oyama, R, Ravasi, T, Lenhard, B, Wells, C, et al. 2005. The transcriptional landscape of the mammalian genome. *Science (New York, NY)* **309**(5740):1559.
- Davis, J and Goadrich, M. 2006. The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240. ACM.
- Durbin, RM, Altshuler, DL, Durbin, RM, Abecasis, GR, Bentley, DR, Chakravarti, A, Clark, AG, Collins, FS, De La Vega, FM, Donnelly, P, et al. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467**(7319):1061–1073.
- Flicek, P, Amode, M, Barrell, D, Beal, K, Brent, S, Chen, Y, Clapham, P, Coates, G, Fairley, S, Fitzgerald, S, et al. 2011. Ensembl 2011. *Nucleic acids research* **39**(suppl 1):D800.
- Graveley, B, Brooks, A, Carlson, J, Duff, M, Landolin, J, Yang, L, Artieri, C, van Baren, M, Boley, N, Booth, B, et al. 2010. The developmental transcriptome of *Drosophila melanogaster*. *Nature* .
- Griffith, M, Griffith, OL, Mwenifumbo, J, Goya, R, Morrissy, AS, Morin, RD, Corbett, R, Tang, MJ, Hou, YC, Pugh, TJ, et al. 2010. Alternative expression analysis by RNA sequencing. *Nature Methods* **7**(10):843–847.
- Grzechnik, P and Kufel, J. 2008. Polyadenylation linked to transcription termination directs the processing of snoRNA precursors in yeast. *Molecular cell* **32**(2):247–258.
- Hardcastle, T and Kelly, K. 2010. baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics* **11**(1):422+.
- Illumina: Preparing Samples for Sequencing mRNA Illumina, INC. 2009. <http://icom.illumina.com/>.
- Kim, V, Han, J, and Siomi, M. 2009. Biogenesis of small RNAs in animals. *Nature Reviews Molecular Cell Biology* **10**(2):126–139.
- Leinonen, R, Sugawara, H, and Shumway, M. 2011. The sequence read archive. *Nucleic Acids Research* **39**(suppl 1):D19.
- Lemay, J, D’Amours, A, Lemieux, C, Lackner, D, St-Sauveur, V, Bähler, J, and Bachand, F. 2010. The nuclear poly (A)-binding protein interacts with the exosome to promote synthesis of noncoding small nucleolar RNAs. *Molecular cell* **37**(1):34–45.
- Li, N, Ye, M, Li, Y, Yan, Z, Butcher, L, Sun, J, Han, X, Chen, Q, et al. 2010. Whole genome DNA methylation analysis based on high throughput sequencing technology. *Methods* **52**(3):203–212.

- Locke, DP, Hillier, LW, Warren, WC, Worley, KC, Nazareth, LV, Muzny, DM, Yang, SP, Wang, Z, Chinwalla, AT, Minx, P, et al. 2011. Comparative and demographic analysis of orang-utan genomes. *Nature* **469**(7331):529–533.
- Marioni, JC, Mason, CE, Mane, SM, Stephens, M, and Gilad, Y. 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research* **18**(9):1509–1517.
- Mortazavi, A, Williams, BA, McCue, K, Schaeffer, L, and Wold, B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5**(7):621–628.
- Nagalakshmi, U, Wang, Z, Waern, K, Shou, C, Raha, D, Gerstein, M, and Snyder, M. 2008. The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science* **320**(5881):1344–1349.
- Oshlack, A, Robinson, M, and Young, M. 2010. From RNA-seq reads to differential expression results. *Genome Biology* **11**(12):220+.
- Oshlack, A and Wakefield, M. 2009. Transcript length bias in rna-seq data confounds systems biology. *Biology Direct* **4**(1):14+.
- Park, P. 2009. ChIP-seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics* **10**(10):669–680.
- Robinson, MD, McCarthy, DJ, and Smyth, GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**(1):139–140.
- Robinson, MD and Oshlack, A. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* **11**(3):R25+.
- Shi, L, Reid, L, Jones, W, Shippy, R, Warrington, J, Baker, S, Collins, P, De Longueville, F, Kawasaki, E, Lee, K, et al. 2006. The MicroArray Quality Control (MAQC) project shows inter-and intraplatform reproducibility of gene expression measurements. *Nature biotechnology* **24**(9):1151–1161.
- Slomovic, S, Laufer, D, Geiger, D, and Schuster, G. 2006. Polyadenylation of ribosomal RNA in human cells. *Nucleic acids research* **34**(10):2966.
- Srivastava, S and Chen, L. 2010. A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. *Nucleic Acids Research* **38**(17):e170.
- Sultan, M, Schulz, MH, Richard, H, Magen, A, Klingenhoff, A, Scherf, M, Seifert, M, Borodina, T, Soldatov, A, Parkhomchuk, D, et al. 2008. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* **321**(5891):956–960.
- Trapnell, C, Pachter, L, and Salzberg, S. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**(9):1105.
- Trapnell, C, Williams, BA, Pertea, G, Mortazavi, A, Kwan, G, van Baren, MJ, Salzberg, SL, Wold, BJ, and Pachter, L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* **28**(5):511–515.
- Velasco, R, Zharkikh, A, Affourtit, J, Dhingra, A, Cestaro, A, Kalyanaraman, A, Fontana, P, Bhatnagar, SK, Troggio, M, Pruss, D, et al. 2010. The genome of the domesticated apple (*Malus domestica* Borkh.). *Nature Genetics* **42**(10):833–839.
- Young, MD, Wakefield, MJ, Smyth, GK, and Oshlack, A. 2010. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biology* **11**(2):R14.
- Zheng, D, Frankish, A, Baertsch, R, Kapranov, P, Reymond, A, Choo, S, Lu, Y, Denoeud, F, Antonarakis, S, Snyder, M, et al. 2007. Pseudogenes in the ENCODE regions: Consensus annotation, analysis of transcription, and evolution. *Genome research* **17**(6):839.



Differential expression in RNA-seq: A matter of depth

Sonia Tarazona, Fernando García-Alcalde, Joaquin Dopazo, et al.

Genome Res. published online September 8, 2011

Access the most recent version at doi:[10.1101/gr.124321.111](https://doi.org/10.1101/gr.124321.111)

Supplemental Material <http://genome.cshlp.org/content/suppl/2011/09/08/gr.124321.111.DC1.html>

P<P Published online September 8, 2011 in advance of the print journal.

Accepted Manuscript Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.

Open Access Freely available online through the *Genome Research* Open Access option.

Creative Commons License This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see <http://genome.cshlp.org/site/misc/terms.xhtml>). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 3.0 Unported License), as described at <http://creativecommons.org/licenses/by-nc/3.0/>.

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).

Advance online articles have been peer reviewed and accepted for publication but have not yet appeared in the paper journal (edited, typeset versions may be posted when available prior to final publication). Advance online articles are citable and establish publication priority; they are indexed by PubMed from initial publication. Citations to Advance online articles must include the digital object identifier (DOIs) and date of initial publication.

To subscribe to *Genome Research* go to:
<http://genome.cshlp.org/subscriptions>
